S

SABIN VACCINE. See **Poliomyelitis; Virus.**

SABLE. See **Mustelines.**

SACCHARIDE. See **Carbohydrates.**

SACCHARIMETER. An instrument for the measurement of sucrose solutions. A saccharimeter differs from a polarimeter in that the saccharimeter uses white light, whereas a polarimeter is operated with sharply monochromatic light. Consequently, a saccharimeter can only be used with sugar solutions (in which case the quartz compensates for the rotatory dispersion of sucrose). Conversely, a polarimeter is suitable for the measurement of optical rotation of any solution, including sugar. However, saccharimetric sugar determinations are the basis of internationally accepted *sugar degrees* (°S), and saccharimeters are appropriately calibrated. When a polarimeter is employed that utilizes monochromatic light rather than a quartz wedge, deviations from sugar degrees will be found in some solutions, and they may not be inconsequential. The International Sugar Scale assigns 100°S to a pure sucrose solution of normal weight (26 grams in 100 milliliters of pure water) at 20°C and a 200-millimeter light path, measured in a saccharimeter with white light and a dichromate filter. Of course, an exact numerical conversion from sugar degrees to angular degrees is possible with pure sucrose (100°S corresponds to $\alpha = 34.6°$), but in practical, more or less impure solutions, the relationship is not exactly predictable, and the sugar scale is conventionally and legally binding. It is not surprising, then, that virtually all sugar laboratories use visual saccharimeters (also called polariscopes). They differ in vintage, the half-shade presentation, and in construction features, but they all use the quartz-wedge compensating principle, and they read out in sugar degrees. A typical instrument will have a split half-shade field or a triple field for observation. The quartz wedge is equipped with a fine scale that is read off a second observation tube. It is graduated in °S (e.g., -30 to $+110$°S) and comes with a vernier, readable to $\frac{1}{10}$°S.

See also **Photometers; Polarimetry;** and **Polarized Light.**

SACCHARIN. See **Sweeteners.**

SACCHAROMYCES. See **Yeasts and Molds.**

SACCHAROSE. See **Carbohydrates.**

SACRUM. The portion of the spinal column of vertebrates, usually formed of several fused vertebrae, with which the pelvis is articulated.

SADDLE POINT. A point (x_0, y_0) on a surface $f(x, y)$ where $f(x, y_0)$ is a maximum at $x = x_0$ and $f(x_0, y)$ at the same time is a minimum at $y = y_0$. A familiar example is the hyperbolic paraboloid, which has a saddle point at the origin if its standard equation is taken as $x^2/a^2 - y^2/b^2 = 2cz$. A person walking toward the origin in the XZ-plane would be ascending a mountain peak while he would be descending into a valley if he walked in the YZ-plane. It is also called a *minimax* or a *col.*

See also **Game Theory; Paraboloid;** and **Surface.**

SAFETY (Intrinsic). See **Intrinsic Safety.**

SAFETY SWITCH. See **Limit Switch.**

SAFETY VALVE. The common form of the safety valve is the pop valve held against its seat by a heavy spring and having a "huddling chamber" to make it open quickly and remain open until a predetermined pressure drop (2–4% of the working steam pressure) has occurred. The A.S.M.E. Boiler Construction Code requires boilers having more than 500 square feet (~46 square meters) of heating surface, or those generating better than 2000 pounds (907 kilograms) of steam per hour, to have two or more safety valves. The safety valves should have sufficient relieving capacity to prevent more than 6% pressure rise at maximum rate or combustion. Required discharge capacity of a safety valve may be based either on the heat units in the fuel consumed or on the amount of steam generated.

In case more than one safety valve is used, the smaller one can be set to pop at the desired maximum pressure and the larger at 2 or 3 pounds per square inch (0.1 or 0.2 atmospheres) higher.

The relief valve is a form of safety valve, but usually intended for less severe service and of less importance from the safety viewpoint. Relief valves are applied to air, water, and steam lines and also to tanks, heaters, and so on. Among them could be mentioned the back pressure valves and atmospheric relief valves.

SAFFLOWER (*Carthamus tinctorius*; *Compositae*). This plant, a native of the East Indies, is now widely cultivated in tropical Asia and Egypt and to a lesser extent in southern Europe and elsewhere. It is a low annual plant with yellowish-red flowers that have tubular corollas.

SAFFLOWER SEED OIL. See **Vegetable Oils (Edible).**

SAFFRON (*Crocus sativus*; *Iridaceae*). *Crocus sativus* is a perennial herb, the native home of which is the eastern Mediterranean region. The stem is an underground flattened corm, the surface of which is covered by a few scaly leaves. At the top of the corm is a terminal bud, which develops into linear leaves 5–9 inches (12.5–22.5 centimeters) long and flowers. The flowers are white or lilac-tinted, with the perianth six-parted and with a very long tube, so that the ovary remains below the surface of the ground. The three stigmas are bright red. These, when dried, are known as saffron, an orange-yellow dye with a considerable percentage of volatile oil present. See also **Flavorings.**

SAGE. See **Antioxidant; Flavorings; Labiatae.**

SAGEBRUSH (*Artemisia tridentata*; *Compositae*). A number of other species of this genus are also called sagebrush, but *A. tridentata* is more prominent and of wider distribution than the others. This shrub has an extensive root system and may attain heights up to 7 feet (2.1 meters). The gray-green foliage and aromatic odor of this shrub are distinctive. The flowers occur in inconspicuous heads. This plant is distributed from the Black Hills to southern British Columbia to southeastern California to northern Arizona. It reaches its best development, however, in the Great Basin region, where it may occur over large areas in nearly pure stands. See Also **Artemisia.**

SAGITTARIUS (the archer) This large constellation is the ninth sign of the zodiac. Lying as it does in a particularly rich portion of the Milky Way, it contains a large number of star clusters and gaseous nebulae of great beauty visible in a moderate-sized telescope. From the large number of faint stars, cepheid variables, and globular clusters that seem to congregate in this region, we can deduce that the stellar galactic system has its greatest extension and hence its center in this direction. Long-exposure photographs indicate that large numbers of dark or obscuring nebulae lie in this portion of the Milky Way. The only way to penetrate these clouds is by means of radio astronomy. (See map accompanying entry on **Constellations.**)

SAGITTA (the arrow). A northern constellation located next to Aquila.

SAIGA. See **Goats and Sheep.**

SAILFISH. See **Billfishes.**

SAILINGS (The). The position of a vessel at sea or in the air is defined by the latitude and longitude. The position at any particular instant is connected with any other position, either the one just left or the one toward which the vessel is proceeding, by means of the true course and distance.

Any given course and distance may be resolved into two components at right angles to each other: the northing or southing and the easting or westing, each expressed in nautical miles. The northing or southing may be immediately converted into difference of latitude, expressed in angular units, for the nautical mile is, by definition, approximately equal to a minute of arc along a great circle. However, the conversion from easting or westing, commonly known as departure, into difference of longitude, can be accomplished only after taking into account the shape of the earth and the approximate latitude of the ship.

The navigator is continually faced by one of two problems. (1) Given the difference of latitude and longitude between two points on the surface of the earth, to find the course and distance between them. (2) Given the course and distance followed by a ship, to find the difference of latitude and longitude between the point of starting and the destination. The different methods of solving these problems are known as the sailings and include plane, parallel, middle latitude, mercator, great-circle, and composite sailings.

See also **Course**; and **Navigation**.

SAILPLANE. A sailplane is a highly efficient glider. Being designed for the use of expert glider pilots, it is unsuited for primary training. It is characterized by very low sinking speed, nearly flat glide and perfection of construction. It is capable of rising flight on weak thermal air currents and is the type of aircraft employed for cross-country motorless flights of a sporting nature. The sailplane has high aspect ratio (about 20), careful streamlining, clean and smooth external surfaces, and minimum weight consistent with structural safety. The gliding angle in still air is often as flat as 22:1 and the sinking speed as small as 2 feet (0.6 meter) per second.

SAINT ELMO'S FIRE (or Corona Discharge). A brush-like, luminous, and often audible discharge from charged objects in the atmosphere. It occurs on ship masts, on aircraft propellers, wings, and other projecting parts, and on objects projecting from high terrain when the atmosphere is charged and a sufficiently strong electrical potential is created between the object and the surrounding air. Aircraft most frequently experience St. Elmo's fire when flying in or near cumulonimbus clouds or thunderstorms, in snow showers, and in dust storms.

SAITHE. See **Codfishes.**

SAKI. See **Monkeys and Baboons.**

SALAMANDER (*Amphibia, Urodela*). A vertebrate with a slender body, short legs, and a long tail. The moist skin of the amphibians limits them to protected habitats, either near water or under some protection on moist ground, usually in the woods. Some species are aquatic throughout life, some take to the water intermittently, and some are entirely terrestrial as adults. The salamanders resemble the lizards superficially, but they are easily distinguished by the moist skin, without scales. See also **Hellbender.**

SALICACEAE. See **Willow Trees.**

SALICYLIC ACID. Salicylic acid or $C_6H_4(OH)(COOH)$ is a white solid, melting points 159°C, sublimes at 76°C, insoluble in cold water, soluble in hot water, alcohol, or ether. With ferric chloride solution, salicylic acid solutions are colored violet (distinction from benzoic acid).

Salicylic acid may be obtained (1) from oil of wintergreen, which contains methyl salicylate or (2) heating dry sodium phenate C_6H5ONa plus carbon dioxide under pressure at 130°C and recovering from the resulting sodium salicylate by adding dilute sulfuric acid. Salicylic acid is a mild disinfectant and antiseptic and has been used as a food preservative. Salicyclic acid and certain salicylates are used in medicine as anti-rheumatics.

SALINOMETER. An instrument for determining salt concentration (salinity), particularly one based upon electric-conductivity measurements. See **Electrical Conductivity.**

SALIVA. See **Caries and Cariology.**

SALK VACCINE. See **Poliomyelitis; Virus.**

SALMON (*Osteichthyes*). The order *Salmoniformes* consists of eight sub-orders, with some quite unlike others. Most important commercially is the suborder *Salmonoidei*. These fishes are primarily migrating species and associated with the freshwater of the northern hemisphere. The three subfamilies of *Salmonoidei* include the *ayus* and *smelts*. Many of the salmon species are good eating not only because of their flavorful, fatty meat, but also because they lack those bones that in most fishes are embedded in the cartilaginous walls between the muscular segments. Salmonidae contains such familiar fishes as *salmon, trout,* and *chars*.

Atlantic Salmon. The species *Salmon solar,* prior to extensive pollution of certain waters, was one of the most prevalent fishes in the Atlantic drainage areas. Its distribution extended from Kara in northeastern Russia along the coast of Europe to Douro in the northwestern part of the Iberian peninsula, and on to Iceland, the southern tip of Greenland, and across Newfoundland to Cape Cod in the northeastern United States. Salmon migrate extensively. The early part of the salmon's life is spent in the upper courses of large rivers. Then they migrate into the sea, where they grow relatively quickly, and then return to swim up rivers for spawning. During their stay in the oceans, salmon traverse great distances. Thus, salmon marked off the European coast have been recovered in waters of western Greenland. Generally, however, the salmon stays near the shore. Feeding grounds are primarily in the southern Baltic Sea and off northwestern Norway. When preying upon other fishes, they are found in the upper water levels to a depth of about 32 feet (10 meters), but may penetrate deeper.

Some studies indicate that their distribution at various depths depends upon daily and seasonal changes. During their period in the sea, salmon grow at a remarkable rate, often exceeding 2.2 pounds (1 kilogram) per month. They spend 1 to 3 years in the ocean before returning to the rivers to spawn. During this period, they have stored great quantities of fat, so much that their skin is orange-red. They leave freshwater when they are from 4 to about 8 inches (10 to 20 centimeters) long. After a year in the ocean, they measure nearly 20 to 26 inches (51 to 66 centimeters) in length and weigh from 3.3 to 7.7 pounds (1.5 to 3.5 kilograms). After 2 years, their length is from about 28 to 36 inches (71 to 91 centimeters), with a weight of from 9 to 17.5 pounds (4 to 8 kilograms). After 3 years, salmon are from 36 to 41 inches (91 to 104 centimeters) long and weigh from 17 to nearly 28 pounds (8 to 13 kilograms). Salmon probably reach a maximum age of 10 years. Occasionally, old males up to 41 inches (104 centimeters) in length and weighing as much as 80 pounds (36 kilograms) are caught. Females are generally smaller and rarely exceed a length of 39 inches (100 centimeters). Their greatest weight is 44 pounds (20 kilograms). See Fig. 1.

Salmon from various rivers meet at the feeding grounds. When the spawning season comes, they separate once again and each salmon seeks out the river in which it was born. The exact manner in which salmon find their way back to their home river is not understood. It is only known for certain that their olfactory sense plays a crucial role in the second phase of their ascent up the river. This ascent takes place

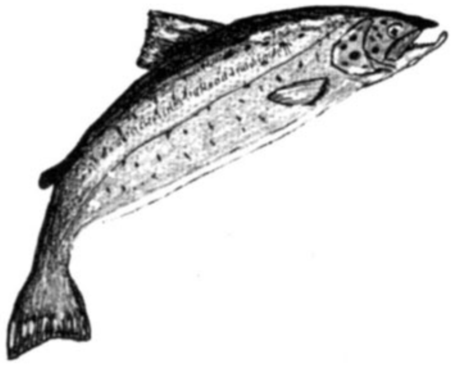


Fig. 1. Atlantic salmon.

throughout the year in some rivers, while in others it is found only at certain seasons. Often, larger salmon are seen ascending a river at one period, while smaller ones are found at some other time. Four major types are distinguished—large and small summer salmon and large and small fall or winter salmon. When they meet at the river mouths, the salmon can be distinguished by the development of their gonads. The individual groups then seek out their various spawning sites. If these sites are far from the mouths of rivers, it is generally the large fall salmon that ascend the river. Their germ cells are still immature when they begin their ascent. The ascent is interrupted by the onset of frost; at this time, the salmon winter somewhere upriver and will reach their spawning sites in the following fall.

While the energetic large salmon are capable of swimming great distances upstream, the smaller species generally find spawning sites near the river mouths. Each day these salmon cover greater distances. In small Scottish rivers, daily distances of up to 34 miles (58 kilometers) have been recorded. It must also be recalled that the salmon only migrate for 5 to 6 hours per day, during which time they swim with great strength and endurance. Their highest speed has been estimated at 10 miles (13 kilometers) per hour. They can also swim through rapids. Against a current of 19.6 feet (6 meters) per second, a salmon can push ahead at a rate of over 3.3 feet (1 meter) per second. Small waterfalls are passed by jumping over them, and leaps of up to 10.8 feet (3 meters) high and 16.4 (5 meters) long have been observed. To jump out of the water, the salmon swim up through the water surface on a slant, and with a particularly strong beat of the tail they gain additional acceleration at the surface. While jumping, salmon typically show a distinct lateral arching of the body. If the first attempt fails, the leaps are repeated constantly. This often results in skin injuries in stony waters, which sometimes become infected and cause death prior to reaching the spawning site.

During the entire migration, the salmon virtually cease feeding from the time they enter the river. During the journey, their fat reserves are converted into energy, and the orange-red hue of the skin disappears. As the germ cells mature, the salmon also alter their appearance. In the ocean, salmon have rather plain coloration, with a gray-green back, silvery sides, and white belly. X-shaped spots are above the lateral line and round black spots mark the head. However, when migrating upstream, a brilliant coloration develops. The back becomes considerably darker, while the sides take on a bluish shimmer and the stomach becomes reddish. Purple-red spots appear beside the black ones, and even the lower sides of the pectoral, anal, and caudal fins take on reddish hues. The color change is accompanied by a major anatomical modification of the lower jaw. The jaw points upward and develops cartilaginous growths from which a hook-shaped appearance develops.

The salmon spawning period in central European waters occurs generally from mid-November to mid-December. In the north, the onset of spawning is in mid-September and in some groups may last until February. Spawning sites are located in regions with clear, cold, oxygen-rich, fast-flowing water and a clean gravel bottom. Salmon typically seek gravel banks in the upper water levels at a depth of about 1.5 feet (0.5 meter). Once the female arrives at the spawning site, she prepares the nest. The female digs up the floor with powerful rump and tail motions and constructs a depression some 4 to 8 inches (10 to 20 centimeters) deep and often well over 3 feet (1 meter) long. Bohemian fishermen have said that a horse could fit in one of these nests! During the entire period, there are usually several males in the vicinity, but they do not assist with construction of the pit. All salmon species carry out their spawning habits in a similar manner. After maneuvers of rivalry among males, the male and female salmon will swim side-by-side, pressing close together with mouths open just above the spawning pit, and the eggs and sperm are released. Eggs are released several times, interrupted by additional rivalry fights and more courtship display behavior. A single female lays a total of from 10,000 to 30,000 eggs. It lays about 500 to 2000 eggs per kilogram (2.2 pounds) of body weight. This is actually a low number when compared to some fishes, such as the carp. The yolk-rich, sticky eggs are from 5 to 7 millimeters large and, according to water temperature, will lie between the stones in the spawn pit for some 70 to 200 days. The young hatch in April or May. As long as the larvae still have yolk upon which to feed, they remain hidden in the pit. When that supply is exhausted they move into the water and initially feed on small crustaceans and insect larvae. As they get older, the young salmon increasingly feed on fishes and, at the end of their juvenile period, their diet is exclusively small fishes. The young stay in freshwater for 1 to 2 years (in the north, up to 5 years), after which they enter the oceans and take on their typical oceanic coloration.

The *effects of pollution* on the Atlantic salmon population have been marked and have been going on for decades. These fishes have almost completely disappeared from the Rhine, Weser, and Elbe Rivers in Germany. A few isolated salmon still appear in the lower Rhine, but they are inedible because their fat is contaminated with phenol and other wastes in the river. In Europe, the decline of the salmon began in the mid-1890s as the rivers became polluted. In England, industrialization began earlier, and thus the Thames, once a salmon river, became contaminated still earlier. The last salmon caught in the Thames was recorded in 1833. Until the early part of the twentieth century, the East Prussian coast was the most productive one for German commercial fishing. Large catches no longer are made. Currently, most European salmon are caught in Norway and Denmark. Canada has the greatest catch in the world. Salmon are generally smoked.

Inevitable dying after the first spawning is a rare phenomenon among vertebrates. Death of the adult salmon occurs in nearly all instances within 1 to 2 weeks after spawning. The immediate cause of the post-spawning death is not well understood. Physical exhaustion from the long migrations does not appear to be the principal cause of death, since salmon running up short streams may reach the spawning site in very good condition, but they still undergo degeneration that rapidly leads to death after shedding their sex products.

Pacific Salmon. There are several species: *chinook* (*Oncorhynchus tshawytscha*); the *coho* (*O. kisutch*); the *sockeye* (*O. nerka*); the *chum* (*O. keta*); and the *pink* (*O. gorbuscha*). These are generally known as the North American species. There is an additional Asian species (*O. masou*), the common name of which is *masu* salmon.

The various species of salmon vary greatly in size reached at maturity (Harry 1969). The chinook is the largest of the Pacific salmon, with a record weight of 126 pounds (57.2 kilograms), but with an average weight of about 20 pounds (9.1 kilograms). Coho salmon range in size up to 30 pounds (13.6 kilograms), with most of the fish being in the 8-to-12-pound (3.6-to-5.4-kilogram) category. Chum salmon weigh as much as 33 pounds (15 kilograms) and the average weight is about 8 pounds (3.6 kilograms). Sockeye salmon are usually between 5 and 7 pounds (2.3 and 3.2 kilograms), but weights up to 15 pounds (6.8 kilograms) have been recorded. Pink salmon usually weigh between 3 and 5 pounds (1.4 and 2.3 kilograms) and specimens up to 12 pounds (5.4 kilograms) have been taken. The masu salmon has an average weight of about 10 pounds (4.5 kilograms) and a maximum weight of 20 pounds (9.1 kilograms).

Chinook Salmon. This fish can be distinguished from the other species by the heavy black spotting on the back, the dorsal fins, and both lobes of the caudal fin, as well as by the black pigmented skin along the base of the teeth. Young chinook salmon can be recognized in freshwater by strongly developed parr marks (marks characteristic of young fishes). The chinook salmon enter the Sacramento-San Joaquin system of California to spawn, but are found only rarely in streams south of San Francisco Bay. Spawning occurs in rivers north to the Bering Sea and on the Asiatic side down to the Amur River, although rarely that far south, and in Japan in the rivers' of northern Hokkaido. In the North Pacific Ocean, chinook salmon are found generally to the north of 46° north latitude in the eastern half of the ocean, but west of the 180th meridian they occur about as far south as 42° north latitude. Chinook salmon generally ascend the larger streams to spawn, and they are abundant in such rivers as the Sacramento, Columbia, Fraser, and Yukon. The center of abundance of this species is the Columbia River, in which adults return to spawn during every month of the year. Each female chinook salmon carries from 2,000 to 13,000 eggs. After the eggs are deposited, they are protected by a cover of gravel. Hatching may take as long as 4 months, depending upon water temperature. From a commercial tonnage standpoint, the chinook salmon represents only about 4% of the total salmon catch of the major fishing countries (Canada, the United States, Japan, and the U.S.S.R.).

Coho Salmon. This fish is found in North American rivers and streams from Monterey Beach, California, in the south, to the Chukchi Sea the north. In Asia, they are rarely found in the Anadyr River, but occur in large numbers in Kamchatka and south almost to the Amur River. The species is also present on Sakhalin Island and in Hokkaido. During their fast-growing period in the ocean, coho salmon are distributed across the northern Pacific Ocean and in the Bering Sea. They are found as far south as California waters in the eastern half of the northern Pacific and in the western North Pacific, they are found as far south as 42° north latitude.

Adult coho salmon begin the freshwater migration between September and December, often coincident with a freshet (stream of freshwater flowing into the sea). This species enters the larger rivers, but also is common in the very small coastal streams throughout its range. Spawning takes place often in tiny tributaries only 3 to 4 feet (1 meter plus) wide. The young emerge from the gravel in the early spring. Coho salmon usually remain in freshwater for about 1 year after fry emergence and then begin their downstream journey to the sea. In Alaskan streams, coho commonly remain in freshwater for 2 years. Larger coho salmon feed principally on squid, small fish, and euphausiids. Adult coho salmon return to spawn late in the year following that in which they entered the ocean. In their last summer of ocean life, coho salmon grow very rapidly and commonly double their weight during this period. Of the commercial tonnage taken by the four major salmon fishing countries previously mentioned, the coho catch represents about 7% of the total. See also entry on **Aquaculture.**

Chum Salmon. This fish is distributed along the North American coast from the Klamath River in California, north to the Arctic coast of Alaska and even in the MacKenzie River. In Asia, chum salmon are abundant in the Amur River and are found on the island of Sakhalin and in Hokkaido streams and south almost to Tokyo. Chum salmon are in the Pacific Ocean north of 45° north latitude in the eastern part of the ocean, and south almost to 36° north latitude in the western Pacific, as well as in the Bering Sea.

Chum salmon enter freshwater to spawn from July to January. To the north, in Alaska and northern British Columbia, the runs are primarily between July and early September, while south of Vancouver Island the runs are from October through January. Chum salmon usually spawn in smaller coastal streams or in the lower portions of larger rivers. However, they ascend several hundred miles (kilometers) up the Yukon to spawn and also spawn in the headwaters of some Asian rivers. Chum salmon frequently dig their nests in the same area as used by pink salmon. The young emerge from the gravel in March through May. Newly emerged chum salmon are a little over 1 inch (2.5 centimeters) long. In most North American streams, young chum salmon almost immediately move downstream to salt water, but in some Asian streams and the Yukon River, it may be several weeks before the young reach the ocean. Chum salmon usually mature at 4 years, but 3- and 5-year-old fish are common. Of the commercial tonnage taken by the four major salmon fishing countries, the chum salmon catch represents about 24% of the total.

Sockeye Salmon. This fish occurs rarely in North American coastal streams south of the Columbia River, but it can be found north to the Yukon River of Alaska, along the Asian coast from the Anadyr River, in the rivers of Kamchatka, where it is the principal species in the Kamchatka River, and south to northern Hokkaido, where it is very rare. Sockeye salmon are distributed in the ocean throughout the northern North Pacific and the Bering Sea.

The mature sockeye salmon male in freshwater becomes a brilliant red and the female is a dark red. This species is distinguished from others by having 28 to 40 long, slender, closely set gill rakers on the first gill arch. There are no black spots. Young sockeye salmon have oval parr marks, which extend only slightly below the lateral line. Sockeye salmon are typically lake-dwelling during the brief freshwater part of their life. After migrating from the lake they usually spend 2 to 3 years at sea. Adult sockeye may begin the upstream migration as early as May and in some areas the migration may extend into October. Sockeye salmon spawn in outlet and inlet streams of lakes and also along some lake shores. After leaving the protection of the gravel, the young move into lakes, where they remain for 1 to 3 years before migrating to the ocean at a length of from 3 to 6 inches (7.5 to 15 centimeters). The downstream migration takes place from April to June, usually under the protection of darkness. In the North Pacific, sockeye salmon feed heavily on amphipods, copepods, euphausiids, pterepods, fish, squid, and similar animals. In the ocean, growth is rapid and those fish that return to freshwater after spending 2 years in the ocean are about 21 inches (53 centimeters) long.

In the ocean, there is a considerable overlap of sockeye salmon from Kamchatka and North America. Maturing sockeye salmon in the high seas begin moving shoreward in May and June. The very important Bristol Bay (Alaska) runs travel an average of about 24 miles (39 kilometers) per day when heading toward their home stream.

Of the commercial tonnage taken by the four major salmon fishing countries, the sockeye salmon catch represents about 24% of the total.

Pink Salmon. This fish occurs occasionally in streams south of Puget Sound and from there northward into Arctic Ocean streams of Alaska. Pink salmon are also found in the MacKenzie River. On the Asiatic side of the Pacific, pink salmon have been reported from the Lena River and south into the streams of the Sea of Okhotsk, the Kurile Islands, Sakhalin, Hokkaido, and on the northeastern coast of Hondo. Pink salmon are found usually north of the 40th parallel across the entire north Pacific Ocean.

Pink salmon move from the ocean into the streams where they spawn from July to November. Adults usually migrate only a short distance from salt water and sometimes spawn in streams in areas that are affected by the tide. In larger rivers, pink salmon may move considerable distances upstream. The young migrate to salt water as soon as the yolk sac is absorbed, in March through early June, but principally in April. In many of the smaller coastal streams of Alaska, the young pink salmon emerge from the gravel and migrate to the sea in one night. Downstream movement begins as darkness approaches and ceases before morning. In the estuaries, schools of young fish migrate along the shore near the surface where the currents gradually carry them toward the ocean. Here their food consists of euphausiids, amphipods, pterepods, small fish, crustaceans, larval squid, and copepods. Pink salmon migrate from the stream to the estuaries at about 1.5 inches (3.8 centimeters) in length and, when they return to spawn after 15 to 17 months of life in the ocean, their average length is about 20 inches (51 centimeters). See Fig. 2.

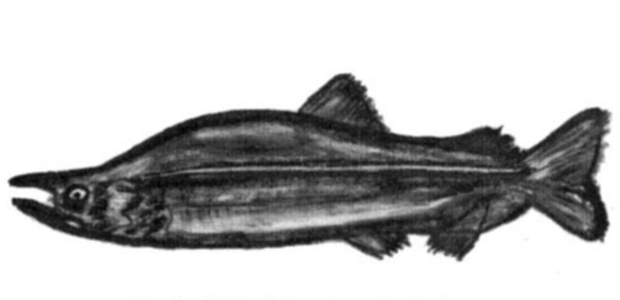


Fig. 2. Male pink or humpback salmon.

Of the commercial tonnage taken by the four major salmon fishing countries, the pink salmon catch represents about 41% of the total.

Masu Salmon. Masu salmon are found only in streams on the Asiatic side of the Pacific Ocean, from the Amur to the Pusan River, in Sakhalin streams, and on the island of Hokkaido and Hondo. They are also occasionally taken on the western coast of Kamchatka.

Salmon Products. Salmon are sold as fresh fish and are also canned, smoked, and fresh-frozen. Salmon eggs are used as bait for sport fishermen or are eaten as red caviar or salted as food. Salmon for the fresh market are usually troll caught in the ocean within a few miles of the port of landing. Coho salmon is the principal species used in the fresh market, but fresh chinook and pink salmon are also marketed. By far the greatest percentage of sockeye, pink, and chum salmon are processed in cans weighing 1 pound (0.45 kilogram) or less.

Commercial salmon fishing, as an industry, commenced in North America in the early 1860s, with the first commercial pack produced at Sacramento, California. The industry soon moved north where canneries on the Fraser River in British Columbia, on Puget Sound, and on Prince of Wales Island in Alaska were established by the late 1870s. The salmon catch in North American waters reached its first peak in the 1930s, after which production has been cyclic. Japanese salmon fisheries commenced in the early 1950s and developed rapidly. Somewhat later, the U.S.S.R. entered into serious commercial salmon fishing operations and is now a major factor.

Freshwater Salmon. An important sport fishery for salmon has been developed in the Great Lakes.

Salmon Technology. Much research has been devoted to protecting and expanding the traditional salmon fisheries as well as developing sport fisheries for salmon in freshwater areas, such as the Great Lakes. Experimental stockings have been carried out with coho and chinook salmon in these lakes, notably Lake Michigan and Lake Superior.

In the area from British Columbia to California, and to some extent in Alaska, an increasing percentage of adult salmon found passage hindered or blocked by power, irrigation, or flood control dams. The Columbia River and its wild tributary, the Snake River, have increasingly become polluted by wastes as well as thermally. Fishways have been constructed over the low dams of the main Columbia and Snake Rivers, and these have been generally successful in passing adult salmon. Downstream migrating salmon pass through the turbines of the low dams or over the spillways with usually only a small percentage of loss.

In recent years, fish passage facilities have been provided at most dams, but at some high dams, fishway construction has not proved feasible and other methods have been provided for maintaining the runs. Chief among these has been the production from salmon and steelhead hatcheries. In recent years, research efforts have resulted in a great improvement in the success of rearing salmon in hatcheries. Many hatchery diseases can now be controlled, and nutritious diets have been developed for young salmon. Scientists have also given attention to improving survival of salmon eggs and fry by constructing spawning channels. An artificial spawning channel generally consists of a dam at the head to control the flow of water and an artificial channel of appropriate width and slope with gravel of optimum size for best egg and fry survival. Many of the conditions detrimental to the survival of salmon eggs and fry in nature can be controlled in such artificial spawning areas. The optimum number of adults can be allowed to spawn. Flood waters, which wash eggs and fry from the gravel, can be prevented. Predators can be controlled. Gravel can be selected to allow optimum circulation of water with its life-giving oxygen, and the amount of water needed to ensure best production of fry can be maintained. At Jones Creek, a tributary of the Fraser River in British Columbia, one of the first artificial spawning channels for salmon was constructed as early as 1954, where it was found that the survival of pink salmon eggs and fry was from 4 to 6 times greater than that of the natural environments (Harry 1969). See also **Aquaculture.**

The smelt, closely related to the salmon, is described in the entry on **Smelt.**

Literature references mentioned in this entry will be found listed at the end of the entry on **Fishes.**

SALMONELLA. See **Toxin.**

SALMONELLOSIS. See **Foodborne Diseases.**

SALPINGITIS. Infection of the fallopian tubes. This common disease is most frequently due to a gonorrheal infection. More rarely, it may be due to tuberculosis, *streptococcus*, or *pneumococcus* infection. Gonorrheal salpingitis is not only the most prevalent but is the most disabling in its aftereffects. Salpingitis may not occur for months or even years after the original gonorrheal infection.

SALSIFY. See **Composite Family.**

SALT. A compound formed by replacement of part or all of the hydrogen of an acid by one (or more) element(s) or radical(s) that are essentially inorganic. Alkaloids, amines, pyridines, and other basic organic substances may be regarded as substituted ammonias in this connection. The characteristic properties of salts are the ionic lattice in the solid state and the ability to dissociate completely in solution. The halogen derivatives of hydrocarbon radicals and esters are not regarded as salts in the strict definition of the term.

In the classical concept of the process of neutralization, whereby an acid and a base in solution react to form a salt, the proton of the acid and hydroxyl ion of the base react to form water, leaving the cation of the base and the anion of the salt by recombination.

Upon evaporation of the solvent, the salt is obtained as such, frequently as crystals, sometimes with and sometimes without water of crystallization. A salt, when dissolved in an ionizing solvent, or fused (e.g., sodium chloride in water), is a good conductor of electricity and when in the solid state forms a crystal lattice (e.g., sodium chloride crystals possess a definite lattice structure for both sodium cations (Na^+) and chloride anions (Cl^-), determinable by examination with x-rays).

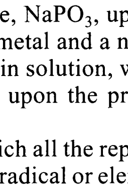
A broader definition than that confined to solutions is demanded in some fields of chemistry (e.g., in high temperature reactions of acids, bases, and salts). In the formation of metallurgical slags, at furnace temperatures, calcium oxide is used as base and silicon oxide and aluminum oxide as acids; calcium aluminosilicate is produced as a fused salt. Sodium carbonate and silicon oxide when fused react to form the salt sodium silicate with the evolution of carbon dioxide. In this sense:

$$\left[\begin{array}{l}\text{Oxide of any}\\ \text{element functioning}\\ \text{as a metal—that is,}\\ \text{as a base}\end{array}\right] \text{ plus } \left[\begin{array}{l}\text{Oxide of any}\\ \text{element functioning}\\ \text{as a nonmetal—that}\\ \text{is, as an acid}\end{array}\right] \text{ yields [Salt]}$$

Iron and sulfur when heated react to form the salt ferrous sulfide. In this sense:

metal plus nonmetal yields salt

Salts therefore, are prepared (1) from solutions of acids and bases by neutralization and separation by evaporation and crystallization; (2) from solutions of two salts by precipitation where the solubility of the salt formed is slight (e.g., silver nitrate solution plus sodium chloride solution yields silver chloride precipitate [almost all as solid], and sodium nitrate present in solution as sodium cations and nitrate anions [recoverable as sodium nitrate, solid by separation of silver chloride and subsequent evaporation of the solution]); (3) from fusion of a basic oxide (or its suitable compound—sodium carbonate above) and an acidic oxide (or its suitable compound—ammonium phosphate), since ammonium and hydroxyl are volatilized as ammonia and water. Thus, sodium ammonium hydrogen phosphate



yields sodium metaphosphate, $NaPO_3$, upon heating. (4) Salts also are prepared from reaction of a metal and a nonmetal.

Reactions of salts as such in solution, without decomposition of cation or anion, are dependent upon the presence of the cation and the anion of salt.

An *acid salt* is a salt in which all the replaceable hydrogen of the acid has not been substituted by a radical or element. These salts, in ionizing, yield hydrogen ions and react like the acids (e.g., $NaHSO_4$, $KHCO_3$, Na_2HPO_4).

An *amphiprotic* (also called *amphoteric*) *salt* is a salt that may ionize in solution either as an acid or a base, and react either with bases or acids, according to the conditions.

A *basic salt* is a salt contains combined base as $Pb(OH)_2$ $Pb(C_2H_3O2)_2$, a basic acetate of lead. These salts may be regarded as formed from the basic hydroxides by partial replacement of hydroxyl (e.g., HO—Zn—Cl). They react like bases and, when soluble, ionize to yield hydroxyl ions.

A *complex salt* is a saline compound having the structure of a combination of two or more salts and that is regarded as the normal salt of a complex acid. Complex salts do not split into a mixture of the constituent salts in solution, but furnish a complex ion that contains one of the bases (e.g., potassium molybdophosphate and potassium platinochloride).

A *double salt* is a substance consisting of two simple salts that crystallize together in definite proportions and exist independently in solution (distinction from complex salts). The alums are representative double salts.

An *inner salt* is a member of a special class of internal salts in which an acid group and a neutral group coordinate with metals to form a cyclic complex. These salts occur widely in analytical chemistry, (where they are formed between metallic ions and organic reagents) in dyestuffs, in life processes (chlorophyll and hematin belong to this class of compounds), and in many other fields.

An *internal salt* is a compound in which the acidic or basic groups that react to produce the salt linkage (which may or may not entail the formation of water) are in the same molecule. This particular salt linkage may consist of a polar or a nonpolar bond.

A *mixed salt* is a salt of a polybasic acid in which the hydrogen atoms are replaced by different metallic atoms or positive radicals.

A *pseudo salt* is a compound that has some of the normal characteristics of a salt, but lacks certain others, notably the ionic lattice in the solid state and the property of ionizing completely in solution. The absence of these properties is due to the fact that the bonds between the metallic and nonmetallic radicals are covalent or semicovalent instead of polar. Because these salts do not ionize completely, they are also called *weak salts*.

SALTATION. This term, as proposed by McGee, in 1908, is used by geologists to designate the particular mode of the stream transportation of clastic sediments by intermittent leaps or bounds. Probably an important factor in the ultimate transportation of the coarser fragments by streams and rivers.

SALT BRIDGE. A type of liquid junction used to connect electrically two electrolytic solutions. It consists commonly of a U-tube filled with a strong solution, and provided with porous plugs. It is used for such purposes as to connect electrolytic half cells in making measurements of electrode potential.

SALT (NaCl). See **Sodium Chloride.**

SALTPETER. Potassium nitrate. Sodium nitrate is often called Chile saltpeter, and calcium nitrate is sometimes called Norway saltpeter.

SALT RIVER. Although many rivers of the world contain relatively high contents of salt, a salt river is usually identified as a river that penetrates into a huge deposit of rock salt and that forms glistening white cliffs of salt in its course. One of these formations occurs in a broad bend of the Huallaga River in eastern Peru. The salt cliff extends some 8 kilometers in length and is about 100 meters in height. The Huallaga is a major tributary of the Amazon River. Several exposed salt deposits of this kind are found in the region of the Amazon headwaters. It is reasoned that during the formation of the Andes, compression forces caused rock salt to flow in a way somewhat analogous to the manner of flow of glacial ice. Salt deposits in the Andes region greatly affect the chemical composition of the Amazon tributaries and it is estimated that dissolved salt in these tributaries accounts for about half of the sodium and chlorine content in the Amazon River.

SALVE BUG (*Crustacea, Isopoda*). A marine crustacean, *Aega psora*, parasitic on various fishes. It is elongate oval in form and is a little more than one-half inch long. Found on both sides of the Atlantic.

SAMARIUM. Chemical element symbol Sm, at. no. 62, at. wt. 150.35, fifth in the Lanthanide Series in the periodic table, mp 1,073°C, bp 1,791°C, density 7.520 g/cm^3 (20°C). Elemental samarium has a rhombohedral crystal structure at 25°C. The pure metallic samarium is silver-gray in color, retaining a luster in dry air, but only moderately stable in moist air, with formation of an adherent oxide. When pure, the metal is soft and malleable, but must be worked and fabricated under an inert gas atmosphere. Finely divided samarium as well as chips from working are pyrophoric and ignite spontaneously in air, burning at 150–180°C. There are seven natural isotopes of samarium ^{144}Sm, ^{147}Sm through ^{150}Sm, ^{152}Sm, and ^{154}Sm. Eleven artificial isotopes have been identified. The natural ^{147}Sm isotope is weakly radioactive with a half-life of 2.5×10^{11} years. The samarium isotope mixture is the second highest (after gadolinium) of all elements in terms of its thermal-neutron-absorption cross-section (5,800 barns at 0.025 eV). The cross-section of ^{149}Sm is about 40,000 barns, but no chain reaction exists because of separation by low-cross-section isotopes. Samarium ranks 62nd in abundance of the elements in the earth's crust, exceeding tantalum, mercury, bismuth, and the precious metals, excepting silver. The element was first identified by Lecoq de Boisbaudran in 1879. Electronic configuration

$$1s^2 2s^2 2p^6 3s^2 3p^6 3d^{10} 4s^2 4p^6 4d^{10} 4f^5 5s^2 5p^6 5d^1 6s^2.$$

Ionic radius Sm^{2+} 1.11 Å, Sm^{3+} 0.964 Å. First ionization potential 5.6 eV; second 11.1 eV. Other important physical properties of sarmarium are given under **Rare-Earth Elements and Metals.**

The principal sources of samarium are monazite (4.5% Sm_2O_3) and bastnasite (0.5% Sm_2O_3). Current demands for the element are met by the coproduction with europium and gadolinium from these minerals. The residues of uranium mining (Canada) also contain about 4.5% Sm_2O_3. Unlike the other light rare-earth metals, the salts and oxide of samarium do not reduce to metal using barium, calcium, or lithium, nor can electrolytic processes be used. The most effective reducing agent is lanthanum, which is mixed with Sm_2O_3 and heated under vacuum in a tantalum crucible. The samarium metal volatilizes and is condensed as powder or sponge on coiled tantalum or copper condenser plates. Subsequently, the samarium must be remelted under an argon or inert atmosphere before it is cast into graphite molds.

Samarium has been alloyed with gadolinium and aluminum to produce nuclear reactor hardware that will absorb neutrons for short periods. The use of samarium in intermetallics, cermets, and other chemical forms for use in nuclear applications holds promise. Small quantities of Sm_2O_3 are used in optical-glass filters and to encase lanthanum borate glass rods which then are drawn into fine fibers for fiberoptics applications. The element has been used as a coding agent for inks used in data handling systems. Small amounts also have been used for activating phosphate-type phosphors. The addition of samarium oxide produces a strong narrow emission in the near-infrared spectral region. The most significant use of samarium is in the permanent-magnet alloy $SmCo_5$. The strength of these magnets is five times that of other previously developed magnetic materials. A new co-reduction process has reduced the cost of $SmCo_5$, making it competitive with other magnet alloys. See also **Cobalt.** Current and possible applications include electric motors, line printers, frictionless bearings, jewelry, and hospital surgical techniques.

See references listed at ends of entries on **Chemical Elements;** and **Rare-Earth Elements and Metals.**

Additional Reading

Sax, N. R., and R. J. Lewis, Sr.: "Dangerous Properties of Industrial Materials." 8th Edition, Van Nostrand Reinhold, New York, 1992.

Staff: "ASM Handbook—Properties and Selection: Nonferrous Alloys and Pure Metals," ASM International, Materials Park, Ohio, 1990.

Staff: "Handbook of Chemistry and Physics," 73rd Edition, CRC Press, Boca Raton, Florida, 1992–1993.

SAMPLE-AND-HOLD AMPLIFIER. Also known as a track-and-hold amplifier, this device has an output that is proportional to the input until a "hold" signal is received. Upon receipt of that signal, the amplifier output is maintained essentially constant even though there may be changes in the input signal. The input and output waveforms of a sample-and-hold amplifier are shown in Fig. 1.

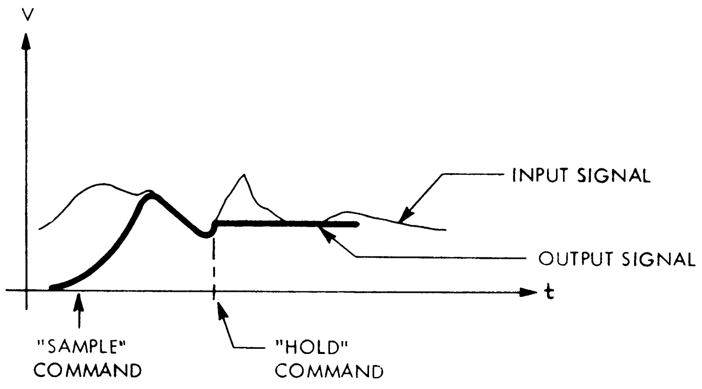


Fig. 1. Action of sample-and-hold amplifier.

As shown in Fig. 2, the conceptual design of a sample-and-hold amplifier comprises two independent amplifiers connected by a switch. With sampling switch S_1 closed, holding capacitor C is charged when an input signal is applied to the first amplifier. Upon receipt of the "hold" command, switch S_1 is opened, thus leaving capacitor C charged at the instantaneous value of the input signal. Capacitor C is not discharged because the second amplifier has a high input impedance. The output of the second amplifier remains essentially steady for a period of time. The "hold" signal may be generated by an external circuit (coupled to a process or experiment) or by a computer or digital control unit under control of a stored program.

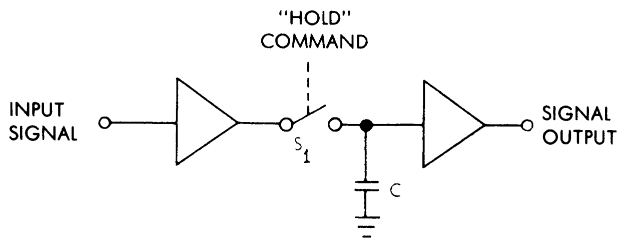


Fig. 2. Schematic circuit of a sample-and-hold amplifier.

Sample-and-hold amplifiers meet certain specialized needs in digital-data acquisition systems. Usually one or both of the following requirements exist: (1) the value of a single signal must be determined at a precise instant of time, or (2) the values of two signals must be compared at a precise instant of time. These requirements cannot be met by a time-shared amplifier and analog-to-digital converter system because these systems require a finite settling and conversion time. Where a sample-and-hold amplifier is used, it is possible to retain the instantaneous value of one or more input signals over the time interval required to convert them to digital values.

Thomas J. Harrison, International Business Machines Corporation, Boca Raton, Florida.

SAMPLE (Statistics). A sample is a collection of individuals drawn from a population. Ordinarily, inferences are to be made from the sample to the population, and the one must be in some way representative of the other.

When sampling from an actual population, the simplest method is to draw a random sample. The members of the population are numbered off, and the sample selected with the aid of a table of random numbers or some similar device. To ensure more even coverage, stratified sampling may be adopted. The population is divided into a number of homogeneous groups (the strata) and a random sample is selected from each. If only a random selection of the strata are sampled, we have a two-stage sample; three or more stage samples are similarly constructed. In two-phase sampling one type of observation is taken only on a small calibrating sample, while another (which may be easier or cheaper to obtain) is taken on a larger sample; regression or ratio relationships derived from the small sample are used to predict values of the first measurement for individuals in the large sample.

The introduction of a random element can ensure the absence of bias from an estimate based on a sample, and also makes possible a valid estimate of the error to which such an estimate is subject. (For various types of samples, and various methods of sampling, see **Sampling (Statistics).**

SAMPLING CONTROLLER. An automatic controller using intermittently observed values of a signal, such as the setpoint signal, the actuating error signal, or the signal representing the controlled variable, to effect control action. Also termed *sampled data system.*

Applications of sampled data systems include those applications in which the actuating or error information is only available in sample form, as well as those arrangements in which deliberate conversion of continuous data to sampled data is made. Examples of the first situation include an automatic tracking radar system and a gas chromatograph. The radar system scans in two coordinates and thus can furnish information on a particular target only at the discrete time intervals when the antenna direction permits radio frequency every to intercept the target. In the chromatograph, information on a particular variable is available at the completion of each sample analysis. An illustration of the second possibility arises in direct digital control. In this application, the digital computer is used to operate, simultaneously, numerous feedback control systems. By use of sampled data techniques the computer may be used periodically to process data from one system, then from the second, and so on until it completes the operating cycle and accepts data from the first system again. Figure 1 is a typical schematic for this type of system.

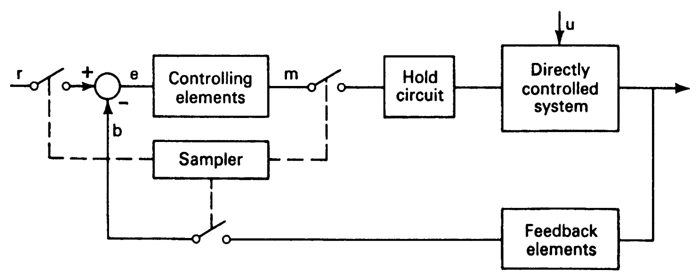


Fig. 1. Block diagram of sampled data control system.

The hold circuit shown in Fig. 1 is an integral part of any sampled data control system. Its function is to hold or remember the value of assigned variable at the sampling instant for a finite portion of the sampling period *T*. See Fig. 2. Its position in the loop will depend on the particular application. Wave form for a zero order hold following a

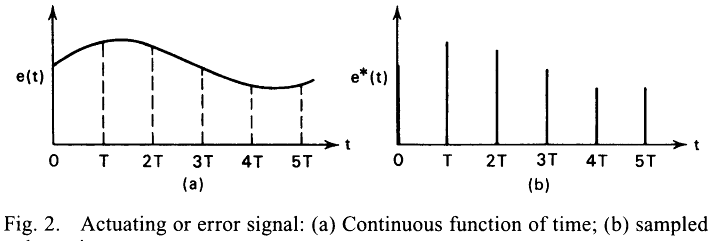


Fig. 2. Actuating or error signal: (a) Continuous function of time; (b) sampled pulse train.

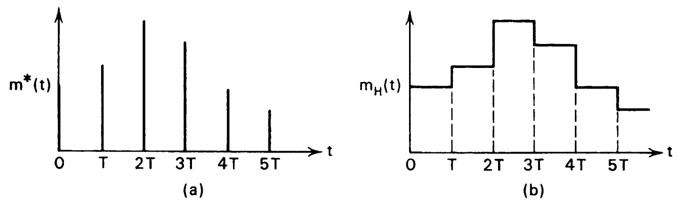


Fig. 3. Output waveforms of sample-and-hold circuit.

sampler is shown in Fig. 3. Although higher order holds are possible, they are seldom used in practice due to their complex nature.

See also **Control Action;** and **Sample-and-Hold Amplifier.**

SAMPLING (Statistics). The process of taking a sample. If repeated samples are drawn from a probability distribution and the value of some statistic is calculated for each sample, the resulting set of values will define a new probability distribution known as the *sampling distribution* of the statistic. The *sampling fraction* is the proportion of the total number of sampling-units in the population, stratum, or higher-stage unit within which simple random sampling (with multiple counting of sample-units when sampled with replacement) is made. There are thus sampling fractions corresponding to different strata and different stages of sampling. Exactly the same definition is sometimes (loosely) applied to other sampling schemes, e.g., in sampling with variable probability or *multi-stage sampling* (ratio of total number of ultimate units included in the sample to total units in the population). However, for other general application it appears desirable to define it as the reciprocal of the raising-factor of the sample when it exists, i.e., when the sample is self-weighing. A *uniform sampling fraction* is obtained when a sample is selected from a population which has been grouped into strata, in such a way that the number of units selected from each stratum is proportional to the number of units in that stratum. If from a stratified population a simple random sample is selected from each stratum in such a way that the proportion of units sampled in each stratum varies from stratum to stratum, the sample is said to be selected with *variable sampling fraction*. Applicability of the term to other sampling schemes rests upon the general definition of sampling fraction. There are many methods of sampling depending upon the procedure to be followed and the distribution of samples, which are determined in turn by the character of the population to be sampled.

Area Sampling is a method of sampling used when no complete frame of reference is available. The total area under investigation is divided into small sub-areas which are sampled at random or by some restricted random process. Each of the chosen sub-areas is then fully inspected and enumerated, and may form a frame for further sampling if desired. The term may also be used (but it is not to be recommended) as meaning the sampling of a domain to determine area, e.g., under a crop.

Bulk sampling is the sampling of materials which are available in bulk form, that is to say, it is the population which is in bulk; the term does not mean the drawing of a sample in bulk. Examples of such sampling would be the sampling of a shipment of coal for ash-content, or tobacco for moisture content.

Capture release sampling is a method of sampling specially suited to the estimation of the size of total populations of wild animals. It is also known as capture-recapture sampling. The method was practiced by Lincoln (1930) and involves capturing, marking and releasing a random sample, say, of animals of a particular kind. Subsequently, a further random sample is taken and the proportion of marked animals in this sample forms the basis of estimates of total population.

Cluster sampling. When the basic sampling unit in the population is to be found in groups or clusters (e.g., human beings in households) the sampling is sometimes carried out by selecting a sample of clusters and observing all the members of each selected cluster.

Direct sampling is a term used when the sample units are the actual members of the population and not, for instance, some kind of record relating to such numbers, such as census form, ticket, or registration card. The term relates to the directness of the observation of the units which enter into the sample, not to the process by which they are selected.

Sampling error is that part of the difference between a population value and an estimate thereof, derived from a random sample, which is due to the fact that only a sample of values is observed; as distinct from errors due to imperfect selection, bias in response or estimation, errors of observation and recording, etc. The totality of sampling errors in all possible samples of the same size generates the sampling distribution of the statistic which is being used to estimate the parent value.

Extensive sampling is a term used to denote sampling where the subject matter, or geographical coverage, of a sample is diffuse or widespread as opposed to intensive, where it is narrowed to a small field. Extensive sampling may refer either to a case where a wide variety of topics are covered superficially (rather than a few topics in detail) or a large area is surveyed broadly (rather than a small area studied in detail). The term could also be used with reference to time, that is to say, of sampling covering a long period.

It would be convenient to distinguish the cases as space-extensive, item-extensive and time-extensive, respectively.

Grid sampling is a form of cluster sampling, the clusters being individual areas of a grid and hence consisting of groups of basic cells arranged in some standard geometrical pattern. The term *configurational sampling* is also used in the same sense.

Indirect sampling is sampling from documents, or some record of the characteristics of a population, rather than the recording of information obtained at first hand from units of the population themselves. For example, it is becoming customary to obtain preliminary information on the results of, say, a national census by analyzing a sample of the census forms before the full analysis is undertaken; the population is then subject to indirect sampling. (See also Direct sampling.)

Intensive sampling. Like extensive sampling this expression may mean two different things: Either (a) sampling in a particular area with a dense scatter of sampling points, or (b) sampling wherein information on a restricted range of topics is sought by probing on them very deeply with an intricate schedule of questions. (See also Extensive sampling.)

Inverse sampling is a method of sampling which requires that drawings at random shall be continued until certain specified conditions dependent on the results of those drawings have been fulfilled, e.g., until a given number of individuals of specified type have emerged. In this sense it is allied to sequential sampling. The term is not a good one.

Lattice sampling is a method of sampling in which sub-strata are selected (for the sampling of individuals) according to some pattern analogous to the allocation of treatments on a lattice experimental design. For example, if there are two criteria of stratification, each p-fold, so that there are p^2 sub-strata, it is possible to choose p sub-strata so that none occurs in more than one "row" or "column" of the array representing the p^2 possible sub-strata; in short, in the manner of a Latin square. Similar schemes are possible for three-way, or more, classification. Various schemes of the lattice type are known under the name of "deep stratification."

Line sampling is a method of sampling in a geographical area. Lines are drawn across the area and all members of the population falling on the line, or intersected by it, are included in the sample. If the lines are straight parallels equally spaced across the area concerned, then the sampling becomes one form of systematic sampling. If, instead of all intercepts on the lines, a series of evenly spaced points are chosen on each line, the sampling is equivalent to choosing the points on a lattice and may also be regarded as two-stage line sampling.

Lottery sampling is a method of drawing random samples from a population by constructing a miniature of the population (e.g., by inscribing the particulars of each member on to a card) and drawing members at random from it (e.g., by shuffling the cards and dealing a set haphazardly). It is the method usually employed at a lottery—hence its name— but suffers from the disadvantage that the preparation of the cards entails considerable labor and strict precautions must be taken in the shuffling process to guard against bias.

Mixed sampling is where a sampling plan envisages the use of two or more basic methods of sampling. For example, in a multi-stage sample, if the sampling units at one stage are drawn at random and those at another by a systematic method, the whole process is "mixed."

Usage is not uniform, but where samples at one stage were drawn at random with replacement and at another stage were drawn at random without replacement, it would seem better not to describe the whole process as "mixed," the essential basic method of random selection being employed throughout.

Multi-phase sampling. It is sometimes convenient and economical to collect certain items of information from the whole of the units of a sample and other items of (usually more detailed) information from a sub-sample of the units constituting the original sample. This may be termed *two-phase sampling*, e.g., if the collection of information concerning variate, y, is relatively expensive, and there exists some other variate, x, correlated with it, which is relatively cheap to investigate, it may be profitable to carry out sampling in two phases. At the first phase, x is investigated, and the information thus obtained is used either (a) to stratify the population at the second phase, when y is investigated, or (b) as supplementary information at the second phase, a ratio or regression estimate being used. Two-phase sampling is sometimes called *double sampling*. Further phases may be added if desired. It may be noted, however, that multi-phase sampling does not necessarily imply the use of any relationships between the variates x and y. The expression is not to be confused with multi-stage sampling.

Probability sampling. Any method of selection of a sample based on the theory of probability; at any stage of the operation of selection the probability of any set of units being selected must be known. It is the only general method known which can provide a measure of precision of the estimate. Sometimes the term random sampling is used in the sense of probability sampling.

Proportional sampling is a method of selecting sample numbers from different strata so that the numbers chosen from the strata are proportional to the population numbers in those strata.

Quota sample is a sample (usually of human beings) in which each investigator is instructed to collect information from an assigned number of individuals (the quota) but the individuals are left to his personal choice. In practice, this choice is severely limited by "controls," e.g., he is instructed to secure certain numbers in assigned age-groups, equal numbers of the two sexes, certain numbers in particular social classes and so forth. Subject to these controls, which are designed to make the sample as representative as possible, he is not restricted to the contracting of assigned individuals as in most forms of probability sampling.

Representative sample. In the widest sense, a sample which is representative of a population. Some confusion arises according to whether "representative" is regarded as meaning "selected by some process which gives all samples an equal chance of appearing to represent the population"; or, alternatively, whether it means "typical in respect of certain characteristics, however chosen." On the whole, it seems best to confine the word "representative" to samples which turn out to be so, however chosen, rather than apply it to those chosen with the object of being representative.

Route sampling. A procedure similar to line sampling and used in surveys of crop acreage in districts which are well provided with roads. A route which adequately covers the area is chosen and the roadside lengths of the different crops recorded. Since the location of roads is unlikely to be random, estimates of acreage so obtained are likely to be biased, but changes in acreages may be estimated by using the same route for a number of years. The method of route sampling as a form of systematic sampling can also be applied to crop estimation.

Sampling with replacement. When a sampling unit is drawn from a finite population and is returned to that population, after its characteristic(s) have been recorded, before the next unit is drawn, the sampling is said to be "with replacement." In the contrary case, the sampling is "without replacement."

A different usage occurs in sample-surveys when samples are taken on successive occasions. If the same members are used for successive samples there is said to be no replacement; but if some members are retained and others are replaced by new individuals there is "partial replacement."

Unbiased sample. A sample drawn and recorded by a method which is free from bias. This implies not only freedom from bias in the method of selection (e.g., random sampling) but freedom from any bias of procedure, e.g., wrong definition, non-response, design of questions, interviewer bias, etc. An unbiased sample in these respects should be distinguished from unbiased estimating processes which may be employed upon the data.

See also **Quota Sample;** and **Random Selection.**

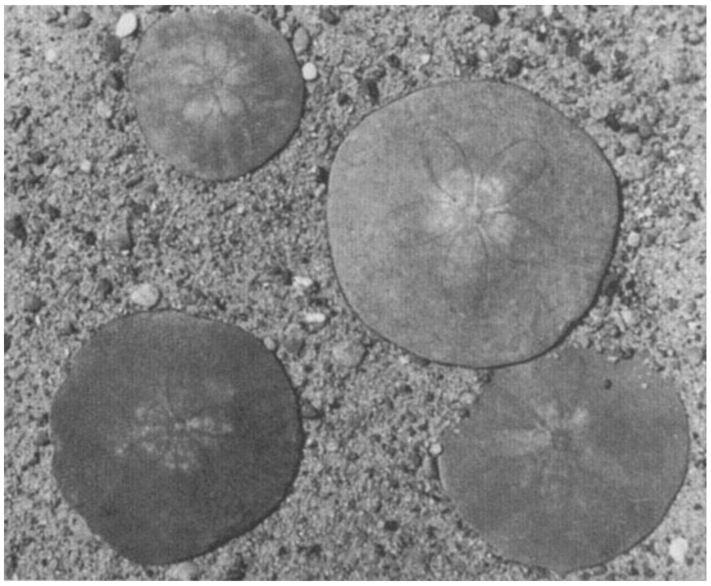
SAN ANDREAS FAULT. See **Earth Teconics and Earthquakes.**

SANDBOX TREE. See **Euphorbiaceae.**

SAND CRICKET (*Insecta, Orthoptera*). *Stenopelmatus.* Thick-bodied clumsy insects with large heads and long slender antennae. They live in loose soil, usually under some protective object, in the western United States. They are not true crickets but are more closely related to the long-horned grasshoppers.

SAND DAB. See **Flatfishes.**

SAND DOLLAR (*Echinodermata, Echinoidea*). A sea urchin with a very thin body, almost circular in outline and with a diameter of less than 3 inches (7.6 centimeters). It is common on both coasts of North America.



Sand dollar. (*A. M. Winchester*)

SAND FLY FEVER. See **Phlebotomus Fever.**

SAND GROUSE. See **Pigeons and Doves.**

SANDPIPER. See **Shorebirds and Gulls.**

SAND SHARKS. See **Sharks.**

SANDSTONE. Sand grains cemented by such substances as silica, carbonate of lime or iron oxide, so as to form a solid rock is called sandstone. It occurs usually in beds of varying thickness, depending upon the conditions under which the original sediments were laid down. Because it is normally well-jointed and easy to work, sandstone has been much used for building purposes. Unfortunately, however, as most sandstones are quite porous, the weathering action of the atmospheric agencies may have a very deleterious effect upon them.

SANDSTONE DIKE. Sandstone occurring in fissures which have been filled from above, or from beneath. The latter type are usually the result of earthquake fissures in great flood plain or delta deposits in which the sands have been injected from below.

SAPAJOUS. See **Monkeys and Baboons.**

SAPANWOOD. See **Brazilwood.**

SAPODILLA (*Achras zapota, Sapotaceae*). A large tree native to the forests of Central and tropical South America, the fruit of which is very desirable. The yellow-brown flesh is translucent and very sweet and wholesome. The greatest value of the tree is in its latex product, which yields chicle. The chicle-gathering industry is centered in Yucatán and Central America. The tapping is done in the rainy season. The tapper climbs to a height of 30 feet (9 meters), and with a machete cuts a series of connecting zig-zig diagonal gashes in the bark as he descends. At the bottom of this series of cuts he attaches a cup, into which the latex flows. The crude substance is collected, boiled down to eliminate much of its water and the coagulated product pressed into 20–25 pound (9-11.3 kilogram) blocks. This substance, chicle, varies in quality from the best grade, which is milk-white in color, to pinkish or darker grades, which have received less care in preparation. Each tree yields about $2\frac{1}{2}$ pounds (1.1 kilogram) of chicle during one season and may be tapped every 3–4 years. The blocks of chicle are shipped largely to the United States, where they are melted and cleaned, flavored and sweetened, and then marketed as the familiar chewing gum. This use of the latex of the Sapodilla is not new, since the Aztecs and their predecessors knew of it and used it. When first introduced into the United States it was tried as a rubber substitute, but proved unsuitable.

The record sapodilla tree growing in the United States is located in Miami, Florida. As compiled by the American Forestry Association, this specimen has a circumference (at $4\frac{1}{2}$ feet; 1.4 meter above ground level) of 99 in (252 cm), a height of 50 feet (15.2 meters), and a spread of 29 feet (8.8 meters).

SAPPHIRE. See **Corundum.**

SAPROPHYTES. These are plants which obtain their food from non-living organic material. Most of the saprophytes are fungi. Among the higher plants, a small number of flowering plants and perhaps a few mosses are also saprophytes. It is characteristic of these saprophytic plants that they have little or no chlorophyll, and so are not able to carry on photosynthesis. Their energy is derived from the complex organic substances which they absorb. In many instances the absorption of these substances is greatly advanced by the presence of mycorrhizae.

Especially is this the case with various species of saprophytic orchids which have mycorrhizae within the cells of the roots or rhizomes. The various species of Carol-roots (*Corallorrhiza*) are common saprophytic orchids of American woods. These orchids have no roots, absorption occurring in the much-branched fleshy rhizome which gives the plant its name. In this rhizome the mycorrhizae are found. These plants have erect stems, leaves reduced to scales, and no chlorophyll. Other saprophytic orchids occur in the continents of the Old World.

SARCOMA. A malignant tumor originating in connective tissue. These growths are composed of densely packed cells, diffusely imbedded in a homogeneous ground substance. Their degree of malignancy varies greatly. They spread by local infiltration and by blood stream invasion. The most frequent sites in which sarcomas develop are bone, lymph nodes, and subcutaneous tissue.

Sarcomas are much less common forms of malignant tumors than are carcinomas.

SARDINE. See **Herring.**

SARDONYX. See **Agate.**

SAROS. The fact that eclipses occur in periodic intervals was known to the ancient Chaldeans, and probably even in prehistoric times. This period of 18 years, $11\frac{1}{3}$ days ($10\frac{1}{3}$ days if there happen to be 5 leap years in the interval) is known as the Saros. If an eclipse should occur on January 1, 1977, at noon, another similar eclipse would occur on January 12, 1995, at eight o'clock in the evening. The eclipse would not occur at the same point on the earth but would be about 8 hours farther west in longitude.

During the course of a Saros there are about 29 lunar and 41 solar eclipses, each repeated during the next Saros, but not at the same portion of the earth. See also **Eclipse.**

SARSAPARILLA (*Smilax* sp.; *Liliaceae*). The genus *Smilax* contains some 200 species, most of which are tropical, though a few such as the carrion flower, *Smilax herbacea*, and the cat briar, *Smilax rotundifolia*, occur as far north as the New England states. The tropical species are mostly climbing shrubs or vines, usually with prickly stems. The leaves are entire and of oblong to ovate shape. At the base of the leaf is a pair of tendrils which are perhaps to be interpreted as modified stipules, though such structures are not usually found in monocotyledons. The flowers are small, dioecious and borne in umbels. The fruit is a berry. Some of the South American species are the source of Sarsaparilla, which is obtained from the dried roots.

SASSABY. See **Antelope.**

SASSAFRAS. See **Laurel Family.**

SATELLITE (Astronomy). The term as used in astronomy usually refers to small, planetlike objects that are revolving about the individual planets in orbits. The moon is the satellite of the earth and has been known from remotest antiquity. Satellites over the years have served a useful purpose to astronomers since the mass of a planet can be determined accurately only if the planet has a satellite. By application of the rigorous expression for the harmonic Keplerian laws of planetary motion the mass of any planet and satellite may be found in terms of the mass of the earth-moon system after the distance of the planet from the satellite and its period of revolution are known. The problem of the determination of the masses of the satellites themselves is a more difficult problem.

Further information regarding the different satellites will be found in **Moon (Earth's);** and **Planets and the Solar System.** See also entries on the specific planets.

SATELLITES (Communication and Navigation). In terms of positive impact on world civilization, the communication satellite must be ranked among the very highest of scientific and engineering inventions and achievements of the last half of the 20th century. As of 1994, *just one* of numerous modern communication satellites orbiting the earth can handle the equivalent of 120,000 telephone calls and three television channels at any one time. When first conceived, a given communication satellite was expected to enjoy but a relatively short useful life of about 2 to 3 years. With design and materials improvements, such satellites frequently are operable for considerably longer periods.

The world's first commercial communication satellite, the *Early Bird* designed by Hughes Aircraft Company, was launched in 1965 and became the first satellite to provide international telephone service for the International Telecommunications Satellite Organization (INTELSAT). The *Early Bird* weighed but 75 pounds (34.5 kg), with a height of only 2 feet (0.6 m), and could carry only 240 telephone calls or one TV channel at any one time. The satellite was designed for an 18-month useful life, but provided reliable service for nearly 4 years. The several *Intelsat* satellites operational as of the early 1990s are nearly four stories tall and weigh up to 3 tons or more.

Basics of Satellite Technology

A communication satellite is a signal relay in space and may be (1) *passive*, i.e., the satellite simply reflects an incoming signal back to some location on ground; or, as almost exclusively applied in modern telecommunication systems, (2) *active*, where the satellite receives, regenerates, processes, and retransmits incoming signals. The path from ground to satellite is the "up link." The path for the return signal is the "down link." The electrical characteristics of the up and down links differ.

Satellites also may be classified by the nature of their orbits. A *geostationary* satellite has a circular orbit that lies in the plane of the Earth's equator. The satellite moves about the Earth's polar axis in the same direction and with the same period (24 hours) as the Earth's rotation. The satellite remains above a fixed point on the Earth's equator. The geostationary satellite height for Earth (this would vary with different planets) is about 35,784 km (22,240 mi) above the equator. At this height, no internal force to overcome gravity for remaining in position is required. (The centrifugal force, i.e., the radial outward force of acceleration, equals the centripetal force, i.e., the radial inward force of gravitational attraction between the Earth and the satellite.)

A *synchronous satellite* is one that orbits about the center of the Earth in a circle in exact synchronism with the rotation of the Earth, although it is not necessarily in the equatorial plane. If it is in the equatorial plane, then it fully satisfies the prior definition of geostationary satellite. If it is not in that plane, the satellite will oscillate back and forth in a figure-eight pattern and, in fact, may never be directly over certain ground locations. A stationary orbit must be synchronous, but a synchronous orbit need not be stationary.

Because of small thrusters aboard a satellite, ground controllers can reposition a satellite in orbit. Generally, ground-based satellite operations centers will periodically issue commands to fire these thrusters, usually for very short intervals, to maintain a satellite in its assigned orbital slot. Much less frequently, a relatively major orbital location change may be needed, as in the case of moving a weather satellite to gain a better view of North America when a companion satellite failed; or to accommodate any major communication network redesigns.

Communication Satellite Chronology

Essentially a half-century ago, Dr. A.C. Clarke, who has become a world-renowned science fiction writer, stated in the February 1945 issue of *Wireless World,* "An 'artificial satellite' at the correct distance from the Earth would make one revolution every 24 hours; i.e., it would remain stationary above the same spot and would be within optical range of nearly half the Earth's surface. Three repeater stations, 120 degrees apart in the correct orbit, could give television and microwave coverage to the entire planet." See Fig. 1. This was twelve years prior to the successful launch of *Sputnik* by the former Soviet block (October 1957).

Generally, Dr. Clarke's writings were met with skepticism from the scientific community. However, the predictions were not fully brushed aside because by the late 1950s communications engineers were exhibiting growing concern over the rapidly increasing telecommunications load that for many decades depended on solid-conductor channels for transmitting and receiving overland and overseas messages. With the accelerated expansion of television, coaxial cables were becoming overloaded and remained costly. Radio communications were also becoming over-crowded. High-frequency radio waves (3-30 MHz) can progress from one point on the Earth's surface to another at any distance only by repeated reflections from the ionosphere to the ground. Such signals are interrupted or distorted by ionospheric disturbances. Radio waves at higher frequencies (30-3000 MHz) are not reflected by the ionosphere, and so are reliable only between stations in line of sight of each other, this distance being severely limited by the curvature of the Earth and numerous obstacles of the terrain. Thus, severe technologic and economic problems motivated consideration of satellite communications, with some credit given to *Sputnik* for its demonstration, even if simplistic, of the feasibility of space communication.

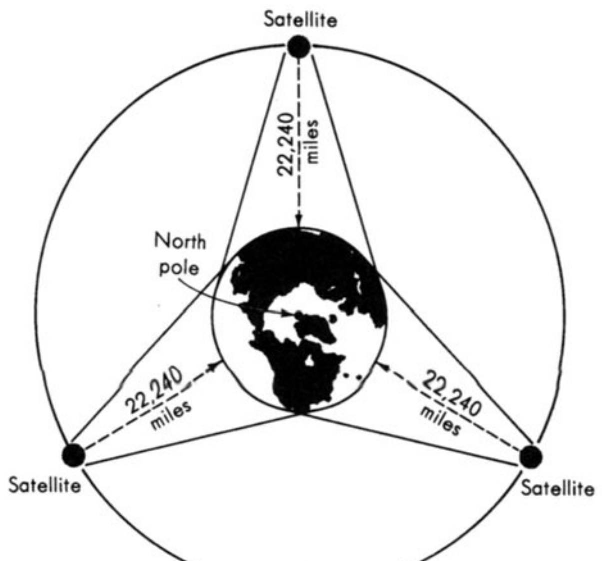


Fig. 1. Early proposed arrangement of how three communication satellites could cover the globe. Note: 22,240 mi = 35,784 km. The figure of 22,300 mi (35,881 km) usually is used.

By contrast with the general reactions, Dr. Clarke's[1] observations were considered seriously by Dr. Harold Rosen, who at the time was designing airborne radars at the Hughes Aircraft Company. Rosen convinced his associates to take the matter seriously. Within a few years, Rosen and Thomas Hudspeth and Donald Williams, also of Hughes Aircraft, proposed such a satellite to the National Aeronautics and Space Administration (NASA).

In October 1958, the National Aeronautics and Space Administration (NASA) in the United States launched a "catch up" space program. Following the demonstration of a privately funded prototype of a synchronous satellite, Hughes Aircraft Company received a NASA contract in 1961 for development of what was to become the *Syncom.* The first successful version of *Syncom* (*II*) was launched on July 26, 1963 and achieved a synchronous (not geostationary) orbit some 35,800 km (22,300 mi) above the Atlantic Ocean. (*Syncom I* was launched in February 1963, but the satellite blew up during the firing of the rocket's final stage.)

Syncom II proved that a communication satellite could guarantee around-the-clock access to ground stations located within the satellite's transmission beams. Because of this constant capability, live television coverage of events on the other side of the world were to become commonplace and the cost of overseas telephone conversations could be reduced.

The first communication through a synchronous satellite was between a U.S. Army crew at Lakehurst, New Jersey and a Navy crew aboard the U.S. Naval Service *Kingsport* (a satellite communication experiment ship) stationed in the Lagos, Nigeria harbor. Not so dramatic as the first telegraph message tapped in Washington, D.C. and received in Baltimore, Maryland ("What hath God wrought?") in 1844, or when Alexander Graham Bell conversed successfully by telephone on March 10, 1876, the first words carried by *Syncom II* simply were, "Kingsport, this is Lakehurst. How do you read us?" The real drama of the event was the clarity of the signal as compared with static-prone radio telephone links of that period.

Syncom II was the first communication satellite to transmit wire service coverage of a sports event (Gene Fullmer-Dick Tiger championship boxing bout in Nigeria). President John Kennedy and Nigerian Prime Minister Abubakar Balewa held the first two-way live telephone conversation between heads of state by way of *Syncom II.* In the first month of operation, ending October 18, 1963, *Syncom II*, hovering over Brazil, transmitted for 470 hours, an average of more than 15 hours per day. During that first year, *Syncom II* traveled more than 101 million km (63 million mi), obeyed more than 12,000 electronic commands, performed nearly 3000 communication experiments, and carried the voices of nearly 4000 people.

Because *Syncom II*'s orbit was inclined relative to the equator, the satellite did not remain stationary over a single spot on Earth, but described the expected figure-eight pattern. *Syncom III*, launched a year later by a somewhat more powerful rocket, achieved equatorial orbit and became the world's first geostationary satellite.

Meanwhile, in August 1962, the U.S. Congress passed the Communications Satellite Act, which established a national policy for the organization of a global satellite system in cooperation with other nations. This Act authorized private as well as governmental participation

[1]Dr. Clarke also is well known for his short story, "The Sentinel," which became the basis for Stanley Kubrick's film, "2001: A Space Odyssey."

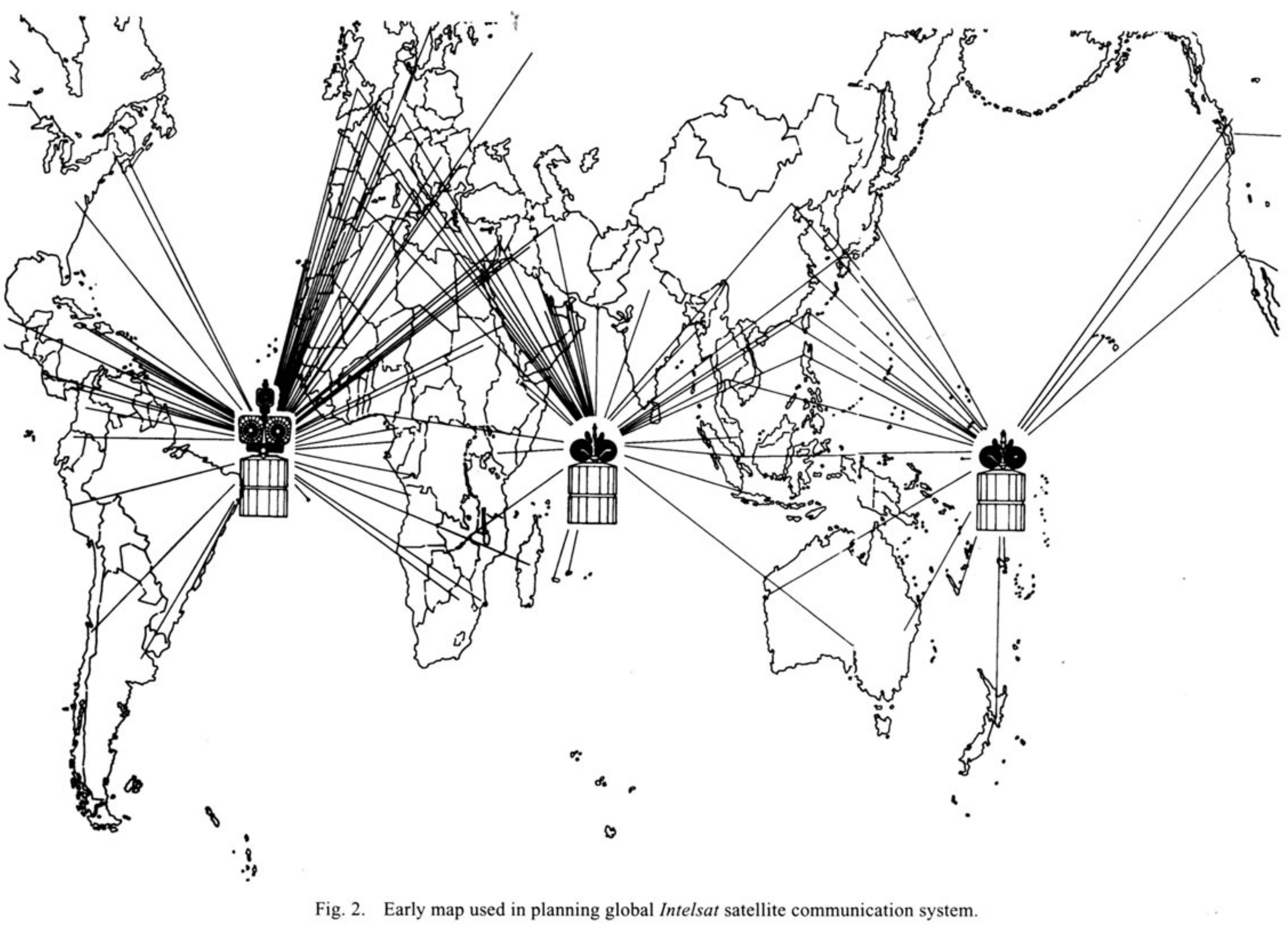


Fig. 2. Early map used in planning global *Intelsat* satellite communication system.

and control of the program. Many other governments joined in this participation in the manner of a cartel. In August 1984, eleven countries signed agreements to form a single global commercial communication satellite organization to become known as *Intelsat* (The International Telecommunications Satellite Organization). Over the years intervening, *Intelsat* has expanded to include 112 nations.

The details of finance and control pertaining to *Intelsat* (and similar satellite operating firms) are far too detailed to include here. Generally, the ground stations are owned and operated by entities within the countries in which they are located. To date, the earth sations handling the majority of satellite communications are the United States, the United Kingdom, Germany, France, Italy, and Japan. Marked increases in traffic also have been noted for ground stations in Brazil, Australia, Argentina, the Netherlands, Singapore, and Venezuela, among others. Although *Intelsat* remains very important to total worldwide satellite communications, in many countries private interests have made large inroads into participation of the field. An early map indicating how the world would be divided in terms of satellite communication is given in Fig. 2. It was realized almost from the outset, however, that heavy communication traffic would require many times more than three satellites.

As of the early 1980s, the problems associated with very heavy communication traffic emerged once again—this time the concern was with what may be termed "satellite saturation." Thus, in recent years, much greater attention has been given to assignment of channels and frequency domains. See Fig. 3.

As experience was gained with the early *Syncoms*, the need for a research program was manifested. Thus, the *Syncom* satellites led to the *Applications Technology Satellite* (*ATS*) program, which commenced its developmental stage in 1964. The ATS program was designed to flight test experimental payloads, investigate synchronous satellite operation in numerous modes, test antenna-pointing concepts that involved fixed conical beams pointed at the Earth rather than blanket coverage used by the *Syncoms*, and simultaneous communication with one satellite by several ground stations. Launched on December 6, 1966, *ATS 1* was sent on a three-year scientific mission for NASA, making history when it provided the first wide-angle pictures of the Earth's full disk, an illustration that was featured time and again by the video and print media.[2]

From a historical viewpoint, the progress made in communication speed is depicted in Fig. 4. Diagrams of second and third generation satellite system electronics are given in Figs. 5 through 8. Because communication satellites are designed for a comparatively long useful life (several years), very early advantage must be taken of improvements and innovations in electronics and communication technology. Even so, most satellites at any given time will lag technologically after insertion into orbit simply because alterations, as by retrieving and modifying

[2]Of considerable interest is the fact that the *ATS 1*, designed for a useful life of just a few years, remains in orbit and still may prove useful. NASA, which operated *ATS 1* for 19 years, turned command over to the public for nonprofit use in August 1985, a few months after the satellite ran out of positioning fuel. Depleted of onboard fuel, *ATS 1* drifts in a synchronous orbit unable to maintain a position over a fixed spot. Its single transponder, however, continues to provide that much esteemed link. Research vessels located off the coast of Africa have been linking up to *ATS 1* for approximately six hours of free communications (value estimated at $12,000/month) per day since April 1986. The University of Miami acquired signals from *ATS 1* when the satellite drifted eastward from its post over the Pacific basin to within sight of the University's Earth station. A favorite of its users, *ATS 1* is one of the few remaining satellites that operates in the VHF (very high frequency) range, making it possible to communicate via relatively simple antenna systems. Although its scientific instruments have long been shut down, *ATS 1* has been supplied with ample power to keep it alive through solar cells that coat its drum-shaped body.

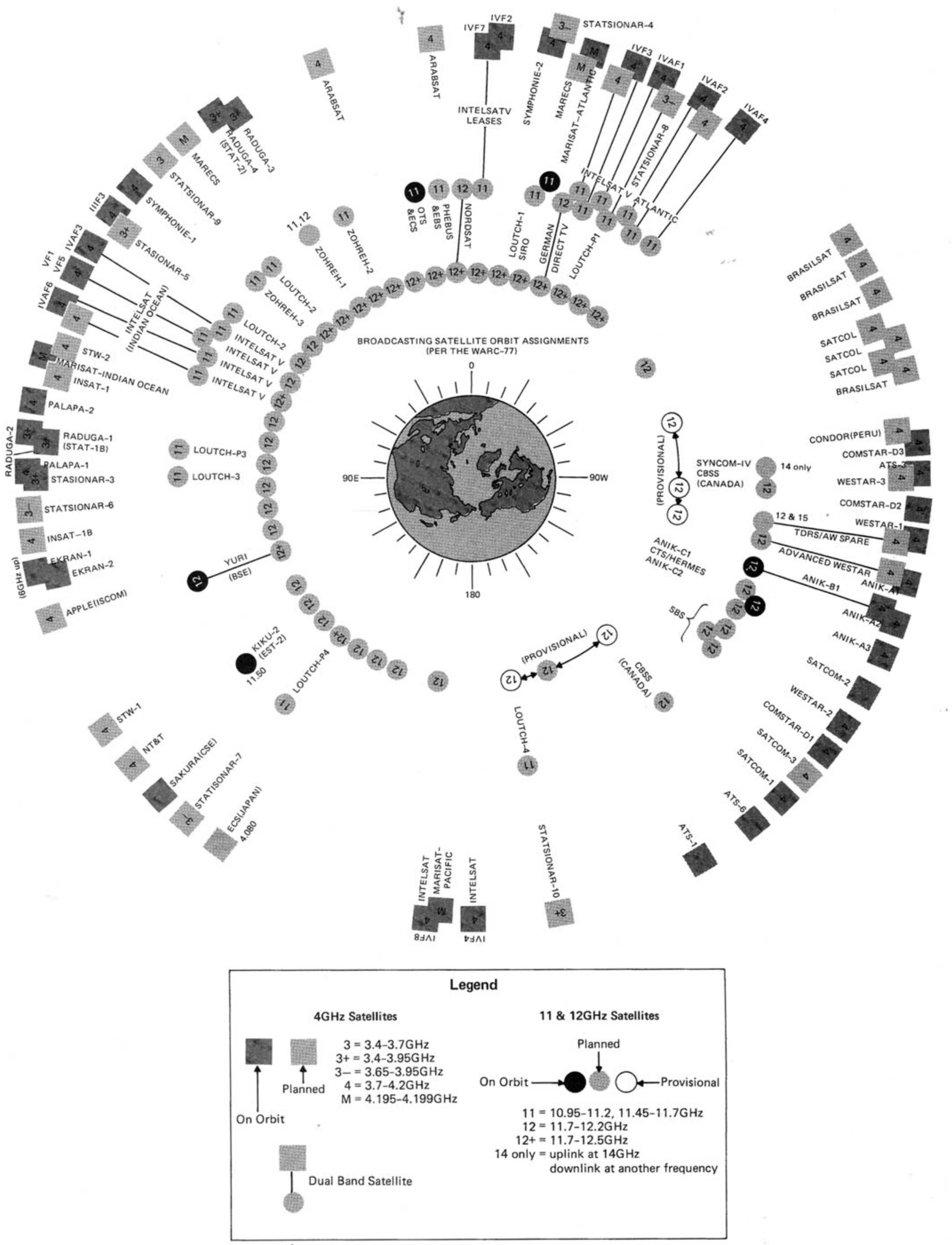


Fig. 3. As depicted by this diagram, concern over communication satellite saturation (overcrowding) of available orbit slots commenced with the proliferation of satellite placements in the late 1980s. (Adapted from an original color diagram prepared by Walter L. Morgan, Communications Center of Clarksburg, Maryland, for *Satellite Communications* magazine.)

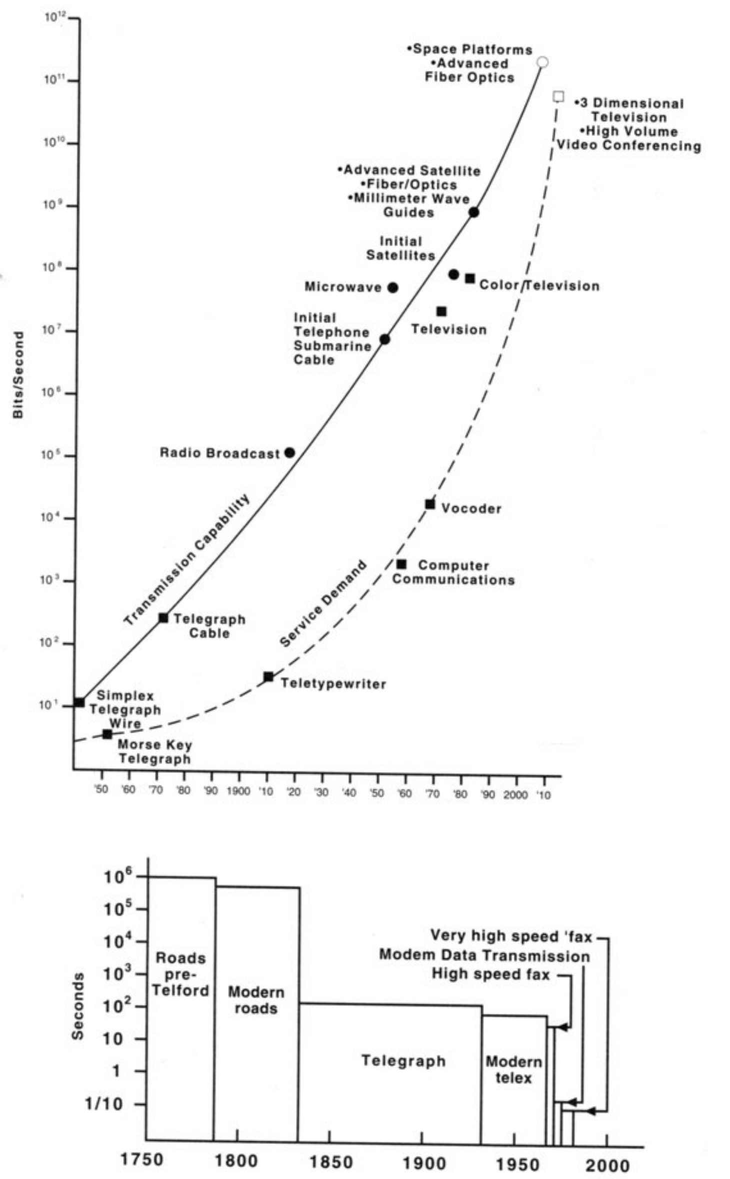


Fig. 4. Graphical depictions of telecommunications progress made over many years. (*Top*) From 1840 through the year 2000 (information rate in bits/second); (*bottom*) from 1750 to late 1990s (seconds required to transmit a single-page document). (*Charts prepared by J. N. Pelton for Satellite Communications magazine.*)

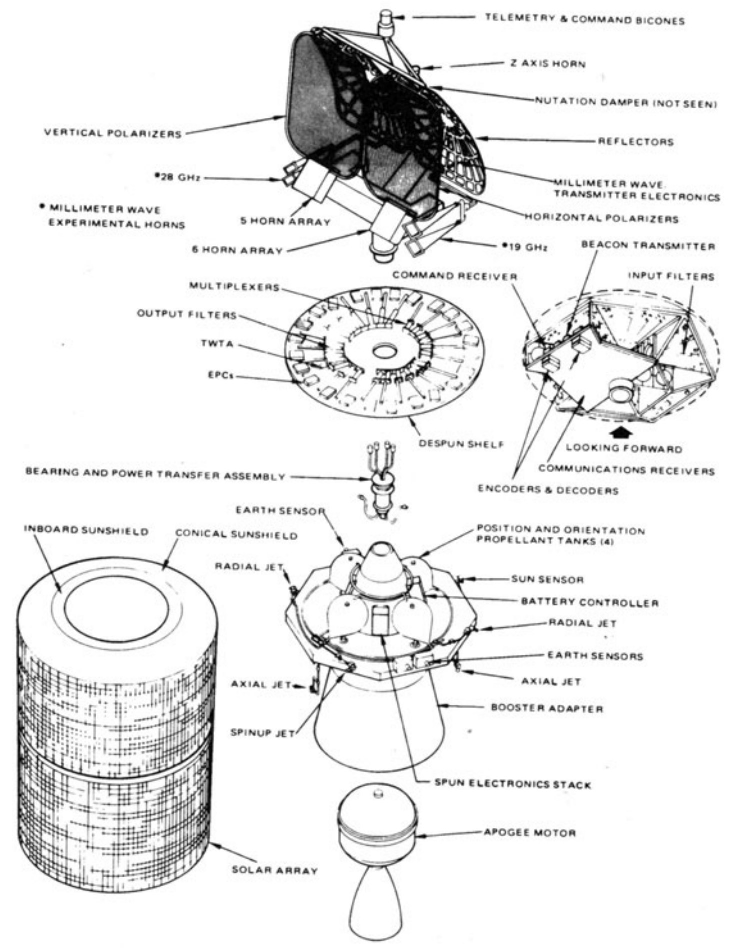


Fig. 5. Exploded view of communication satellite (circa late 1980s). The satellite utilizes a dual-spin design, with the spin axis perpendicular to the orbital plane.

The two main elements of the spacecraft are the spinning rotor and the despun earth-oriented platform containing the communications repeater and its antennas. The spinning rotor provides the basic gyroscopic stability to the spacecraft. The positioning and orientation (reaction control) subsystem is mounted on the rotor. Redundant jets are provided for orbital station-keeping, attitude control and spin-up with sufficient fuel (hydrazine) for the projected lifetime of the satellite. The directional communication antennas are pointed to the proper point on the earth by the despin control system. Redundant sensor information (three earth sensors, two sun sensors, all rotor-mounted) is used by an on-board processor to establish the inertial attitude of the spacecraft and control the antenna platform. The apogee motor is carried in the aft half of the rotor section. The burn of this motor provides the needed impulse to place the satellite in its final synchronous orbit.

The satellite temperature is controlled passively by selecting in the design phase the proper ratio of the solar energy absorptivity of the various external surfaces to their infrared sensitivity. Active heaters are also provided on certain temperature-sensitive elements of the satellite equipment. The power subsystem is located on the spinning rotor; included are two cylindrical solar arrays and two nickel-cadmium batteries. The batteries' capacity is sufficient to provide continuous service of all 24 transponders during eclipse seasons throughout the projected life span. Electric power for the satellite in orbit is provided by an array of several thousand solar cells mounted on the cylindrical body of the spacecraft.

The despun section of the satellite contains the communications repeater electronics and telemetry and command electronics. Spacecraft commanding is performed through two cross-strapped command systems. Telemetry information from the spacecraft is provided by redundant pulse-code modulated systems with a frequency modulated real-time mode available for transmission of certain data during spacecraft maneuvers. Other specifications: Satellite receive frequency band, 5925-6425 MHz; satellite transmit frequency band, 3700–4200 MHz; EIRP per transponder (at beam edge), 33.0 dbW (31 dBW for CONUS/Alaska combined); *G*/*T* (at beam edge), −9 dB/°K; usable rf bandwidth, 34 MHz per transducer.

through the use of a space shuttle, would have to be exceptionally important to defray the modification costs.

Satellite System Performance Parameters

In continuous attempts to improve the performance of satellite communication systems, designers pay particular attention to earth station equipment, path loss, frequency, noise, and networking concepts. The principal RF components of the earth station that contribute to link performance and station costs are antennas, the high-power amplifiers (HPAs), and the low-noise amplifiers (LNAs). These components affect the up-link radiated power transmitted by the earth station. This power, in dBW, is the sum of the HPA power output at the transmitting antenna and the antenna transmit gain. A trade-off study is usually required to identify the minimum-cost combination of HPA, LNA, and antenna size.

The distance traveled by the transmitted signal in a satellite link can range from 36,000 to 40,000 km (22,465 to 24,960 mi), which is a function of earth station location and satellite view angle. The path loss of any given signal is directly proportional to the square of *Rf*, where *R* is the path distance and *f* is the operating frequency.

The performance of a communication satellite is usually expressed by a transfer characteristic that relates the satellite power radiated (down-link EIRP—effective isotropic radiated power—toward the receiving ground station) to the flux density of the signal received at the satellite. The latter, in turn, is related to the radiated power, that is, the up-link EIRP of the transmitting ground station. The higher the satellite

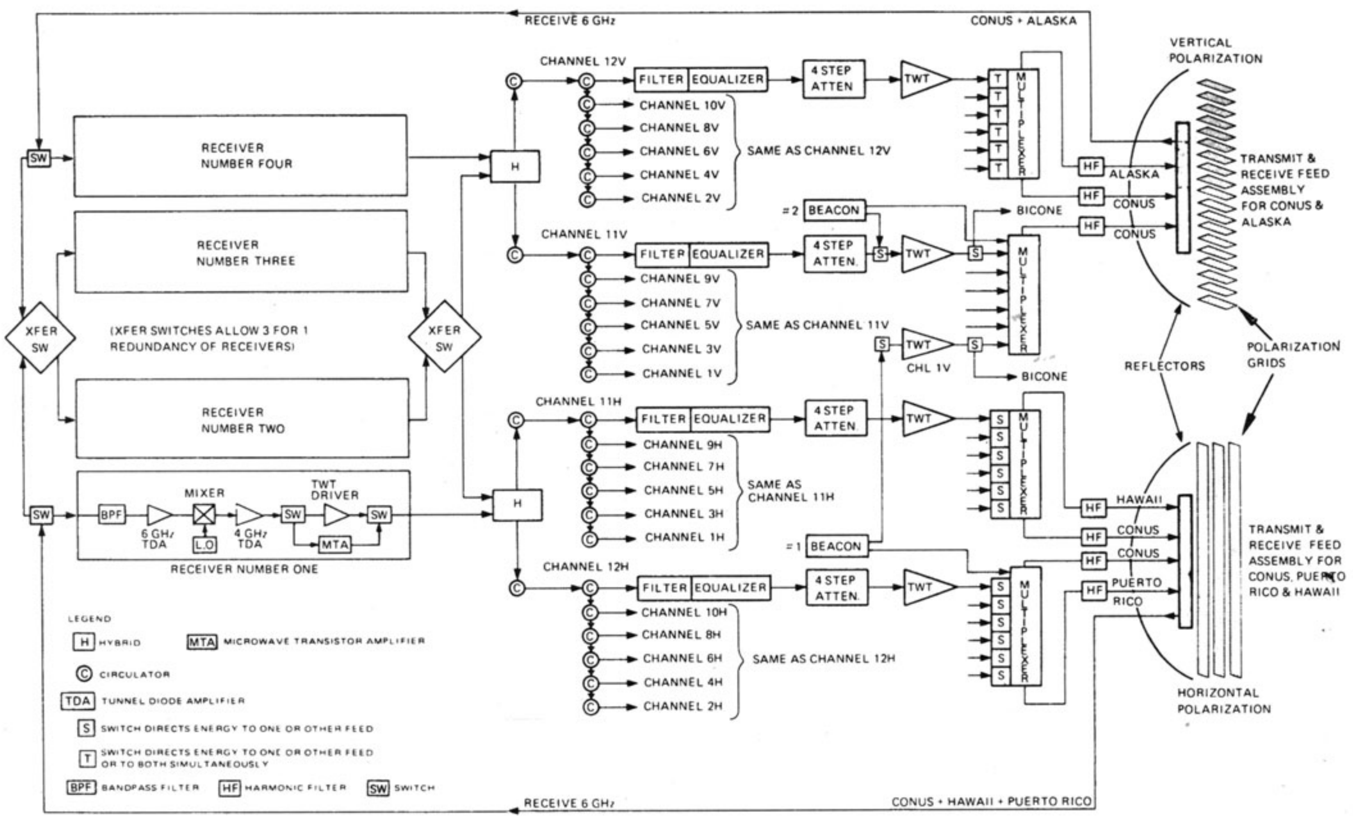


Fig. 6. Communication subsystem of satellite shown in Fig. 5. The satellite contains 24 transponders. The receiving portion of the subsystem consists of two primary receivers with two back-up receivers, each capable of replacing either of the primary receivers by ground command. Each receiver amplifies 12 channels and converts them to 4 GHz. The output of each receiver then passes to a 12-channel input multiplexer where each channel is separately filtered, amplified, and then fed to the output multiplexer, where the channels are combined in banks of six and fed to the transmit antennas.

transfer gain, the lower is the EIRP necessary from the transmitting earth station for the same down-link EIRP. The transponder nonlinearity characteristic also is an important parameter in determining the noise contributions from intermodulation caused by multiple carriers sharing a satellite transponder.

As pointed out by Sharma (see reference), the earth station receive G/T (antenna gain-to-noise temperature), in decibels, is the difference of the receive antenna gain and the system noise temperature, which depends on the noise temperature of the LNA. New K_u-band satellites are designed with higher satellite receive G/T, higher transponder EIRP, and larger transfer gains, as compared with C-band equipment. This allows smaller antenna sizes in the ground stations. Sharma further reports that for similar-sized antennas, the earth station antenna gain advantage at K_u-band is exactly offset by the increased path losses, but the narrower antenna beamwidths at the K_u-band provide greater interference discrimination.

A fundamental parameter in communication satellite systems is the ratio of carrier power to thermal noise density. Carrier power is defined as the power of the modulated rF carrier at the input of a receiver. Thermal noise is defined as that noise which occurs in all communication systems and the transmission media, resulting from random electron motion.

Networking. Because of a satellite's wide area of visibility, the several communications links that can access a given satellite require

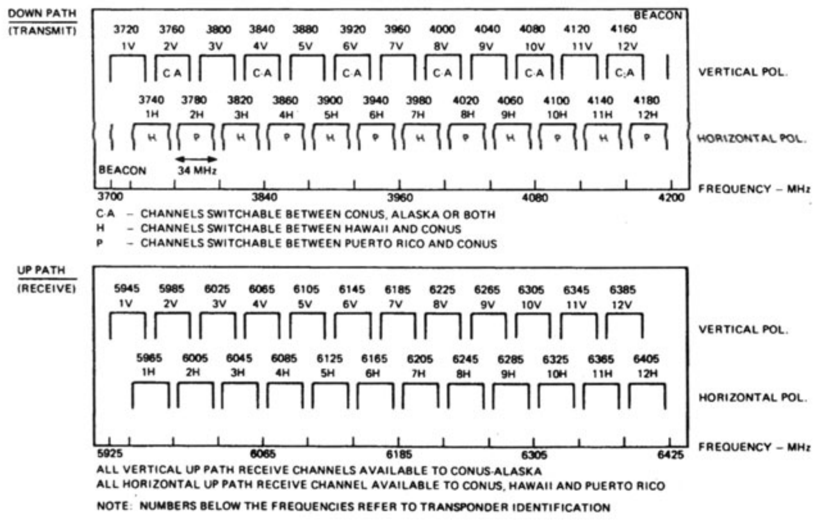


Fig. 7. Frequency plan of satellite shown in Fig. 5. To provide frequency reuse within the same satellite, one set of 12 channels is crosspolarized and frequency staggered by 20 Mhz to the other 12 frequency-sharing transponders. The isolation provided by the spacecraft antenna between the two polarizations is greater than 33 dB over the earth coverage areas of interest.

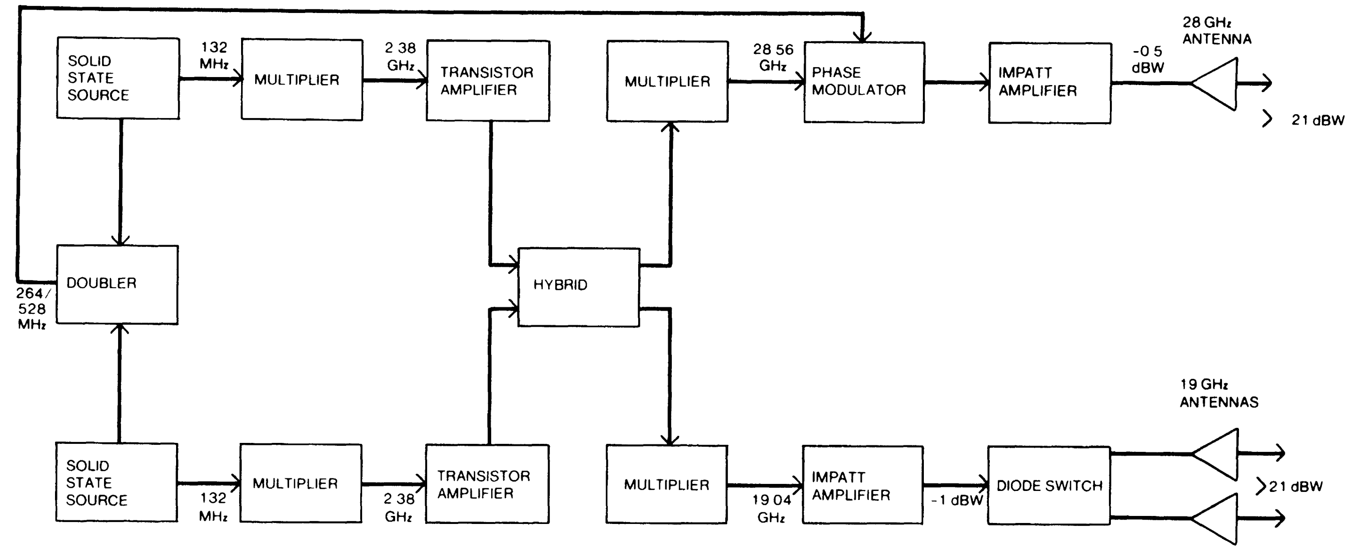


Fig. 8. The satellite shown in Fig. 5 also contains a $\frac{19}{28}$ GHz beacon package to provide super high radio frequency signal sources for use in gathering data on space-to-earth signal propagations as part of experiments by users of the system. Attenuation, depolarization, and phase coherence are measured at 19.0 and 28.6 GHz. These data are helpful in determining minimum power and other performance margins needed for improvement of satellite communication systems operating above 10 GHz.

separation, one from the other. Multiple access and capacity assignment methodologies fix the manner and efficiency that results when several earth stations share a common satellite repeater. Most commonly used in the past for interconnecting multiple earth stations simultaneously by way of a common satellite transponder have been frequency-division multiple access (FDMA) and time-division multiple access (TDMA). The implementation of FDMA is usually accomplished by one of two schemes: (1) single channel per carrier (SCPC/FDMA), or (2) multiple channels per carrier (MCPC/FDMA). In the case of SCPC/FDMA, multiple modulators/-demodulators (one for each unique up-link RF frequency) are required at the earth stations; also the bandwidth and power allocated to each SCPC carrier is proportional to the data rate of the port to which it is assigned. In the case of MCPC/FDMA, only one modulator for transmission is required for each earth station. However, several demodulators are needed for each frequency that must be received. The bandwidth and power of each MCPC carrier is proportional to the sum of the data rates from all ports of the earth station.

In terms of time-division multiple access (TDMA), Sharma explains that while TDMA earth stations need only one modulator/demodulator, operating at a high burst data rate, they do require additional control electronics for burst formatting and frame synchronization in order to properly position a burst within a TDMA frame. (For a large network, the lesser amount of modern hardware needed can favor the TDMA as compared with FDMA ground stations. This improved cost effectiveness, however, is not realized for smaller networks.) Provided that additional ports and additional nodes can be incorporated within the TDMA burst rate, the TDMA offers more flexibility than the FDMA. In TDMA systems, each ground station must have the capability of transmitting (and receiving with a mesh network) at a data rate proportional to the total network data rate regardless of its individual data rate requirements.

Capacity Assignment. Independent of access methods used, there are three main mechanisms for assigning satellite capacity: (1) *Fixed assignment* is most efficient when ground stations have a continuous stream of traffic at relatively high levels of use, as contrasted with bursty traffic of short duration (as may be encountered with interactive data or voice communication). (2) *Demand assignment* provides economies of scale for low-duty cycle traffic, where network resources are allocated only upon demand. These systems are more costly to implement than fixed assignment systems. Delay and contention also can be experienced in establishing links. (3) *Semivariable or dynamic assignment* (DAMA), as portrayed by Sharma, provides earth station sharing in a manner similar to a telephone Class 5 circuit-switching facility. Engineering of the earth station is based upon allowing a finite probability of traffic blocking (or message queuing) due to the unavailability of free channels at the ground station. Switching control is located at the ground station. In a VSAT-hub network, described later, this system can be used at the hub station to reduce the hardware required at the hub and the satellite capacity required for the hub-to-VSAT link. The dynamic assignment of satellite capacity is referred to as *fully variable DAMA*, and a parallel with a Class 1 switching office in a telephone network can be made.

VSAT Technology. As stressed by Parker (see reference), the early application of satellite communication sought to replace high-capacity, point-to-point terrestrial lines with a celestial route. This approach required costly earth stations and dedicated satellite capacity. The situation improved with the introduction of demand assignment strategies, coupled with technology which reduced the size of the earth station. Ground facilities, however, remained relatively expensive. In 1981, the introduction of the first personal earth station with a 2-foot (0.6 meter) dish for data reception introduced a promising era in satellite usage. After a comparatively slow start, considerable research is now being directed toward further development of the very small aperture earth terminal (VSAT). Manufacturers of VSATs currently are addressing three main issues—cost, size, and simplicity. To keep the costs low, the disk must be small and the electronics relatively simple. In the case of the VSAT, traditional discussions of optimizing the use of the satellite become less important. The optimal use of the satellite (least cost per bit) is attained by the larger, high-duty cycle earth stations, but experience has shown that these represent a small market. To keep VSAT prices attractively low and assure a market, it is evident that they will have to be built in high volume production. Parker observes that part of the cost/volume equation is recognition that this satellite communications approach represents a global marketplace. In fact, the ease of installation of a VSAT may prove even more valuable in the third world than in developed countries. VSAT could offer a rapid and economical way to introduce communication where none existed or where expansion of existing facilities is required. It should be remembered that, indeed, this was one of the goals of satellite communications in its early beginnings.

Beginnings of the Future

The largest and most complex communication satellite conceived to date (*Intelsat VI*) and mentioned briefly in the introduction to this article was launched in October 1989 from French Guiana on a *Ariane* booster. A second in the series was launched aboard a Martin Marietta Titan III rocket from Cape Canaveral on March 14, 1990. Due to a booster separation problem, the satellite was stranded in low Earth orbit about 250 miles (402 km) high, well below its intended orbit of 22,300 miles (35,881 km) The satellite continued to function well, but was useless in its very low orbit. After much deliberation, Hughes Aircraft

Company and NASA worked out a plan whereby the satellite would be picked up by the first flight of the space shuttle *Endeavour* when it became operational in 1992. The successful operation was accomplished on May 5, 1992. See Fig. 9.

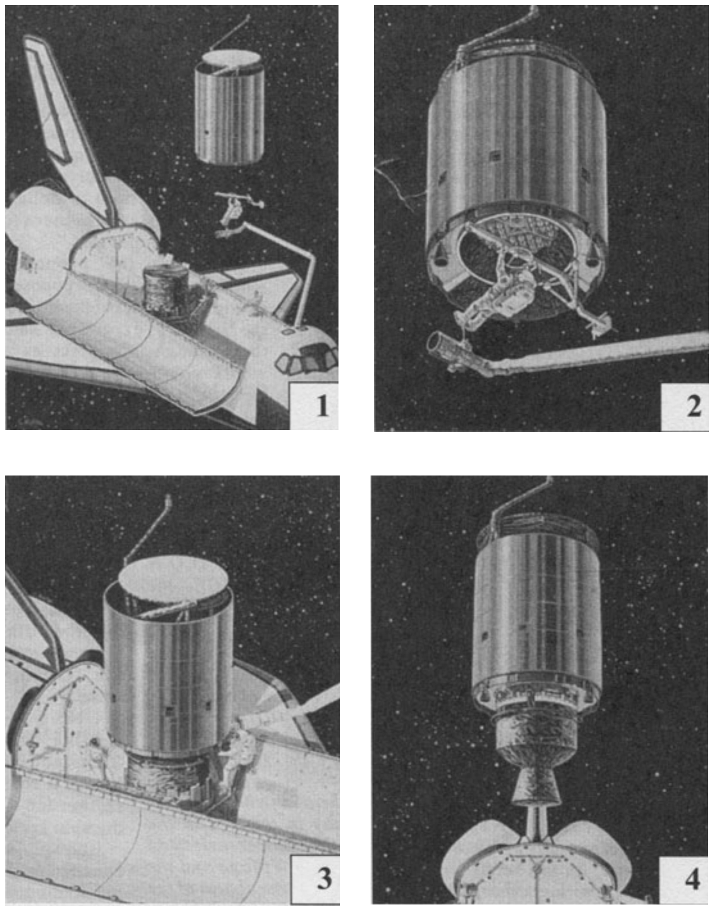


Fig. 9. Rescue of *Intelsat VI* by space shuttle *Endeavour*: (1)About six hours before *Endeavour*'s approach, the *Intelsat VI* was spun down to less than 0.65 rpm. *Endeavour* approached the satellite and stayed within the remote manipulator arm's grapple range. From the arm's platform, an astronaut installed the capture bar across the aft end of *Intelsat VI* and manually halted its slow rotation. In this artist's conception, the aft end of *Intelsat VI,* the new motor and capture hardware can be seen resting in the cradle in the space shuttle's cargo bay. (2) The remote manipulator system moves the astronaut in to "catch" the satellite, attach the capture bar, and bring it into the cargo bay. (3) Once the satellite is inside the space shuttle, astronauts attach the new boost motor, activate timers, and move into the airlock before the spacecraft is deployed. (4) Space shuttle crew members in the cabin fire the separation system to release the satellite. *Intelsat VI* controllers establish a command link with the satellite about 35 minutes later, after it has reached a safe distance from the *Endeavour*. (*Hughes Aircraft Company*.)

Upon completion of in-orbit testing, the satellite was positioned over the Atlantic Ocean region to provide telephone, television, and data communications between countries in North and South America, Europe, and Africa.

Using new technologies that will provide a wide variety of communication services to three ocean regions, the *Intelsat VI* is designed for maximum versatility. New features include in-orbit switching of three sets of zone beams and satellite-switched time division multiple access that will enable interconnection between the satellite's six primary beams. Considered the satellite of the 21st century, the 39-foot (12 meters) high satellite is being built for the International Telecommunications Satellite Corporation, a consortium of 112 member countries, under a contract calling for one prototype and five flight vehicles. The prime contractor is Hughes Aircraft Company (now a part of General Motors). In testing the prototype in late 1987, the satellite withstood vibration (35,000 pounds; 156 kN), 200,000 V, baking at high temperatures and freezing within a vacuum, conditions that are much more severe than those prevailing in orbit. Electrically, the satellite consists of a maze of more than 300 complex units, 400 switches, several thousand cables and waveguides, all connected by a wire harness containing more than 30,000 terminations.

The *HS 601*'s high-power capabilities enable satellite services, such as direct broadcast, mobile communications, and private business networks, through VSAT antenna dishes. It is predicted that these very small aperture terminals will be mounted on homes, automobiles, planes, and business sites for direct satellite access. Prior satellites have relied on spin-stabilization. Because of their operational simplicity, spinners are better suited for low and medium-power applications, such as domestic telephone and television services. But, in the area of power generation, the body-stabilized configuration is quite advantageous. As shown by Fig. 10, elongated solar-panel "wings" capture the sun's energy three times as effectively as the cylindrical panels on spin-stabilized satellites.

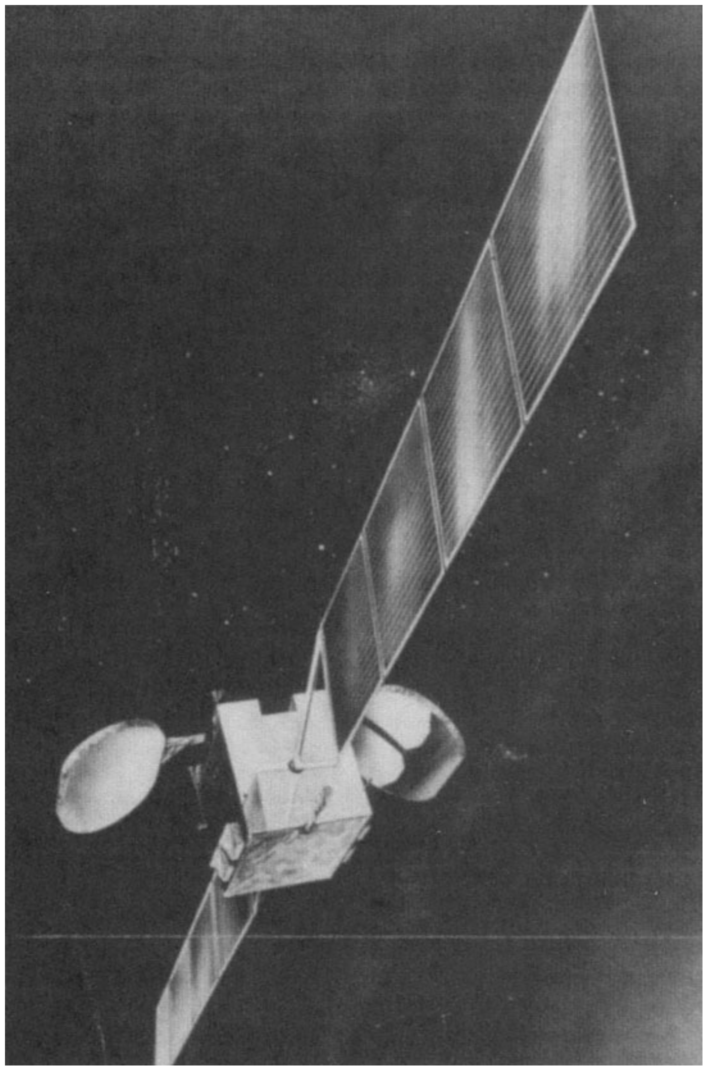


Fig. 10. Design of advanced, body-stabilized communication satellite features elongated solar "wing" panels for capturing solar energy for high-power space communications. Called the *HS 601*, the high-power capabilities of the the new satellite will enable satellite services, such as direct broadcast, mobile communications and private business networks, through very small aperture terminals (VSATs). (*HughesNews*, October 23, 1987.)

The *HS 601* is designed to handle up to 6000 W of power. The spinner's practical limit is considerably lower. If the *Space Shuttle* is not available, the *HS 601* can be launched from the U.S. rockets, *Titan* and *Atlas Centaur*, the European *Ariane*, and the *Long March*, built by the People's Republic of China. Stowed for launch, the satellite folds into an 8-foot (2.4-meter) cube. Depending upon the number of solar panels used, it can unfold to more than 100 feet (30+ meters) in its highest-power configuration. The *HS 601* has three basic nodules; (1) a bus section containing batteries and power control electronics; (2) a propulsion system that has four large tanks with propellant management devices, 12 bipropellant thrusters for orbit and attitude control, and a 110-pound (490 N) thruster for orbit injection; and (3) the payload module containing all elements of the communications payload.

What is beyond *Intelsat*? Currently called "Sky Cable," this project promises to usher in simultaneously the age of direct broadcast, high-definition television, and digital-quality video and audio — all services that can be received in homes by way of a flat, napkin-sized antenna placed, for example, on a windowsill. This program will require a minimum investment of over $1 billion. Among the attractive features for both rural and urban areas is the convenient, low-cost access to a greatly broadened range of programming services and sharply improved television reception. The most powerful space transmitter ever to be launched for commercial purposes, the "Sky Cable" will (1) direct broadcast up to 108 new television channels, (2) beam high-definition television (HDTV) as well as standard video signals, (3) deliver digital-quality audio and video, and (4) be received by inexpensive small antennas.

Direct broadcast allows for TV programming to be taken directly off a satellite by home viewers, with no intervening cable required. An experimental satellite (BSB-1), launched in late 1989, provided some direct broadcast TV services to the United Kingdom in 1990.

Companies worldwide are scrambling to perfect HDTV technology, which will deliver pristine, wide-screen images. Japan commenced experimental HDTV broadcasts in 1990 and 1991.

Many communications authorities consider the "Sky Cable" concept to be a major step forward in the 21st century for U.S. television viewers. For the first time, rural Americans will be afforded the oppourtunity to enjoy truly broad-based cable services like those available by cable in urban communities. The system will offer a variety of programming services, including multichannel packages, à la carte subscription channels, and pay-per-view programming, not to mention improved sound as well. It is estimated that the small antennas will cost about $300 at consumer electronic stores, a dramatic contrast to the conventional 10-foot backyard dish antennas that run from $2,000 upward. See also **Television.**

Design for Longer Life and Repair. Communications systems designers constantly are evaluating new materials and better component production means for lengthening the life of current, very-high-cost satellites. Also, with past successes, the philosophy of designing for "repair in space" is a factor. Rescue of *Intelsat VI* was described earlier. The earlier case of the *Leasat F3* rescue in 1985 was equally impressive. See Fig. 11.

Two other satellites were successfully retrieved by the *Space Shuttle.* The communication satellites *Palapa B2* and Westar VI were retrieved and brought back to Earth for repair during the summer of 1985.

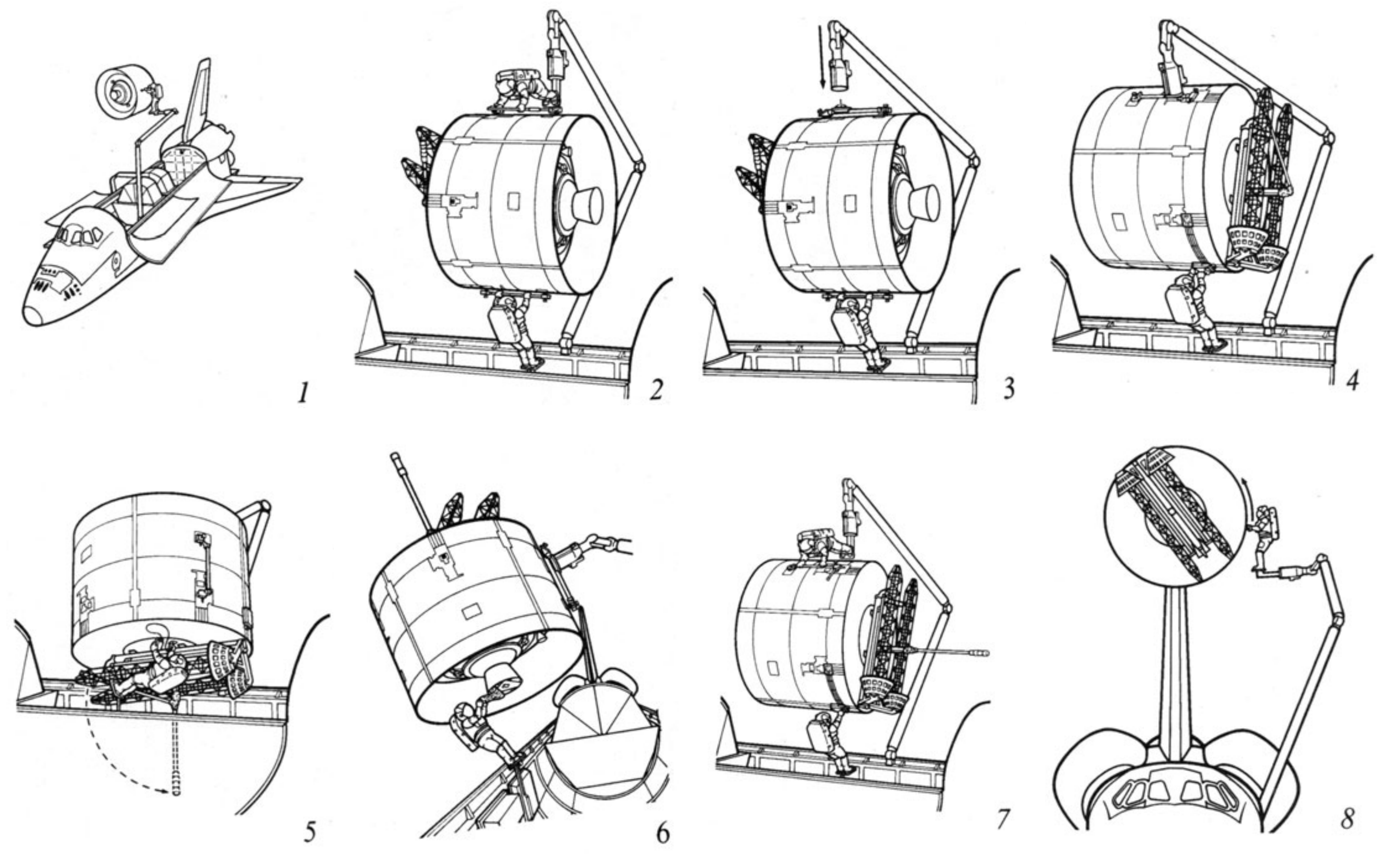


Fig. 11. Repair operations on communication satellite (*Leasat F3*) performed in space in 1985: (1) Astronaut hoisted into space by the *Space Shuttle's* remote manipulator arm, hooks the slowly spinning satellite and stops its spin by using a "capture" bar, one of four. (2) Astronaut holds the satellite over the shuttle bay while a second astronaut at bottom, attaches a special "handling" bar. (3) Astronaut uses the handling bar, at bottom, to hold onto the satellite while the shuttle's robotic arm is manipulated to grab the satellite. Another astronaut, not shown, has replaced the capture bar at the top of the satellite with the "grappling" bar. (4) The handling bar has been removed by astronaut at bottom, who is shown attaching one of two electrical bypass units that will enable ground command for the firing of *Leasat F3*'s solid-fuel rocket motor. Entry into the spacecraft's electronics is provided through a test access area into which plugs of the unit's harness are inserted. (5) Working on the despun portion of the satellite, which holds the antennas pointing stationary relative to Earth, astronaut is shown attaching the second remote unit that will enable deployment of the satellite's omnidirectional antenna and open command relays to and from Earth. (6) A special "greenhouse" plate that captures sunlight and prevents it from escaping is shown being attached by astronaut to the bottom of the satellite's solid-fuel rocket motor. Upon completion of the space operation, the satellite will be spun up by an astronaut so that its rocket motor points toward the sun. A small antenna on the plate will enable ground controllers to monitor the temperature of the motor. (7) The operation is completed. Astronaut at top has removed the grappling bar and attached a "spin-up" bar. Another astronaut at bottom turns on a timer in the electronics bypass unit that will delay ground command of the satellite for 13 hours. (8) With a flip of the arm, astronaut spins the satellite back into space. Long after the shuttle and its crew have departed the scene, ground controllers assumed command of the satellite. (*Hughesnews.*)

The first satellite with specific design features for in-orbit maintenance was the *Solar Max* observatory, launched on February 14, 1980. Ironically, after about ten months of service, three sealed fuels failed in the attitude-control module. Thus, the satellite could not orient itself precisely and, with it, the four instruments for observing solar maxima. The link-up (rendezvous) of the shuttle with the satellite proceeded smoothly. An astronaut using a manned maneuvering unit (MMU) or jet-propelled backpack docked with the satellite, but was not able to stop the satellite's slow spin. In fact, the effort caused the satellite to wobble so badly that the shuttle arm could not grab it. Controllers at the Goddard Space Flight Center, however, managed to employ the satellite's magnetic stabilizers (which interact with the Earth's magnetic field) to kill the unwanted motion. The astronauts then were able to proceed with the repairs and return *Solar Max* to its intended use. During the procedure, an electronics package associated with the satellite's coronagraph/polarimeter was also replaced, thus extending the satellites's useful life.

Communication Satellites and Optical Fiber Systems

A Massachusetts Institute of Technology team concentrating on a research program for communications policy has asked an interesting question: "How can we explain the puzzle of satellite proliferation when the growth of fiber networks could effectively serve the needs of virtually all of the industrialized world?" See Solomon and Anania reference listed.

The researchers point to the cartel that made international satellite communication viable in the first place and the various government policies, including those of the Federal Communications Commission (FCC) in the United States, that have fostered growth in the field. It is interesting to study the geographics of communication. Most major industrial complexes of the world are clustered at or near seaports, and only three nations—the United States, Germany, and Russia—have truly decentralized inland industrial structures. It is noted that the United States and Germany have been building comprehensive domestic fiber networks and that the former East Germany and Russia do not have customers for private satellite services. It is also noted that, in the remainder of the industrialized nations, a comparatively few populated urban areas generate the bulk of the traffic, with these areas easily served by terrestrial trunks. The progress in technology and the ability to manufacture fiber optic equipment in most of the advanced countries also can be cited. For example, the new fiber-optic cables (TAT-8 and 9 and PTAT-1) promise to triple existing capacity on undersea transatlantic routes. And it can be mentioned that these observations do not take into account those technologies that will compress voice and data to further enhance land and ocean fiber capacity.

The MIT team also mentions that world telecommunications systems are tending toward digital transmission and switching. These systems, linked by broadband fiber, will be found in major industrial cities throughout the world. Do these factors indicate that not only fiber optics, but even microwave digitization, may be a major threat to satellite networks? With virtually instantaneous acknowledgments, digital technology makes any terrestrial link more robust, and it is impossible to get instantaneous reverse channels on satellite links, digital or not. Therefore, satellite error-free data rates tend to be slower, although forward error-correction techniques may eventually solve the problem. Other arguments can be given to indicate that, for many situations, economics and technology could tend to favor fiber optics more and satellite communications less. The observations are sobering and are worth pondering by network planners.

Navigation and Maritime Satellites

The long-awaited *global positioning system* was tested and performed well during the Persian Gulf War (1990–1991). The fundamental principles are described in article on **Navigation.** The system sometimes is referred to as *space-age triangulation*. Because many prior systems remain in place and are operational as of the early 1990s, a chronology of the former navigational satellite technology is in order.

The International Maritime Satellite Organization (INMARSAT), comprised of more than 40 seafaring nations, provides worldwide satellite communication to the maritime and off-shore industries. Services include telephone, telex, and data transmission modes. Maritime satellite communications to and from the United States is provided by way of earth stations located in Connecticut and California. As of late 1986, INMARSAT commenced navigational as well as its long-established communications services. With this service, INMARSAT's operational and spare satellites can provide a civil navigation service in mid-latitudes. It also is possible to utilize INMARSAT to broadcast corrections for electronic navigation systems, such as Loran C, thereby improving their accuracy. When fully developed, INMARSAT satellites also can be used to provide an additional line of positions to the Global Positioning System.

Promises of civilian navigational systems, as exemplified by instrumentation in motor vehicles, have been forthcoming for several years. Some of these have been met in a partial way, and it appears that systems with outstanding accuracy and convenience may be available by the year 2000.

Radiodetermination Satellite Service. As reported by Rothblatt (see reference), the U.S. Federal Communications Commission implemented a new satellite radio service in July 1985, which is projected to have a large impact on navigation and personal communication technology. The FCC named this set of techniques the *Radiodetermination Satellite Service* (RDSS), in accordance with existing ITU definitions and allocated choice frequency bands in the lower microwave region to it. In June 1986, the FCC adopted particular technical and operational standards to govern use of the newly allocated RDSS spectrum.

RDSS is a set of radiocommunication and computational techniques that enables users to determine precisely their geographical position and to relay this and similar digital information to any other user. Because the techniques utilize geosynchronous satellites, the positioning and messaging capabilities are available over geographical areas of as much as the entire nonpolar surface area of the Earth. The majority of RDSS applications concern objects in motion. In the most general sense, RDSS is a mobile communication technology.

As detailed by Rothblatt, RDSS works by maintaining and processing a continuous flow of information between a system control center, geosynchronous satellites, and user transceiver terminals. The system control center transmits an interrogation signal (1618 MHz) many times per second through a geosynchronous satellite located 35,000 km above the equator to a coverage area on the surface of the Earth. The population of user transceivers receives the signal and individual transceivers transmit a response at 2492 MHz through at least two satellites if either (1) the user desires a position determination, or (2) wishes to send a message, or the signal contains information addressed to that transceiver.

Because of the varying distances between a user and each of the satellites and considering the constant speed (velocity of light) at which signals travel, the control center will receive identical responses at slightly different times for the signal response at an individual transceiver. A computer measures round-trip signal-transit time by comparing a stored replica of the transceiver's emitted signals, then measuring the associated time delay. This time delay, scaled by the velocity of light, is the measurement of the round-trip range from the system control center to the individual transceiver. The round-trip range measurement is converted to three-dimensional coordinates, such as latitude and longitude and, if a third satellite is utilized, altitude as well. Details of the system and its application for navigation and numerous other purposes are described in the Rothblatt paper.

For additional information on communication satellites, see **Telephony.** Weather (meteorological) satellites are described in the article on **Weather Technology.** Scientific and reconnaissance satellites are described in the article which follows. See also **Space Station and Space Shuttle;** and **Space Vehicle Guidance and Control.**

Additional Reading

Appenzeller, T.: "Hope for Magnetic Storm Warnings," *Sci. Amer.*, *Science*, 922 (February 21, 1992).

Baker, D. J.: "Toward a Global Ocean Observing System," *Oceanus*, 76 (Spring 1991).

Bierman, H.: "Personal Communications, and Motor Vehicle and Highway Automation Spark New Microwave Application," *Microwave J.*, 26 (August 1991).

Butler, R. E.: "The Role of the ITU in Global Telecommunications," *Telecommunications*, 33–36 (August 1987).

Campanella, S. J., and R. K. Garlow: "RF/Optical Interface Design for Optical Intersatellite Links," *Microwave J.*, 85 (October 1991).

Ellowitz, H. I.: "The Global Positioning System," *Microwave J.*, 24 (April 1992).
Frye, D. E., Owens, W. B., and J. R. Valdex: "Ocean Data Telemetry," *Oceanus*, 46 (April 1991).
Griebenow, A.: "VSAT Implementation from the Buyer's Perspective," *Telecommunications*, 41–42 (June 1987).
Luderer, G. W. R.: "Technological Forces and the Network of the 1990s," *Telecommunications*, 74–75 (August 1987).
Marcus, M. J.: "Satellite Security," *Telecommunications*, 61–66 (June 1987).
Massey, H., and M. O. Robins: "History of British Space Science," Cambridge University Press, New York, 1986.
Noreen, G. K.: "Mobile Satellite Communications for Consumers," *Microwave J.*, 24 (November 1991).
Potts, J.: "Satellites and ISDN (Integrated Services Digital Networks)," *Telecommunications*, 117–119 (October 1987).
Sergo, J. R., Jr.: "The Evolution of the All-Digital Network," *Telecommunications*, 52–54 (August 1987).
Sharma, R.: "Satellite Network Design Parameters and Trade-Off Analysis," *Telecommunications*, 36–38 (June 1987).
Solomon, R. J., and L. Anania: "Is There a Role for Satellites in a Fiber World?" *Telecommunications*, 32–34 (June 1987).
Stix, G.: "Trucking Companies Look Over Their Drivers' Shoulders," *Sci. Amer.*, 124 (October 1991).

SATELLITES (Scientific and Reconnaissance). Earth-orbiting satellites fall into three broad classifications: (1) commercially oriented missions, as represented by communication and navigation satellites described in the prior article; (2) scientifically dedicated satellites, such as those found in space astronomy, cosmology, and earth sciences pertaining to the atmosphere, oceans, and terrain of Earth; and (3) military-motivated missions. The latter are not described in this encyclopedia.

With the successful launch of *Sputnik I* on October 4, 1957 into Earth's atmosphere, the so-called *Space Age* was commenced by the former Soviet Bloc. During the early years of satellite technology, the orbiting vehicles were launched by rockets. Many years later, with perfection of the *space shuttle* concept, orbiters within the handling capacity of the shuttle have been launched from that space vehicle. Some payloads in connection with the *Mission to Planet Earth* program just getting underway in the early 1990s may be so large and heavy as to require going back to direct rocket launching. Power required for final orbit positioning or when moving a given satellite to a new location is accomplished by a power source (usually chemically propelled thrustors) that is built into the satellite by design and the control of which is by remote commands from ground stations on Earth.

The word *satellite* also is used for research vehicles that may be designed to orbit around a planet other than Earth. Examples include the *Viking 1* and *Viking 2 Orbiters*, designed to orbit the planet, as contrasted with the *Viking 1* and *Viking 2 Landers*, which actually set down on the surface of Mars. During the earlier *Apollo* lunar expedition, there was a lunar orbiter, used to keep in touch with crew on the surface of the moon and, most noteworthy, with the mission of seeing and taking pictures of the "blind" side of the moon for the first time. Future space probing expeditions will carry planet orbiting satellites, which ultimately may be manned for comparatively long periods. By extension, satellite can refer to *platforms* in space, an early example of which was *Skylab* and more recent examples of platforms or space stations, which have been under development by the United States and the former Soviet Bloc.

Satellite Sensing of Earth Resources

History will probably show that the earliest practical use of an Earth-orbiting satellite was simply that of viewing and studying Earth from the perspective that is attainable only from a position in space far beyond that which can be achieved by a land-based aircraft. In the beginning, the motivation was largely that of simple curiosity, but the value of Earth imagery from space, both scientifically and commercially, could be high. Possibly shared by the weather satellites in terms of historical importance was the the land satellite (*Landsat*) program. The *Landsat* program has persisted for over three decades, but as described here has had a very checkered record (i.e., in terms of its management and not its technological performance, because scientifically the program has enjoyed an excellent record). As of the early 1990s, remote sensing of numerous properties of Earth is indeed central to the *Mission to Planet Earth* program just getting underway, and that doubtless will extend well beyond the year 2000.

Early Remote Sensing Programs. Prior to the formal launching of the first earth resources technology satellites (initially called *ERTS*; later named *Landsat*), the results of earth surveys from high-altitude aircraft and balloons greatly aided scientists and conservationists in charting the features of the Earth. (It is interesting to note, that as of the early 1990s and for some time into the future, weather services in various countries and the National Center for Atmospheric Research, headquartered in Boulder, Colorado, make extensive use of aircraft and balloons.) Earth resources experiments were on the testing agenda of several of the earlier *Gemini* and *Apollo* manned spaceflights. Images were made of the Earth both in the visible light range and in the infrared (IR) and ultraviolet (UV) portions of the electromagnetic spectrum. The results were promising and indicated that such data, properly interpreted, could assist in a wide variety of programs—agricultural planning, charting the movement of sea life, monitoring concentrations of air and water pollutants, and furnishing more fundamental knowledge of the Earth in terms of geography, cartography, forestry, geology, and hydrology, among other topics of social and scientific interest.

The Landsat[1] Program

The original concept of an earth resources monitoring satellite is generally attributed to the U.S. Geological Survey during the 1960s, although scientists at the U.S. Department of Agriculture were also thinking along similar lines during that time frame. In 1968, the U.S. Department of the Interior, which oversees the Geological Survey, made a request for funds to build a new satellite which it called *EROS* (earth resources observation satellite). The request was delayed because, at that time, policy makers considered that any space project was in the sole realm of the National Aeronautics and Space Administration (NASA). After a short interval, NASA agreed to develop and launch the satellite, but with the proviso that it would not play a role in distributing the data the satellite would generate. (The National Aeronautics and Space Act passed by the U.S. Congress in 1958 stipulated that NASA was to be a research and development agency, not a vendor of services.) The Interior Department accepted the data handling responsibility. It was agreed that satellite data would be collected at the Goddard Space Flight Center in Greenbelt, Maryland and turned over to the Geological Survey. The latter organization subsequently established an *EROS* Data Center in Sioux Falls, South Dakota. With the successful launch and operation of the first satellite of this type, the Sioux Falls center operated effectively for a limited number of years.

ERTS-1 (Earth Resources Technology Satellite)

During the period that the first satellite was under development and construction under the aegis of NASA, the name of the spacecraft was changed from *EROS* to *ERTS*. The *ERTS 1* was built into the same spacecraft frame that had been used for the Nimbus weather satellites of that period. As shown by Figs. 1 and 2, the spacecraft resembled an ocean buoy, equipped with two rectangular photovoltaic "wings" to capture solar energy. *ERTS 1* was launched on July 23, 1971 into polar orbit, which took the satellite over any given spot on earth once every 18 days. Each day, the satellite made 14 revolutions of the earth (about one every 103 minutes). Ground coverage proceeded westward until global coverage was completed. An active attitude control subsystem maintained the satellite's observing system within ± 0.7 degree of the local vertical. The two solar panels provided 500 W of electrical energy. Altitude of the craft was 900 km. The payload included:

1. Three television cameras (return beam vidicons) for photographing the earth in color. This system failed shortly after launch and details of the system need not be given here.
2. A multispectral scanner (MSS) was included as a secondary system. The MSS had previously been tested at high altitude in an airplane, but had not experienced orbital flight. The MSS became the primary system and performed well beyond expectations. Details on the MSS are given later in this article.

[1]The name *Landsat* was coined to differentiate the land surveillance satellite from a *Comsat* (communication satellite).

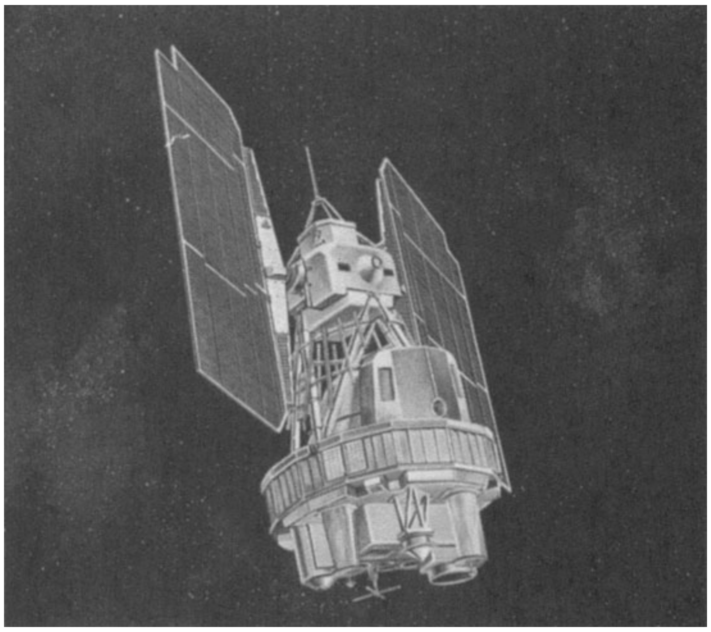
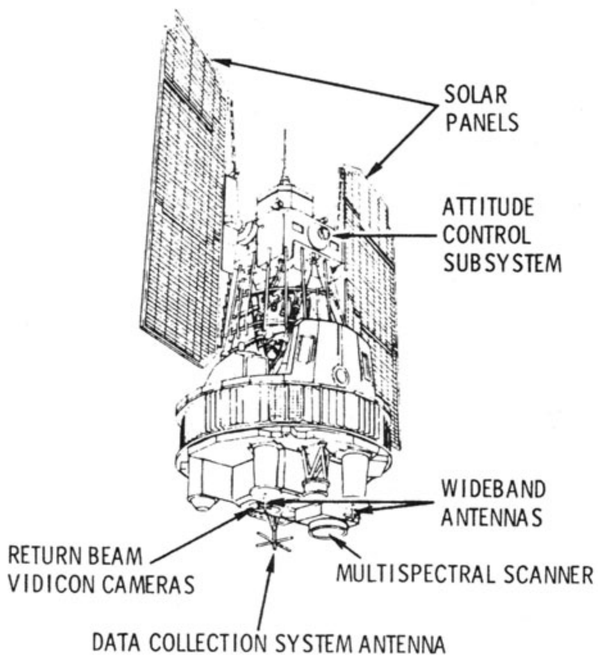


Fig. 1. The *ERTS 1* (later named *Landsat 1*) earth resources technology satellite as it appeared in orbit after launching in 1972. The satellite operated in a polar orbit about 914 kilometers (about 570 miles) above the Earth. The orbiting observatory returned images from a multispectral scanner. A data collection system gathered environmental information from Earth-based platforms and relayed the information to a ground processing facility.


SOLAR PANELS
ATTITUDE CONTROL SUBSYSTEM
WIDEBAND ANTENNAS
RETURN BEAM VIDICON CAMERAS
MULTISPECTRAL SCANNER
DATA COLLECTION SYSTEM ANTENNA
Fig. 2. Principal elements of the *Landsat 1*. Shortly after launching, the vidicon camera system failed. Thus, the satellite depended exclusively on the multispectral scanner (MSS), the results of which received wide scientific acclaim.

Augmenting the viewing data, data were received from automatic data collection platforms located in the United States and Canada concerning such local factors as stream flow, snow depth, and soil moisture (involving as many as eight sensors) and relayed by the satellite to ground stations. Data from any platform were available to users within 24 hours from the time the sensor measurements were made. Data from the MSS were transmitted to tracking stations via dual wideband (2.2 GHz) S-band data links for recording on magnetic tape. Narrowband telemetry provided for satellite housekeeping, the relay of data collection system data, and such payload-related data as attitude and timing. Telemetry, tracking, and command subsystems on board were compatible with tracking stations in NASA's spaceflight tracking and data network installed at that time.

The performance of *ERTS-1* was generally acclaimed successful and, as time permitted careful analysis of the data returned, considerable enthusiasm was shown by users of the information. Thousands of images were collected and cataloged. The rather weird colored views of features of the Earth were frequently seen in both the scientific and lay literature. Most everyone agreed that the satellite surveillance program had gotten off to a good start. Among the early accomplishments of the satellite were:

Geology. Nearly 50% additional unmapped faults found branching off the San Andreas fault in California, previously thought to be an accurately mapped seismologically active area.
Hydrology. Shallow substrate water-bearing rocks detected in Nebraska, Illinois, and New York, polluted-water drift charged off the coast of New Jersey; extent of Mississippi River flooding determined.
Cartography. Inaccuracies in map locations of Brazilian lakes and rivers discovered; unmapped lakes located in Iran.
Oceanography. Uncharted reefs detected off Bahama Islands and Australian coast; navigation charts for Bahamas to depths of 4 to 8 meters (13 to 26 feet) produced more accurately than any prior maps.
Agriculture. Area coverage of Texas and Oklahoma categorized as to range and pasture land, forests and water; up to 11 types of crops identified for fields 20 acres (7 hectares) or larger for San Joaquin Valley and Imperial Valley, California.
Urban Development. Dallas-Fort Worth images showed new roads, reservoirs, suburbs, and airports not on maps made three years prior. Essentially the first year of data gathering by the satellite was devoted by scientists in a variety of fields in learning how to use the data effectively, of literally coping with the immensity of information that can be returned from the satellite and in planning how to make the second satellite.

A second satellite, *ERTS B*, was launched in January 1975, but shortly thereafter was renamed *Landsat 2*. (The original *ERTS 1* was renamed *Landsat 1*.) A third satellite, *Landsat 3*, was launched in March 1978. Design attention was concentrated on improving the imaging system. By 1982, only one of the satellites (*Landsat 3*) remained active, but the first of a second generation of surveillance satellites (*Landsat 4*, renamed *Landsat D*) was launched that same year. The new satellite incorporated an improved multispectral scanner and a thematic mapper. See Figs. 3 and 4.

Politics of Landsat—1980—1993. Plans had been announced in the late 1970s for three additional *Landsats* (D', D'', and D''') to round out the second generation of vehicles. In the very early 1980s and particularly after a change in Administration in 1981, interest in the program by Congress and the Administration waned, although a number of strong supporters remained. The use of data by local and state governments, other governments worldwide, and by private groups did not measure up to earlier expectations. Also, there had been a series of technical difficulties and disappointments. Much of the imagery of the world from space had been accomplished as witness the present frequent use of images made by the earlier Landsats. Thus, with exception of very specific projects, data already on hand served many of the needs. Financial support for the program as derived from the sale of data did not appear to be forthcoming. (For example, the U.S. Department of Agriculture announced a large cut in its support of the program, one reason being that *Landsat* images of a given location, being 18 days apart, were too infrequent to be of help, particularly at the height of the growing season when an insect infestation or a hot, dry wind can kill a crop within days.) But the principal deterrent to the program was the Administration's plan to turn the program over to the private sector. Nevertheless, *Landsat* D', which had been ready for launch for a considerable period, was indeed launched in March 1984. This satellite carried an advanced design of the thematic mapper, although some scientists considered this "old technology" and would have preferred a multilinear array (MLA) that would utilize thousands of silicon charge-coupled devices with the promise of much improved resolution. (The

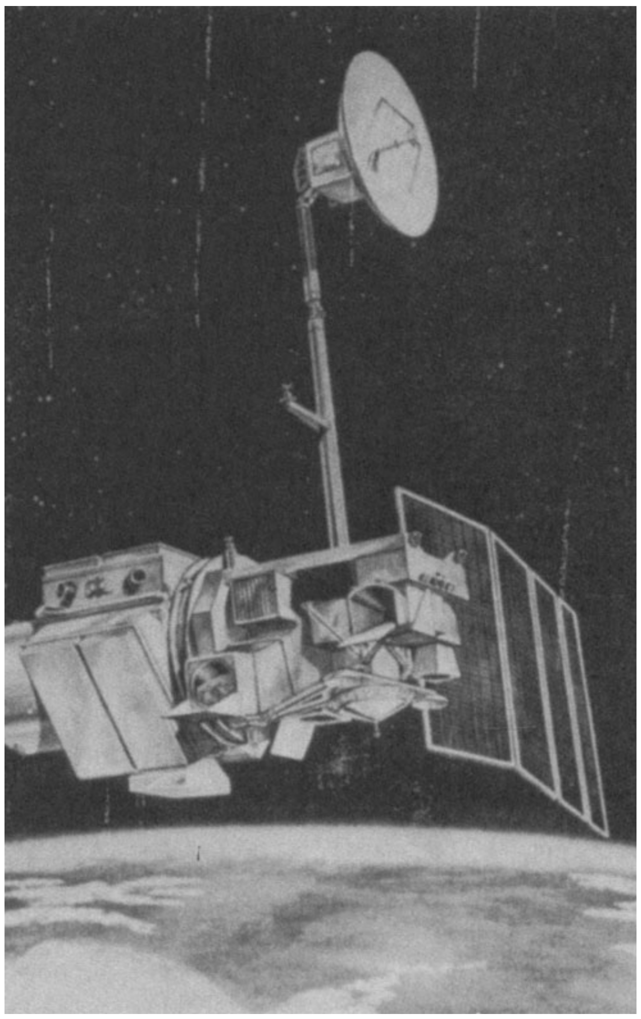


Fig. 3. View of *Landsat D* in orbit (artist's sketch). The satellite was launched into a sun-synchronous, polar orbit approximately 700 kilometers (435 miles) high in 1982. The satellite carried two remote sensing instruments as shown in Fig. 4. The satellite is sometimes referred to as *Landsat 4*. (*RCA Astro-Space Division, General Electric.*)

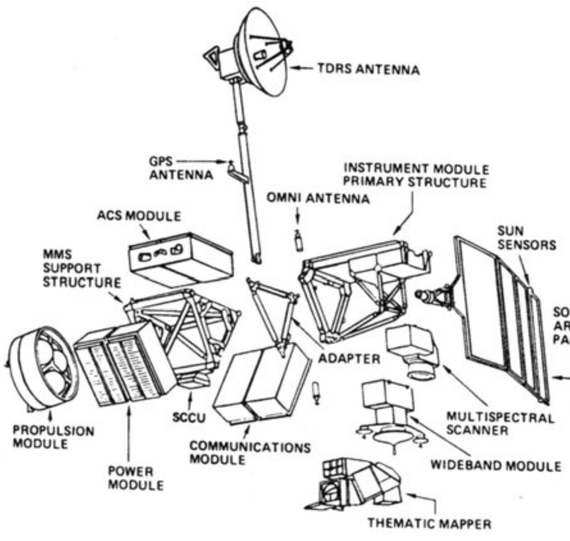


Fig. 4. Expanded view of *Landsat D* (or *4*). The satellite carried a thematic mapper (TM), an experimental sensor of advanced design to provide scenes with 30-meter (about 100 feet) resolution. It is a 7-channel radiometer. Satellite also carried a multispectral scanner (MSS) with 4 channels, 80-meter (262.5-foot) resolution, which is identical to the sensors on *Landsats 1* and *2*. *Landsat D* had the capacity to generate 800 scenes per day (550 MMS; 250 TM) for all ground stations, as compared with 190 MSS scenes from *Landsats 1, 2,* and *3*. (*National Aeronautics and Space Administration.*)

French, for example, as early as 1982 had announced that their SPOT surveillance satellite would feature the MLA technology.)

In 1983, the Administration formally put the Landsat Program up for bid by the private sector. COMSAT (Communications Satellite Corporation) was expected to be one of the bidders. (COMSAT's earlier proposal to take over Landsat and the weather satellites was considered a key step toward commercialization of the satellites). However, COMSAT declined, indicating that it was overextended with other projects. Of seven bids received, the competition was reduced to three firms on technical grounds. These were Space America Corporation, a startup firm in Bethesda, Maryland; EOSAT, a joint venture of RCA and Hughes Aircraft; and a partnership of Kodak and Fairchild. A subsidy from the government to aid operating costs for a period of 5 years was an important part of the negotiations.

Not until early 1986 was an agreement finally reached—with EOSAT (Earth Observation Satellite Company). Hughes announced in February 1986, that EOSAT would operate *Landsats 4* and *5* (the latter previously referred to as *D'*) and would be the primary source for marketing, ordering, and distribution of data from the spacecraft. EOSAT would commence construction of the next-generation commercial Landsat spacecraft to be called *Landsat 6* and *7*. The latter would be built by RCA and would carry imaging devices built by Hughes.

But, in early 1987, EOSAT announced that it had terminated all work on the *Landsat 6* and *7* vehicles. A part of the agreement with EOSAT was a $250 million federal subsidy, spread over six years. However, because of budget pressures, the first installment of $69.5 million was omitted from the 1987 fiscal year budget. As of early 1987, EOSAT announced that some accommodation probably can be reached.

As of mid-1990, Hughes Aircraft announced that testing had commenced on the *Landsat 6* mapper sensor. The spacecraft will orbit Earth at the same altitude, inclination, and equatorial crossing time as *Landsats 4* and *5*. The enhanced thematic mapper sensor will provide improved spatial resolution capable of discerning objects smaller than a tennis court.

French Surveillance Satellite *SPOT.* In preparation for several years, the remote sensing satellite (*Spot 1*) was launched from Kourou, French Guiana aboard the European Space Agency's Ariane rocket on February 21, 1986. The launch of *Spot* dramatized the continuing interest in space remote sensing. The venture is sponsored by CNES, consisting of the French national space agency and 18 public and private institutions in France, Belgium, and Sweden. The images cover four spectral bands in the visible to near-infrared wavelength region, and they offer a spatial resolution of 10 meters by 10 meters on the ground. Early images taken over the Atlas Mountains in northern Algeria clearly showed the tortuous geology of the area. False colors highlighted the various rock types and agricultural activities along the major wadis and streams. Images taken by *SPOT 1* are marketed on a worldwide basis.

Land Surveillance Satellite Instrumentation[2]

Initially, photographs were processed as black-and-white photos, one for each spectral band, with its own information. Later, color composites were prepared by combining processing of black-and-white bands through color filters onto color film. Frequently, they were processed to have the characteristics of infrared photography, a system with which interpreters were familiar because of earlier and some continuing use of infrared photography in conventional aerial missions.

Because surfaces viewed are frequently of high opacity and have radiation-scattering tendencies, data are generally obtained from the upper micrometer or millimeter layers or zones of the materials. Opacity in the visible and infrared portions of the spectrum results from high absorption coefficients of materials. The presence of water in many of the materials viewed restricts penetration in the microwave range because of high conductivity and dielectric constant. However, by considering changes in surface temperature as induced by diurnal solar heating, a trained analyst sometimes can infer certain additional prop-

[2]Although infrared and radar techniques have been introduced to Earth imagery since the earlier days of the *Landsat,* it is helpful to review the earlier optical technology when reviewing the applications of satellite imagery. Improvements in instrumentation simply have improved resolution by a wide margin and accelerated data-processing time.

erties of the imaged materials. But even this technique, where applicable, limits information to a substance depth of ten centimeters or less.

Wavelengths used extend from ultraviolet (0.4 micrometer) to microwaves at 50 centimeters, representing a wavelength range of 10^6. Ultraviolet below 0.4 micrometer is not usually used because of high atmospheric absorption and Rayleigh scattering. Infrared (1 to 3 micrometers) is best suited for determining the composition of observed minerals and rocks. Spectral reflectance of vegetation in the visible and near-infrared differs markedly from that of rock and soils. This spectral region contains intense chlorophyll absorption and the region of high reflectance beyond 0.7 micrometer. Thus, even the small presence of vegetation can alter the spectral signature of rocks and soils. It is interesting to note, however, that the subtle variations in the spectrum of stressed vegetation may effectively outline mineralied areas. Trained analysts also can interpret the distribution of areas of vegetation in terms of soil composition. The mid-infrared region beyond 8 micrometers is valuable for geological mapping because spectral emittance variation can be a basis for discriminating between silicate and nonsilicate rocks. The range from 8 to 14 micrometers is excellent for differentiating types of silicates.

As contrasted with traditional photographic methods, the *Landsat* multispectral scanner (an optomechanical scanner) permits the creation of digital multispectral images at wavelengths outside the sensitivity of film. The *Landsat* MSS provides multiple, spatially registered data sets, each taken at different wavelengths. See Fig. 5. The advantages of digital over photographic processing is that transformations applied to the data are not subject to the problems sometimes associated with the processing of film. Digital processing methodologies used include: (1) contrast enhancement; (2) spatial filtering to embrace morphological or structural information; and (3) arithmetic operations, such as rationing of spectral bands to enhance spectral reflectance differences and to suppress systematic effects, such as topography. Statistical analysis is used to reduce dimensionality in data sets containing many spectral bands or other variables. Sequential rationing of the MSS bands results in black and white ratio images that can be optically combined to produce a color-ratio composite.

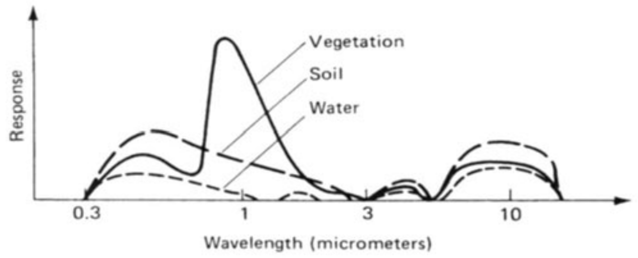


Fig. 5. Light and heat reflected or emitted from the Earth are picked up in selected wavelengths or bands by *Landsat* remote sensing instruments. Each band from the visible to the far-infrared contains specific information about the Earth's surface. In essence, the sensors are "tuned" to the wavelengths of vegetation, soil, water, geologic, and other surface materials to collect the spectral data each material reflects. (*National Aeronautics and Space Administration.*)

Each image displays 34,253 square kilometers (13,225 square miles) with uniform illumination and the scene composition can be optimized through digital processing of the radiance values.

During the earlier phases of the *Landsat* program, large numbers of *Landsat* views of the world were made available, not only to scientists in specialized fields, but also for use by science teachers. *Landsat* views have been very valuable to the teaching of geology, environmental topics, geography, and social and urban studies. A standardized coloring system was developed for these views and the analysis of the views became a challenge to scholars. The color scheme used is:

Red-Magenta—denotes vigorous vegetation, i.e., forests, farmlamds (agriculture) near peak growth. Detailed study of different shades of red aid in classification of different types of vegetation as well as the determination of their growth cycle and vigor.

Pink—depicts less densely vegetated areas and/or vegetation in an early stage of growth. Suburban areas around major cities with the green lawns and scattered trees normally show up pink.

White—areas of little or no vegetation but of maximum reflectance. White areas include snow, clouds, deserts, salt flats, ground scarring, fallow fields, and sandy beaches.

Dark Blue to Black—normally identifies water, i.e., rivers, streams, lakes, reservoirs. In the western United States, particularly Idaho and New Mexico, the very dense black features are ancient lava flows rather than water bodies.

Gray to Steel Blue-Gray—indicates towns, cities, urban, populated areas. The colors are produced by the returns from asphalt, concrete, and other anthropogenic features.

Light Blue in Water Areas—denotes either very shallow water or heavily sedimented water.

Light Blue in Areas of Western United (*and similar global areas*)—identifies desert shrubland capable of supporting limited grazing activities.

Brown, Tan, and Gold—denotes areas comprised primarily of open woodlands (piñon, juniper, aspen, chaparral) and rangeland suitable for grazing.

Although shown here in black and white, Fig. 6 is a *Landsat* view of essentially an agricultural area; the principal object of the view is the Salton Sea. In this reproduction, areas that appear red on the original *Landsat* view are black or nearly black.

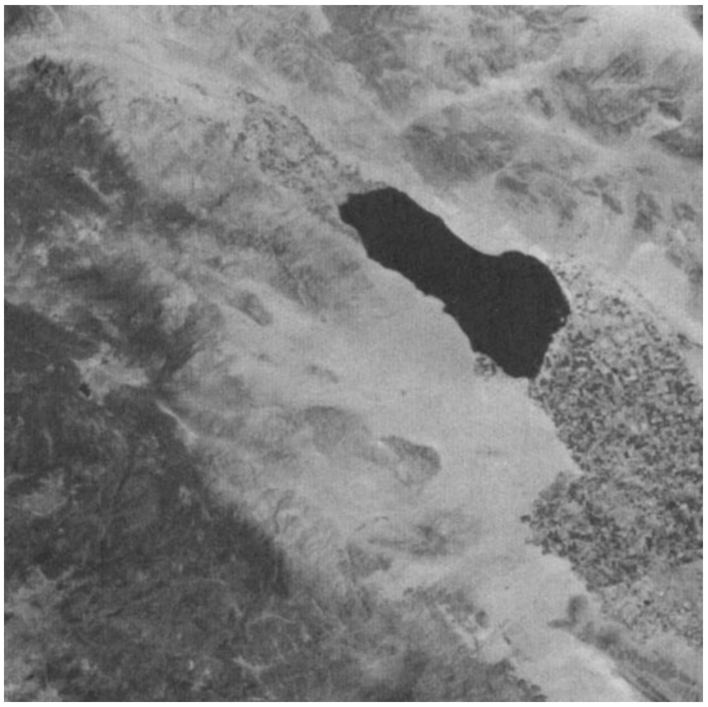


Fig. 6. Imperial Valley, California. The checkerboard pattern on both the northwestern and southeastern margins of the large body of water (Salton Sea) represents irrigated agricultural land in the Coachella Valley. Investigators using images, such as this (they are normally colored), have inventoried more than 25 separate crops. Each crop leaves a "spectral signature." If a number of fields of the same crop are sensed at the same time, the analyst need only make positive identification of one field bearing each crop to identify all crops in the image. (*National Aeronautics and Space Administration.*)

As an aid to researchers as well as scholars, grids in the form of transparencies that can be placed over a Landsat image are very helpful in measuring the area of various features. See Fig. 7.

Color for Color's Sake. In contrast with processing spectral color information as just described, over many years the Russian manned orbiting observers also have stressed the importance of color perception by the astronauts. As pointed out by Vasyutin and Tishchenko (refer-

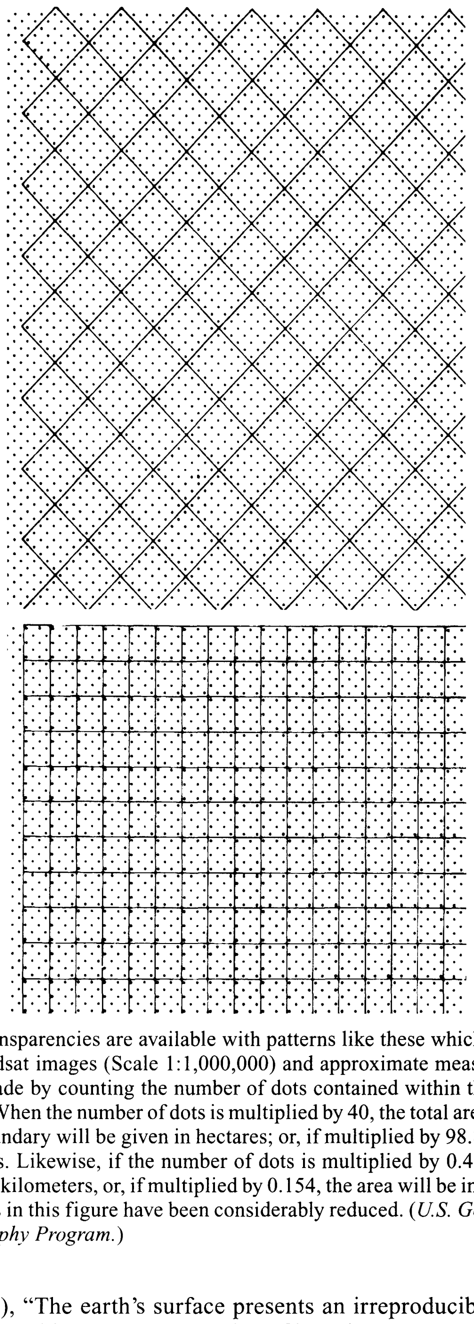


Fig. 7. Transparencies are available with patterns like these which can be overlaid on Landsat images (Scale 1:1,000,000) and approximate measurements can easily be made by counting the number of dots contained within the boundaries of interest. When the number of dots is multiplied by 40, the total area of inclusion within a boundary will be given in hectares; or, if multiplied by 98.8 will give the area in acres. Likewise, if the number of dots is multiplied by 0.4, the area will be in square kilometers, or, if multiplied by 0.154, the area will be in square miles. The patterns in this figure have been considerably reduced. (*U.S. Geological Survey, Geography Program.*)

ence listed), "The earth's surface presents an irreproducible palette to observers in orbit. Accurate reporting of its colors can reveal new facets of both nature and human vision." Russian cosmonauts have been furnished with a color atlas and colorimeter that allows observers to determine the precise hue, saturation, and brightness of natural features. The atlas, developed by the All-Union Mendeleev Research Institute of Metrology, contains 1,000 samples, with a difference in wavelength between hues of about 5 nanometers. In a later development, the atlas was replaced by a sophisticated colorimeter. Too detailed for presentation here, the advantages of instrument-assisted "coloristics" involving human observations, particularly of oceanic characteristics, are well described in this cited reference.

Applications of Land-Observing Satellites

Agriculture. The concept of using remote sensing of agricultural lands and crops was proposed many years ago. The combined use of high-flying aircraft and infrared photography represented an early effort and demonstrated advantages in terms of time and cost as compared with survey crews operating on land. Some uses of satellite surveillance include: (1) early warning of changes affecting production and quality of commodities and renewable resources; (2) specific commodity production forecasts; (3) land use classification and measurements; (4) renewable resources inventory and assessment; (5) land productivity estimates; (6) conservations practices; and (7) pollution detection and impact evaluation, among others. Some of the research has concentrated on the quantitative monitoring of conditions that foretell droughts, and on soil temperature surveillance, including factors that foretell and contribute to grain winterkill. Work on vegetative indexes is being refined, that is, the *green index number* (GIN) which is used to summarize the condition of vegetation within a *Landsat* scene. With this system, digital data from the *Landsat* satellite are put into a scale ranging from 0 for bare soil to 15 for healthy green vegetation. Over time, an understanding of the relationship between these readings and actual yields will allow analysts to quantitatively define production possibilities based on the GIN.

In 1974, jointly sponsored by the U.S. Department of Agriculture, the National Oceanic and Atmospheric Administration, and the National Aeronautics and Space Administration, the Large Area Crop Inventory Experiment (LACIE) was established. During the 1974–1978 period, *Landsat* data were used to estimate wheat production in the United States, Canada, and Russia. It also allowed for technology problem definition in Argentina, Australia, Brazil, India, and China. Working with basic units called sample segments—5 x 6 square nautical miles (25,000 acres; 10,000 hectares) which were further broken down into pixels (1.1 acres, 0.44 hectare)—analysts selected training fields from reconstructed pictures. They identified specific crops in these segments and then programmed the computer to pick out other fields that appeared similar. In this way, area estimates were made. Yield models based upon historical weather information were used together with these area estimates to make production forecasts.

The LACIE estimates were mailed each month ahead of the official U.S. Department of Agriculture crop estimate. Results also were compared directly with those obtained on the ground in test areas, or so-called blind spots, of the United States. The LACIE goal—90/90 accuracy at the country level, or estimates within 10% of the final official estimate 90% of the time—was met during the experiment's third phase for the U.S. winter wheat and the Russian wheat crops. Researchers reported that one of the most critical problems encountered was distinguishing between spring wheat and spring barley—both grown at about the same time in the northern United States and Canada. Moreover, the resolution of the satellite's multispectral scanner was not capable of separating fallow from wheat in the same areas, which were often strip-cropped in widths of 90 meters (300 feet) or less.

In an improved system, retransmission of processed data via satellite can make it possible to deliver data within 60 hours after collection, as contrasted with 29 days during the experiment described. Assessment and interpretation then could follow within a total of 72 hours after data measurement.

Soil-Association Mapping. Soil maps ranging in scales from 1:15,840 to 1:7,000,000 serve such purposes as irrigation and drainage planning, forest management, crop-yield estimates, farm appraisals, and land-use planning. Aerial photography, introduced about 40 years ago, greatly assisted the mapping of soils. Now *Landsat* can image an 8-million acre scene in one frame, allowing comparisons of soil associations over their entire extent, all at the same instant in time and growth. The four spectral bands of *Landsat* and the repetitive coverage made possible by the satellite make subtle differences readily apparent and allow vegetative differences (which are usually a function of varying abilities of the different soil associations to produce vegetation) to be used effectively to help separate soil-association landscapes.

Drought and Desertification. Studies of Landsat imagery of the Sahel region of Africa (critically suffering from drought since 1970) revealed a dark polygonal area bounded by straight lines. It resembled areas known to harbor subsurface water. However, the area was some distance away from any obvious subsurface water source. The unusual linearity of the feature led to the conclusion that it was of artificial origin. The area was later visited and it was found to be a carefully managed, fenced-in ranch, differing noticeably in vegetative cover from the uncontrolled grazing areas outside the fence. This ranch demonstrated that uncontrolled animal grazing, characteristic of this region's agricultural practices, may be at least partly responsible for the advance of the Sahara Desert into the Sahel region. These observations led to suggestions to the governments of the affected nations that careful management of grazing areas may lead to reversal of the Sahara's progress and to the reclamation of some of the land already turned into desert.

Snow and Ice Monitoring. In several areas of the world, such as the western United States, snowmelt and glacial icemelt provide a major part of the annual water runoff needed for irrigation, personal and industrial consumption, and hydroelectric power generation. Results from *Landsat* indicate that snowlines can be observed to within 60 meters (197 feet) under good conditions, that snowcover can be empirically related to water runoff, and that snowcover area can be observed within a few percent. *Landsat* imagery also has been used by cartographers to meet the need for small-scale mapping by national and international polar scientific projects. It has been used to provide numerous and extensive changes to maps of Antarctica, including particularly the regions around the Ross Ice Shelf, Franklin Island, and various ice tongues. Even though maps of the Arctic are much more accurate than those of the Antarctic, several changes have been incorporated in recent maps of the Arctic region. Accurate identification of lake ice and classification of ice types can provide invaluable information for shipping in the Great Lakes, allowing for more efficient routing of shipping through ice-covered waters and the possible extension of the shipping season. Investigators have reported observing pressure ridges, rotten ice, cracks, leads, ice breakup, and other changes on the Great Lakes and also on lakes in New England. By comparing the 0.6 to 0.7 micrometer (band 5) with the 0.8 to 1.1 micrometer (band 7) observations, investigators have been able to delineate the area of melting ice on a lake.

Several investigators have shown that *Landsat* imagery can be used as a survey device to determine the abundance and the dimensional characteristics of Antarctic icebergs as a function of general location. With the polar overlap of *Landsat* imagery, even the prevailing 80% cloud cover yielded sufficient usable imagery for this application. Iceberg heights down to 50 meters (164 feet) have been estimated. Such observations may prove useful later should the concept of towing icebergs to water-starved areas as a source of fresh water become economically practical.

Oil Exploration. Reconnaissance of large sedimentary basins and new exploration provinces provided by *Landsat* imagery has aided rapid interpretation of large features and quickly focused attention on anomalous areas. Types of specific information gathered by *Landsat* have included: (1) data on linear features; (2) general lithologic relationships; (3) closed anomalies of various types correlated with known structural features of oil and gas fields; (4) details of structures controlling hydrocarbon accumulation; (5) overall structure of a basin and its major internal structures; and (6) geologic context of the exploration region.

Ore-Deposit Exploration. Ore deposits tend to be controlled during emplacement by fractures in the earth's crust. One group of investigators studies *Landsat* imagery in detail in central Colorado, well known for its large number of metal-producing mines. Using a snow-covered winter scene of the area, the investigators identified several linears, many of them new discoveries. Some of these linears are nearly straight, but others are curved or arcuate. Most are assumed to be large fractures; the arcuate ones are probably related to underlying intrusions that have cracked the country rock above. The investigators contoured the relative densities of these fractures. Circles 14.5 kilometers (9 miles) in diameter were drawn around the areas of highest fracture density. Of the 10 circles on the *Landsat* image, 5 coincide with major mining districts, indicating a correlation between extensive fracturing and the localization of mineable ore deposits. The remaining 5 circles hopefully will constitute potential targets for exploration. Additional studies have been made of metallic ore deposits in the western Cordillera in Canada and Alaska.

Reports of porphyry copper deposits in central Alaska northeast of Fairbanks have been under investigation. On a *Landsat* image, several of the intrusives have been found along or near a regional linear. The recognition of controlling linears in Landsat imagery can provide an excellent tool in ore-deposit exploration.

In the past, the geologist's basic tool for mineral exploration has been the geologic map. Whenever major improvements in the accuracy and completeness of essential maps can be made, both planning of exploration and eventual field work can be made more efficient. Investigators have shown that *Landsat* images are potentially capable of increasing the rapidity of geologic mapping and reducing its corresponding costs. An African study has provided a good example. Southwestern Africa is noted for its extensive copper-nickel mineralization. Past mapping has been relatively primitive. Investigators in South Africa received several excellent *Landsat* images of this mineral-rich area. By piecing together the images to make a mosaic of the entire region, a rather detailed geologic map of the area was produced for the first time.

Flood-Area Assessment and Flood-Plain Mapping. Regional *Landsat* assessments can be valuable in delineating areas of potential flood damage and for focusing relief efforts of local and federal agencies, of mapping plain areas susceptible to flooding, and of locating areas of extensive flooding where additional flood-control works may be necessary. Such matters are of worldwide significance. Some investigators have used image enhancement and transferred these data to base maps at scales up to 1:100,000 to map floods. The flooded areas are easily delineated by standing surface water or by areas of low, near-infrared reflectance due to excessive soil moisture or stressed vegetation. The estimates of flooding over large regions are accurate to within 5%. For better flood-area assessment, additional efforts will use digital products for mapping to obtain the best resolution. The largest scale for this mapping appears to be 1:24,000.

Land Quality. With *Landsat* data, investigators have mapped strip-mined and reclaimed areas, monitored construction practices, sited new construction, and assessed urban quality. An extensive mapping project of strip-mine and reclamation mapping was completed for areas in Indiana, Ohio, Pennsylvania, and Maryland. Specialists with the U.S. Army Corps of Engineers have prepared vegetation and permafrost maps of portions of Alaska with detail not previously available. Such information will be of much value in evaluating sites for oil pipeline and other kinds of new construction.

Water Quality. *Landsat* data are being used to measure and monitor water quality in rivers, lakes, estuaries, and near-shore coastal zones, Suspended sediment, a water pollutant, occurs as a result of natural processes as well as those of unnatural processes. The sediment load of near-surface waters has been clearly identified in some cases. In several studies, attempts have been made to monitor lake eutrophication, out of which some innovative interpretive techniques have evolved. Determining average reflectance from small areas of water having uniform characteristics can be related to suspended solid concentration in the water (between 5 and 100 milligrams/liter). High reflectivity of algae in the near-infrared permits detecting algal blooms in both lakes and rivers. *Landsat* information has been used routinely in making decisions important to the hydrological needs of agricultural and urban areas and national parks in Florida—as, for example, management of the flow of water and its distribution in the southern Florida drainage canal and impoundment system.

Water Resources. Much of the water supply for populations living in the western United States begins as snow. Below the mountain peaks, rivers are trapped by dams and reservoirs. The captured water must be controlled and released in the proper amounts to provide not only for direct water needs, but also for the electrical energy it can generate. For a number of years, *Landsat* operations have assisted in measuring the amount of snow in western snowpacks. With this information, combined with field data obtained by ground inspections, the volume of water in a spring melt runoff can be anticipated with considerable accuracy. This assists both in energy planning as well as in flood prevention. By attaching a transmitter to a stream gage planted in a mountain stream, it is possible to send the information directly from the stream to *Landsat* every few hours during the snow-melt season. The satellite transmits the data to the ground station where engineers can enter it into the data bank immediately. When floods do occur, Landsat imagery can be used to show the extent of flooding and to monitor changes in flood plain management. Images can record rivers at normal levels and at flood states, demonstrating the value of assembling a chronological set of images.

Air Quality and Weather Modification. The ability to detect certain types of air pollution on *Landsat* imagery was noted by a number of investigators. Air pollution over the Great Lakes had been observed to modify weather. Radiance measurements over water have been used to calculate vertical aerosol burden to about plus-or-minus 10%. In a study of haze in the Los Angeles basin, a correlation was found between radiance and visibility, but the investigator concluded that the radiance values over land were not a particularly sensitive measurement of air pollution. Also, cloud-seeding results were monitored in the Colorado

River Basin. The use of analytical models in conjunction with *Landsat* data was studied by several scientists. The Great Lakes area is being modeled to study the interaction of air pollution and the dynamics of cloud behavior. Consideration is being given to the use of *Landsat* imagery for verification of models to predict plume and contrail dispersion.

Wildlife Resources. Under normal population densities, wildlife species are not resolvable with current *Landsat* imagery. However, investigators have been successful in identifying and measuring habitat factors, such as vegetation and water, both having direct and indirect influences on wildlife population. In the past, vegetation maps of certain wildlife habitats have been gross generalizations due to the difficulties of access to remote and sometimes hostile terrain. *Landsat* not only shows remote areas, but does so synoptically. *Landsat* maps can show between eight and ten vegetation classes, and relatively detailed peripheral boundaries, as compared with former maps of two to four classifications having generalized boundaries. Adding surface water gives a good map of a habitat.

Deforestation. Researchers at the Woods Hole Research Center and the Marine Biological Laboratory reported in September 1986 an ingenious method they have found in connection with analyzing *Landsat* imagery. Healthy vegetation absorbs red light and reflects near-infrared radiation. Bare ground is brighter than vegetation in the red spectral band and less bright in the near infrared. When a patch of forest is cut down, or when clear land is cultivated, the change is visible in *Landsat* images. By superposing two images and subtracting the later (chronologically) image from an earlier image, pixel by pixel, the change can be measured with considerable accuracy. (Each pixel, or picture element, in a *Landsat* image covers an area approximately 59 meters by 79 meters.) It has been reported that this technique has been applied, for example, in connection with the Brazilian state of Rondonia (southwestern Amazon basin). From processed satellite imagery, it has been noted that until 1960 the region was largely untouched tropical forest, but in recent years it has undergone rapid development as the Brazilian government has given away 100-hectare plots to settlers from the crowded southern portion of the country. By comparing previous and present satellite data, it was estimated that the settlers had cleared at least 11,400 square kilometers of forest. The researchers estimate that, using their method, a total global deforestation survey could be made. Governments of less developed countries tend to underestimate deforestation for political reasons. The general hazards of deforestation to preserving a large number of terrestrial species of plants and animals are well established and some deforestations could ultimately result in mass extinctions and a reduction of the Earth's genetic inventory.

Additional Remote Sensing Technologies

High-Altitude Aircraft. These are still used, using *X*-band, 3-centimeter and *L*-band, 25 centimeter radars, as well as thermal infrared multispectral scanners. An 8-channel scanner has been used effectively to gather mineral signature information at wavelengths between 8 and 12 micrometers. Results indicate that igneous rock units can be identified from their free silica content and that carbonate as well as clay-bearing units are readily separable on digitally processed images. Considerable use of this tool has been made in the Death Valley region of California and the Cuprite area of Nevada.

Orbiting Laboratories. This type of craft, for example, downed *Skylab*, which had a 13-channel multispectral imager, can be useful, but the future of such configurations remains indefinite.

The Space Shuttle. Particularly during its first few years, this craft conducted a number of Earth-imaging experiments, including radar and multispectral IR radiometric instrumentation.

Orbiting radar imagery has been employed in planetary explorations and much was learned of Mars in this manner. See article on **Mars.** Also see article on **Photography and Imagery.**

Another *Shuttle* experiment, known as feature identification and location (FILE), sensed radiation from Earth in spectral bands centered at 0.65 and 0.85 micrometers. The instrument compares ratios of the reflected solar radiation in the two wavelengths to make real-time classification decisions about four primary features—water, vegetation, bare land, and a cloud-snow-ice class. The instrument was first flown in aircraft. Together with data collected from aircraft flights, the FILE *Shuttle* experiment results indicated that radiance ratio technology has good potential for use in the development of advanced classification decision-making systems for onboard, real-time applications. Similar tests with the *Shuttle* are planned.

An ocean color experiment (OCE), designed to map ocean features with an 8-channel scanning radiometer, was installed on one of the early *Shuttle* excursions. The operating principle of the instrument relied on for this experiment has been known for a number of years. In 1972, NASA began high-altitude aircraft sensor investigations known as the U-2 ocean color scanner program. In October 1978, the coastal zone color scanner (CZCS) was launched on *Nimbus 7* (weather satellite) to make periodic observations of ocean color, primarily in coastal areas. The concept of a shuttle-borne scanner originated as a result of these developments. The primary aim of OCE was to detect phytoplanktonic algae on a global basis and to determine the chlorophyll pigment concentrations. The OCE was set up to focus mainly on deepwater areas in contrast to coastal waters. Ocean images were obtained at three widely separated locations (Yellow Sea, Gulf of Cadiz, and Grand Bahama Bank). The researchers agreed that the method of using chlorophyll concentration as a tracer may be applied for the deduction of oceanic flow patterns in large areas. Plankton patches are natural drifters and can be tracked by satellites. Thus, the color scanner proved its utility is studies of both ocean circulation and biological processes. (See Kim reference.)

Ocean scientists have long recognized the special role that satellites can play in learning more about the atmosphere-ocean interface, global biological processes, particularly near-surface ocean biology, and the biogeochemical cycles of carbon dioxide, among many other poorly understood topics. Several new satellites for studying these topics were launched during 1989–1990. The first was the Navy's Remote Ocean Sensing System (NROSS) which was designed to provide information on ocean waves and eddies and research data on surface winds. The second was a joint United States-French mission (TOPEX/POSEIDON) to carry a high-precision altimeter to measure the topography of the ocean surface, which results from the combined effects of winds, currents, and gravity. The combined data from NROSS and TOPEX/POSEIDON yielded for the first time a synoptic global description of ocean circulations. These studies were extended by a third and fourth satellite, focusing on biology and geodesy, respectively. A geopotential research mission was recommended for launch in the very early 1990s, focusing on the Earth's gravity and magnetic fields. As suggested by Revelle, measurements of the Earth's gravity would enable a determination of the effect of gravity on the topography of the ocean surface (the geoid) and thus, in combination with altimeter measures, a determination of absolute currents.

A few years ago, *Seasat*, equipped with synthetic aperture *L*-band radar, furnished important oceanographic information. See also **Ocean.**

Heat Capacity Mapping

This was performed by a satellite of that name (*HCMM*). It provided thermal and visible images for mapping the thermal inertia of surface materials. See Figs. 8 and 9. The satellite was used in connection with (1) urban heating patterns; (2) freeze damage assessment with development of planting date advisories; (3) evapotranspiration rates; (4) soil moisture assessment; (5) thermal mapping for discrimination of geologic units and energy or mineral resource areas and evaluation of thermal modeling and satellite mapping techniques; (6) detection of high potential groundwater pollution; (7) studies of snow hydrology programs; (8) studies of estuarine currents; (9) a topoclimatological and snow-hydrological survey of various countries, such as Switzerland; (10) monitoring of large scale pollution effects in the North Sea, among several others. Scientists from several countries, including Australia, Canada, France, Germany, Italy, Morocco, Spain, Switzerland, and the United Kingdom, participated in investigations involving the *HCMM*.

Resolution of Reconnaissance Satellites

Based upon the declassification of former defense documents, it has become evident that the military made extensive progress in the development of reconnaissance satellites in terms of the resolution of ground-level images. This topic is reviewed by J. T. Richelson (January 1991 reference listed). Sensors developed for military purposes used

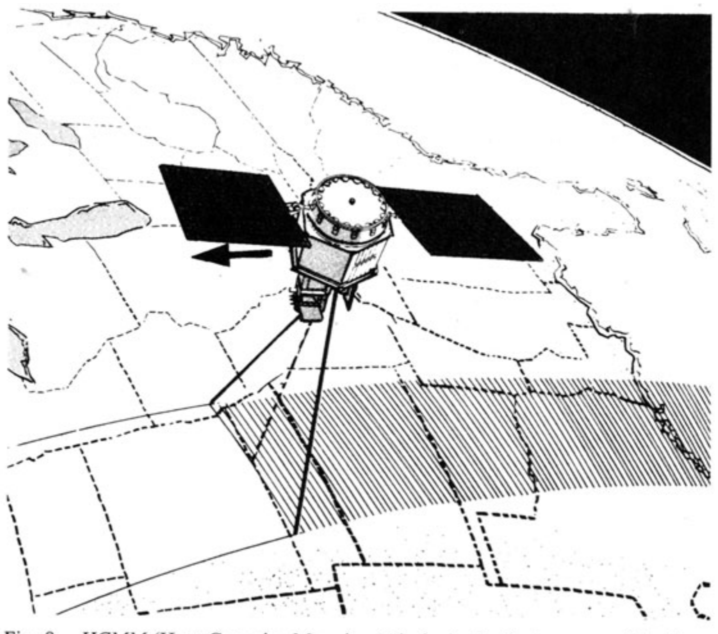


Fig. 8. *HCMM* (Heat Capacity Mapping Mission), the first spacecraft built to test the feasibility of measuring variations in the earth's temperature was launched on April 26, 1978. Purpose of mission—to produce thermal maps for discrimination of rock types, mineral resources, plant temperatures, soil moisture, snow fields, and water runoff. The experimental satellite traveled in a circular, sunsynchronous, 620-kilometer (385-mile) orbit that allowed for measuring midlatitude test areas of the earth's surface for their minimum temperatures and then measuring those same areas for maximum temperatures about 11 hours later. (*National Aeronautics and Space Administration.*)

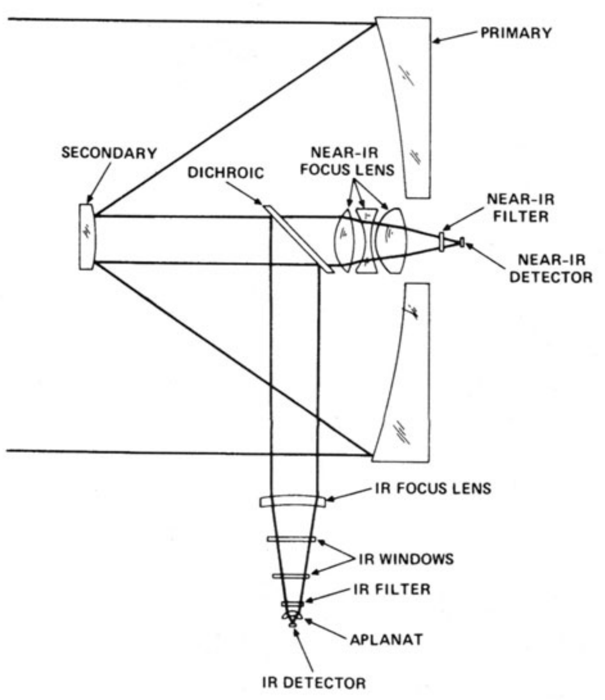


Fig. 9. Optical system of *HCMM* spacecraft. The instrument module was located on the spacecraft so that the Heat Capacity Mapping Radiometer (HCMR) was earth-pointing. The basic radiometer instrument was a modified spare Surface Composition Mapping Radiometer similar to that flown on Nimbus 5. The HCMR had a small geometric field of view (less than 1 | 1 milliradians), high radiometric accuracy, and a wide enough swath coverage on the ground so that selected areas could be covered within the 12-hour period corresponding to maximum and minimum temperatures.

The instrument operated in 2 channels, at approximately 0.5 to 1.1 micrometers (visible and near-infrared) and 10.5 and 12.5 micrometers (infrared), providing measurements of reflected solar energy and equivalent black-body temperatures.

charge-coupled devices that transform varying light levels of a scene into digital signals, which are then transmitted to a ground station by way of a relay aircraft. These sensors made it possible to increase the number of targets from about 20,000 in the early 1970s (mainly in China and the former Soviet Bloc) to some 42,000 targets, about half of which were outside the eastern European countries and the Asian countries, including China. These particular reconnaissance satellites, however, were limited to daylight hours and clear (no clouds) skies.

In late 1988, a satellite termed *Indigo* or *Lacrosse* operated on radar principles, much like those used later by the spacecraft *Magellan*. See **Venus.** Richelson reports that the resolution of *Lacrosse* probably was between 3 and 10 feet (0.9 to 3 m) or roughly ten times better than the resolution of *Magellan*. Much additional information is contained in the Richelson reference. Also see Burrows, Krepon, and Johnson references listed.

"Mission to Planet Earth" Impacts Satellite Programs

As of late 1993, the political and economic factors in the United States and worldwide are not conducive to achieving "good" science in connection with the planning and execution of the "Mission to Planet Earth" program. Uncertainties include the future of the space shuttle, the construction of ever-increasing bulky Earth-orbiting satellites, the vascillating status of the space station concept, and the threat of creating massive quantities of information in the absence of a system that can easily retrieve and analyze data returned from space.

Thus, intentions rather than *fait accomplii* are the essence of current space affairs.

The Sheer Bulk of Proposed Satellites

Over a time span extending well into the next century, massive quantities of data pertaining to nearly all scientific aspects of planet Earth will be returned to ground stations and widely distributed to thousands of scientists worldwide for analysis.

During the late 1980s and early 1990s, the general position taken by many NASA scientists was a preference for fewer, larger satellites (space platforms) that would carry numerous key sensors. These would share a central power supply and other common electronic circuitry components. Initial evaluation of this concept showed some advantages, including cost savings and some design simplification at the satellite and ground support levels.

After the first announcement of a "fewer but larger" satellite design philosophy, considerable resistance within the scientific community developed. Among the factors of concern were:

1. Heightened risk of losing the inputs from numerous sensors upon a failure of the power supply,
2. Inability to repair a large platform in space by a shuttle crew, as has been done successfully several times in the past, and
3. requirements for using large rockets, rather than the shuttle, for launching massive satellites into orbit.

Savings to be derived from scaled-up satellites were scrutinized and became a less-convincing incentive. By 1992, NASA softened its earlier decision and revised some of its long-range plans. See Fig. 10.

Thus, the first stage of the "Mission to Planet Earth" program included the comparatively large UARS (Upper Atmosphere Research Satellite) in mid-September 1991, but the huge EOS-A (the first of two or more very large platforms) is under review. This is scheduled for launch in 1996, but may be delayed for a number of reasons. In the early phases of the "Mission to Planet Earth" program, some projects were scaled down. These include comparatively small, specialized satellites for observing Earth's ozone layer, a joint project with France to measure with precision the height of the world's oceans, from which global ocean circulation can be deduced, and the installation of some sensors on a Japanese satellite for measuring wind stress on ocean surfaces.

Massive Data Problem

Many scientists have been concerned with the great amounts of information that will be created over a comparatively short time span by the "Mission to Planet Earth" program. Some experts have questioned the ability to retrieve and analyze the torrents of information that will be received. First, it is doubted that sufficient thought has been given

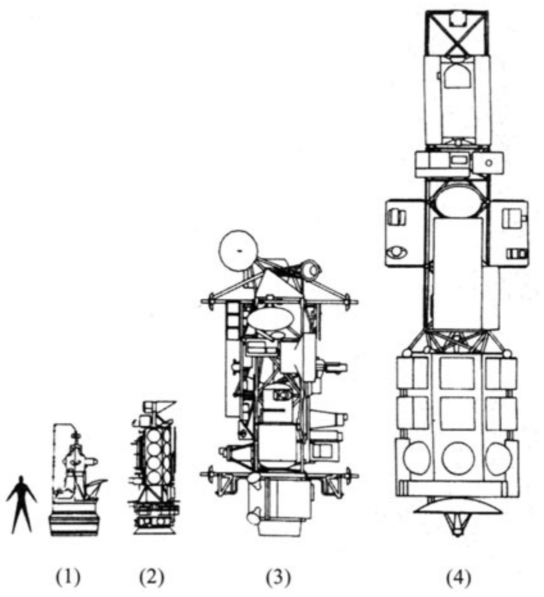


Fig. 10. Comparative sizes of U.S. satellites: (1) *Nimbus-7* and (2) *TIROS* weather satellites of the 1960s, (3) UARS (Upper Atmosphere Research Satellite), launched in 1992, and (4) EOS (Earth-Observing System), originally proposed for 1996 launch. (*National Aeronautics and Space Administration.*)

to how such information will be processed by existing centers using conventional methods. This leads to the query, "Has sufficient funding been included to create an expanded data center?" Also, there is the concern that plans for *analyzing* and *interpreting* the data and the availability of scientists to theorize and create solutions have not been made. In other words, has too much attention been targeted to hardware and data acquisition, as compared with the tasks of data analysis? (A comparison of this situation could be that of the initial enthusiasm in connection with the Human Genome project a few years ago and the fact that the program has suffered from management turnovers during the interim.)

Another factor of concern, not frequently stressed, is simply that of public acceptance. Although the public generally is aware of Earth's environmental problems, it is feared that the "Mission to Planet Earth" may be regarded as pretty "dry stuff" and not comparable to earlier interest in putting a person on the moon.

Long Duration Exposure Facility (LDEF) Satellite

Specifically with long-term satellite performance in mind, on April 7, 1984, NASA launched the LDEF for testing the long-term effects of the space environment on spacecraft materials, components, and systems. Weighing 11 tons and the size of a school bus, the 12-sided, 30-foot (9.1-m) craft contained dozens of trays on its exterior for exposing numerous materials and objects to the direct effects of the space environment. A total of 10,000 test specimens were exposed, representing selections made by scientists from the United States and eight other countries. Also included were millions of tomato seeds for testing the effects of long-term radiation on living tissue. After recovery and subsequent planting, well over half of the seeds were capable of germinating. The effects of some mutations were observed.

Initially, it was planned to recover the LDEF by the space shuttle, but this was delayed because of the *Challenger* tragedy. By January 1990, the LDEF began losing altitude and imminent recovery action was required to prevent the craft from burning up in Earth's atmosphere upon its unassisted return. It was recovered by the shuttle *Columbia* about 200 miles (322 km) above Baja, California, and ultimately returned to the Orbiter Processing Facility at Kennedy Space Center, Florida, in early January 1990.

The major hazards noted from evidence of specimens on the LDEF in low orbit included:

1. Radiation,
2. Meteoroid impacts, and
3. Erosive and corrosive effects of atomic oxygen, which makes up more than 90% of the thin atmosphere of low-Earth orbit.

Atomic oxygen was found to be particularly destructive of polymeric materials, with evidence of broken molecular bonds. Research of the materials aboard the LDEF will be studies for several years, but already has contributed to the selection of superior materials for a possible forthcoming space station. Much of the research is being conducted at the University of Texas at El Paso. See L. E. Murr reference listed.

Space Station

From the initial suggestion of a permanent, peopled laboratory and possible staging station for space experiments and ventures, the concept has encountered some fundamental resistance within the scientific community, particularly pertaining to timing and costs. As of the end of 1993, the socio-political atmosphere has not changed much and, consequently, the program continues to be affected by uncertainties and some lack of support. Also, the general public in the United States to date has not understood nor has it been sufficiently excited to support the program without a number of inhibitions. Further, the "race" for space achievements with the former Soviet Bloc is no longer a factor.

Chronology of Space Station Concept

Considering the present state of technology, a *space station* may be defined as a manned structure in space that is too massive to be launched from Earth in one piece, but that will have to be assembled in space. Such a station would house a half-dozen or more astronauts, scientists, and technicians on a semi-permanent (many months) basis, with relatively infrequent crew changes. In essence, a space station would be a marked extension of the downed U. S. *Skylab* or more recent vehicles operated by the former Soviet Bloc. If it were not so far away, the moon could serve as a station in space except that the moon has gravitational pull, which those scientists who are interested in microgravity experiments wish to elude. See article on **Microgravity and Materials Processing.**

The state of space station science as recently as the mid-1980s is illustrated and described in Figures 11, 12, and 13. It has been variously

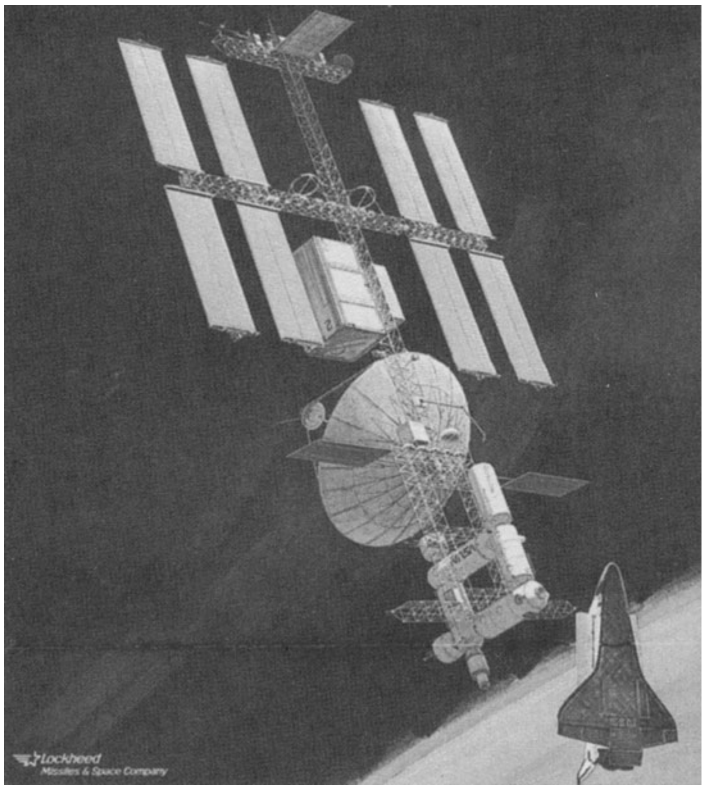


Fig. 11. An early 1984 configuration of the space station envisions a 300 × 500-foot (91 × 152-meter) structure with a single main horizontal beam. Solar pannels to provide power to the station are mounted on the principal horizontal beam. The so-called "race track" configuration at the bottom is the living quarters. From time to time, laboratory modules would be added as needed. (*Lockheed Missiles & Space Company.*)

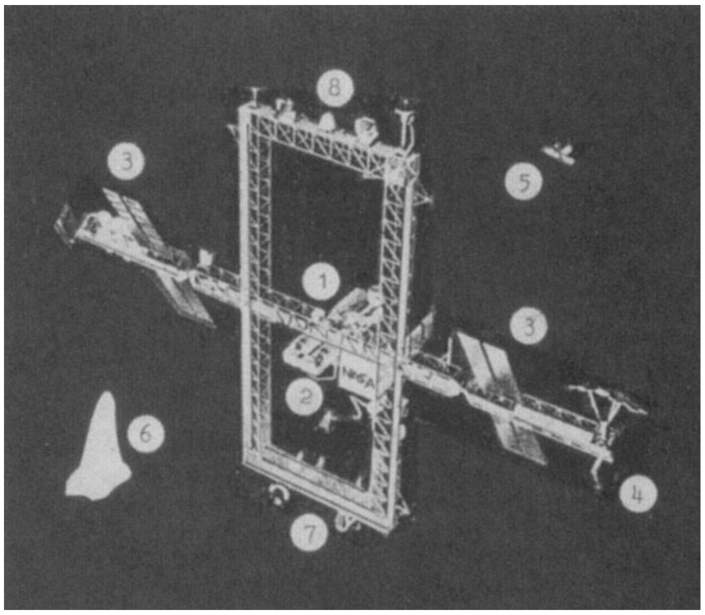


Fig. 12. 1988 configuration of United States space station: (1) central horizontal boom; (2) four habitation and laboratory modules, including one module each from Europe and Japan; (3) two photovoltaic arrays on each end of the boom for delivering 60 kW of power; (4) a Canadian-built, remote-control servicing arm; (5) an unmanned polar-orbiting platform with remote sensing instruments; (6) the *Space Shuttle* shown for size comparison; (7) lower boom to support downward Earth-looking instruments; and (8) an upper boom for supporting astronomical instruments. (*After NASA*.)

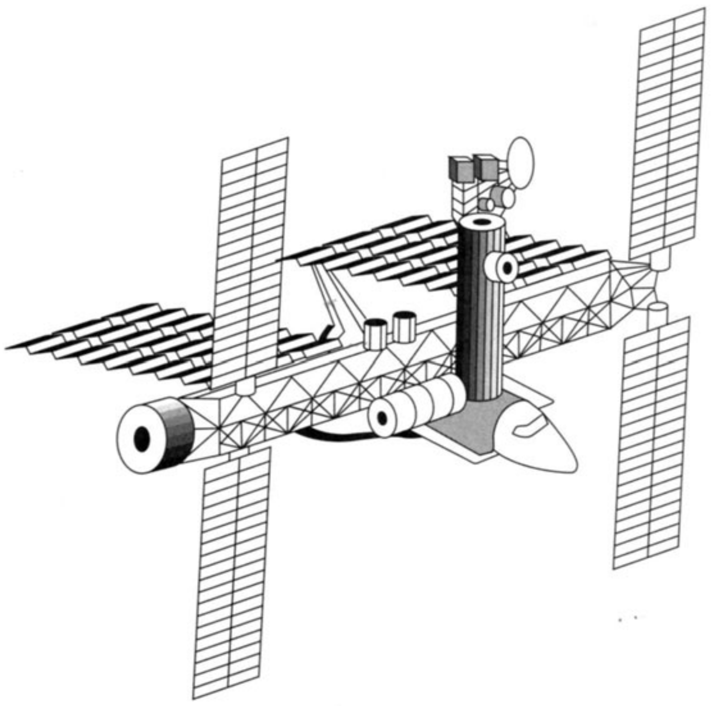


Fig. 14. One of several space station designs submitted to NASA during the early 1990s and that appears to be the most favored at the end of 1993. Scaling down of the space station from prior designs is immediately obvious. (National Aeronautics and Space Administration.)

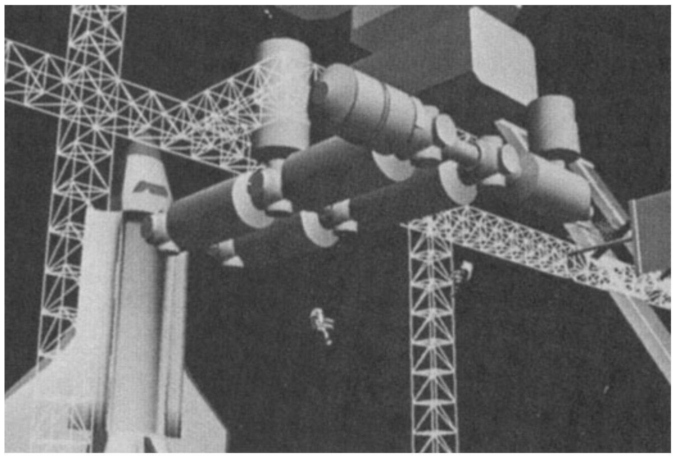


Fig. 13. 1986 space station configuration as proposed by McDonnell Douglas Corporation. This is a computer-generated image, giving a close-up of the shuttle docking area. Scale is indicated by a few astronauts shown in diagram. Several differing design formats have been prepared by prospective constructors of the space station.

estimated that through 1993 the space station program has consumed some \$9 to \$10 billion. Thus, it is no surprise that the current trend is that of reducing station cost, and this inevitably extends to lower station size and less ambitious experimentation. The costs and techniques of constructing (assembling) the station in space, of course, are a key factor. Delays in the program to date have not all been unfortunate, in that numerous changes in construction materials can be made as the result of data received from the LDEF satellite.

The obvious scaling down of the space station is evidenced by a diagram (Fig. 14) of one of the 1993 versions. Even though plans have been drawn, the space station remains vulnerable to the vicissitudes of budgeting and politics.

Space Shuttle

For many years, space scientists and engineers envisioned a spacefaring ship that could be used many times as a ferry for personnel and cargo between various locations in space and the Earth—as contrasted with the high costs of using entirely new equipment for each space mission. After several years in design, construction, modification, and testing, the first successful demonstration flight of such a craft, the *Columbia*, occurred during the spring of 1981. Liftoff occurred at Cape Canaveral, Florida at 7 A.M. (EST) on April 12; touchdown occurred at Edwards Air Force Base, California at 10:21 A.M. (PST) on April 14, for a mission span of 54 hours, 21 minutes in an orbit 198 miles (319 kilometers) from Earth. The orbital speed of the ship was 17,000 miles (27,353 kilometers) per hour. Requiring a crew of two persons, the first demonstration flight was considered by almost everyone as highly successful and well beyond expectations. A second test mission occurred with liftoff from Cape Canaveral on November 12, 1981. Scheduled for a total time in space of about 5 days, the mission had to be shortened because of a malfunction in one of the three fuel cells that generated electrical power for the *Columbia*. However, during the shortened mission, the abilities of the craft's manipulating arm (known as *Canada*) for placing satellites and other objects into a space orbit as well as retrieving them were satisfactorily demonstrated—as was the feasibility of a number of scientific experiments placed aboard the craft. It is envisioned that when the *Space Shuttle* becomes fully operational, it will be capable of a 14-day turnaround, from landing to liftoff. The third test mission was completed on March 30, 1982.

After the January 1986 accident, the shuttle program was put on hold, while design changes could be made. The first successful flight of a redesigned craft, the *Discoverer,* was launched on September 29, 1988. Later, the shuttle *Endeavor* was added to the fleet.

The shuttle consists of an *orbiter,* an *extenal tank,* and two solid-rocket *boosters*. The orbiter and boosters are reusable.

About the size of a commercial DC-9 jetliner, the Orbiter can deliver to orbit single or multiunit payloads up to 29,484 kilograms (65,000 pounds) in a cargo bay that measures 4.5 × 18 meters (15 × 60 feet), and bring back payloads weighing up to 14,515 kilograms (32,000 pounds). See Fig. 15.

Typical Mission. A shuttle mission begins with installation of the mission payload into the Orbiter cargo bay. The payload is checked out and serviced and activated after reaching orbit. The solid rocket boosters and the orbiter main engines fire together at liftoff. The two solid rocket boosters are jettisoned after burnout—when about 44 kilome-

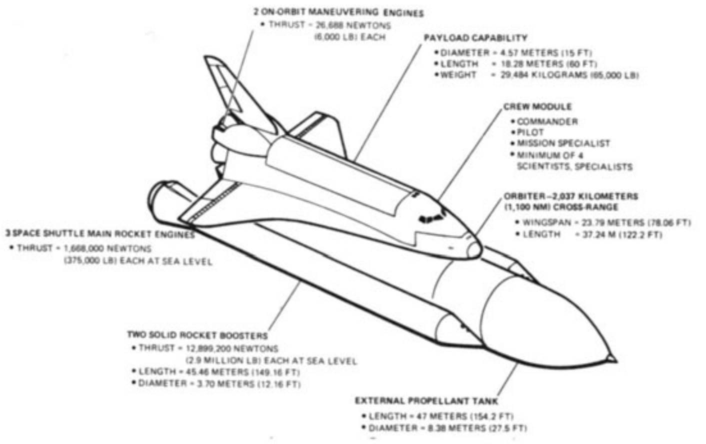

2 ON-ORBIT MANEUVERING ENGINES
• THRUST = 26,688 NEWTONS (6,000 LB) EACH
PAYLOAD CAPABILITY
• DIAMETER = 4.57 METERS (15 FT)
• LENGTH = 18.28 METERS (60 FT)
• WEIGHT = 29,484 KILOGRAMS (65,000 LB)
CREW MODULE
• COMMANDER
• PILOT
• MISSION SPECIALIST
• MINIMUM OF 4 SCIENTISTS, SPECIALISTS
ORBITER–2,037 KILOMETERS (1,100 NM) CROSS-RANGE
• WINGSPAN = 23.79 METERS (78.06 FT)
• LENGTH = 37.24 M (122.2 FT)
3 SPACE SHUTTLE MAIN ROCKET ENGINES
• THRUST = 1,668,000 NEWTONS (375,000 LB) EACH AT SEA LEVEL
TWO SOLID ROCKET BOOSTERS
• THRUST = 12,899,200 NEWTONS (2.9 MILLION LB) EACH AT SEA LEVEL
• LENGTH = 45.46 METERS (149.16 FT)
• DIAMETER = 3.70 METERS (12.16 FT)
EXTERNAL PROPELLANT TANK
• LENGTH = 47 METERS (154.2 FT)
• DIAMETER = 8.38 METERS (27.5 FT)

Fig. 15. Principal features of the space shuttle. (National Aeronautics and Space Administration.)

ters (27.5 miles) high—and are recovered by means of a parachute system. A homing device on each booster guides recovery craft to it for towing the equipment back to shore. The boosters are then refurbished, refueled, and made ready for another flight. The orbiter main engines continue to burn until the orbiter is just short of orbital velocity, at which time engines are shut down and the external tank is jettisoned. During its plunge through the atmosphere, the tank will tumble, break up, and fall into a predetermined ocean area—the Indian Ocean for launches from Florida; the South Pacific Ocean for launches from California.

The orbital maneuvering system is used to attain the desired orbit and to make any subsequent maneuvers that may be needed during the mission. The crew then begins their payload operations by performing numerous assigned tasks, depending upon the purpose of the mission, such as retrieving satellites from Earth orbit, servicing orbiting satellites, conducting experiments in space, and studying the Earth and deep space from a special vantage point high above the atmosphere of the planet.

After orbital operations are completed, normally assumed to be about 7 days, deorbiting maneuvers are initiated. The orbiter begins to reenter the Earth's atmosphere at a high angle of attack. During reentry, portions of the orbiter's exterior will reach temperatures up to 1260°C (2300°F). The orbiter will level into horizontal flight at low altitude for an unpowered aircraft-type approach and will land at a speed of about 335 kilometers (208 miles) per hour. After landing, the orbiter is towed to ground facilities where any returned payloads are removed. Orbiter refurbishment operations will make it ready for another mission in about two weeks.

The Orbiter. Dimensions of the orbiter were given earlier in this entry. The orbiter's exterior is covered with thermal protective materials to protect the spacecraft from solar radiation and the extreme heat of atmospheric reentry. Two types of reusable surface insulation, coated silica tiles and coated flexible sheets, cover the top and sides of the orbiter. The tiles protect the surfaces up to 649°C (1200°F); the flexible insulation provides protection up to 371°C (700°F). The coating on both types of insulation gives the orbiter a nearly white color and has optical properties that reflect solar radiation. On the bottom of the orbiter and on the leading edge of the tail, a high-temperature reusable surface insulation, made of coated silica tiles, protects the aluminum structure up to 1260°C (2300°F). The high-temperature coating is of a glossy black appearance. A reinforced carbon-carbon material is used for the nose cap and the wing leading edges where the temperatures exceed 1260°C (2300°F).

In a normal mission, the orbiter will carry a crew of three and as many as four additional payload and technical personnel, who will occupy a two-level cabin within the crew module at the forward end of the vehicle. The cabin, a combination living, working, and storage area, is pressurized with a nitrogen/oxygen mixture to an atmospheric pressure of 14.7 pounds per square inch (10.1 Newtons per square centimeter) to simulate sea level conditions. It is interesting to note that prior United States manned spacecraft used an atmosphere at a pressure of 5 pounds per square inch (3.45 Newtons per square centimeter).

The upper section (flight deck) contains the controls and displays used to pilot, monitor, and control orbital maneuvering, atmospheric reentry, and landing phases of the mission and to control the mission payloads. Seating for the crew, consisting of the Commander, Pilot, and Mission Specialist, plus one payload specialist, is provided on this level. The Commander and Pilot are seated in the usual pilot/copilot arrangement with duplicate controls that permit the orbiter to be piloted from either seat and returned to Earth by one crew member in an emergency situation.

Seating for three passengers/scientists and the habitability provisions are located on the lower level or deck. The habitability provisions include a galley for food preparation (an oven and hot and cold water dispensers for the preparation of rehydratable freeze-dried foods) and an eating area, personal hygiene facilities for both sexes, and sleeping accommodations.

In the case of need for the shuttle for a rescue mission, the lower deck can be configured to provide three additional seats by replacing the sleeping quarters with seats. Also to support a rescue mission, all shuttle flights will carry extravehicular activity provisions for two trained crew members. These provisions are planned on missions where access to the cargo bay and the payloads is required. The crew members and payload specialists will exit the orbiter through an airlock and hatch in the lower deck. It is also through this hatch that the scientists for Spacelab missions will enter the pressurized Spacelab module to conduct the planned experiments.

Main Engine.[3] The shuttle's main engine, along with two solid rocket boosters, is the most advanced liquid-fueled rocket engine built to date. With variable thrust permitting the engine thrust to be tailored to the mission needs, it can operate effectively at both high and low altitudes. This system has the highest thrust for its weight of any engine yet developed and can operate up to $7\frac{1}{2}$ hours of accumulated firing time before major maintenance or overhaul is required. It is planned that the main engine will be reusable for up to 55 separate shuttle missions.

Three main engines are mounted on the orbiter aft fuselage in a triangular pattern. The engines are so spaced that they can be gimballed during flight and, in conjunction with the two solid rock boosters, are used to steer the shuttle during flight as well as provide thrust for launch.

Fuel for the engines, liquid hydrogen and liquid oxygen, is carried be supplied from the tank at a rate of about 171,396 liters (45,283 gallons) per minute of hydrogen and about 63,588 liters (16,800 gallons) per minute of oxygen.

Numerous features have been designed into the engine to satisfy the performance, life, reliability, and maintainability requirements of the shuttle. One feature includes the use of a stage combustion power cycle coupled with high combustion chamber pressures. In the staged combustion cycle, the liquid hydrogen is partially burned at high pressure and relatively low temperature in preburners and then completely combusted at high temperature and pressure in the main combustion chamber before expanding through the high-area ratio nozzle. The rapid mixing of propellants under these conditions is so complete that a combustion efficiency of about 99% is attained. The engine also uses hydrogen fuel to cool all combustion devices in direct contact with high-temperature combustion products, thereby contributing to long engine life.

Each engine has three primary levels of thrust or power—minimum, rated, and full power. Engine thrust, however, can be varied throughout the range from minimum to full power level, depending upon mission needs. It is planned that most shuttle flights will be launched at rated power level with each engine developing 2,090,560 Newtons (470,000 pounds) of thrust at altitude; or 1,668,000 Newtons (375,000 pounds) at sea level. On some shuttle flights where heavy payloads dictate an extra measure of power, up to full power level thrust can be commanded. This level equals 109% of rated power. During the latter part of ascent, engine thrust is reduced to ensure that an acceleration force

[3]After the *Space Shuttle* accident of January 1986, certain changes were made in the rocket engine design. Details of design changes were not available at the time of publication of this volume. A succinct summary of design problems identified after the 1986 accident are well summarized in *Sci. Amer.*, pp. 62–64 (August 1986).

of no more than three times that of Earth's gravity is reached. This acceleration level, permitted by the throttleable shuttle engines, is about one-third the acceleration experienced in prior manned space flight and is well under the physical stress limits of non-astronaut scientists who will fly aboard the shuttle. The lowest thrust throttle setting, the minimum power level, equals 65% of rated power. See Fig. 16.

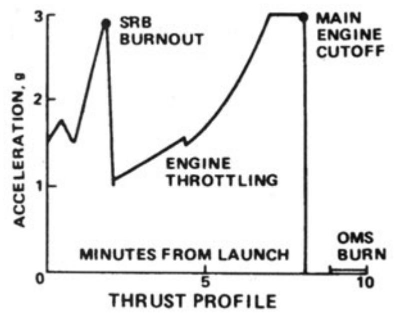


Fig. 16. Thrust profile of space shuttle engines. (National Aeronautics and Space Administration.)

Solid Rocket Boosters. Two solid rocket boosters are used for each shuttle flight to provide, along with the orbiter main engines, the initial ascent thrust to lift the shuttle with its payload from the launch pad to an altitude of about 44 kilometers (27.5 miles). Prior to launch, the entire shuttle weight is supported by the two boosters.

Each solid rocket booster is made up of six subsystems—the solid rocket motor, structures, thrust vector control, separation, recovery, and electrical and instrumentation subsystems. The overall length of the solid rocket booster is 45.5 meters (149.1 feet), and the diameter is 3.7 meters (12.2 feet).

The heart of each booster is the motor, the largest solid rocket motor ever flown to date and the first designed for reuse. The motor is made up of eleven segments joined together to make four loading segments which are filled with propellant at the manufacturer's site. The segmented design permits ease of fabrication, handling, and transportation. The segments are shipped from the manufacturer's site to the launch sites by rail in specially built cannisters carried on a flat rail car. At the launch site, they are assembled to make up a complete motor.

Propellant loading of the motor segments is completed in pairs from batches of propellant ingredients to minimize any thrust imbalances between boosters used for a single shuttle flight. Propellant loading, using different internal propellant shapes (*cores*) is done in such a way that will cause a regressive thrust 55 seconds into the shuttle flight. This prevents overstressing the shuttle vehicle during the critical phase of flight (called the period of maximum dynamic pressure). Each motor, when assembled, contains about 503,600 kilograms (1.1 million pounds) of propellant and at launch develops a thrust of 12.9 million Newtons (2.9 million pounds).

The exhaust nozzle in the aft segment of each motor, in conjunction with the orbiter engines, steers the shuttle during flight. The nozzle can be moved up to 6.65 degrees by the booster's hydraulically operated thrust vector control subsystem. The latter is controlled by the orbiter's guidance and control computer.

Throughout flight, measurements are taken to verify proper booster performance. The signals are routed to the orbiter for data recording and transmission to the ground. Electrical power for solid rocket booster subsystems is supplied from the orbiter fuel cells through interconnect cabling from the external tank.

At burnout, the two solid rocket boosters are separated from the external tank by pyrotechnic (explosive) devices and moved away from the shuttle vehicle by eight separation motors, four housed in the forward nose frustum and four on the aft skirt. Each of the eight separation motors, fired at solid rocket motor burnout, develops a thrust of 97,856 Newtons (22,000 pounds) for a duration of a little more than one-half second, just sufficient to move the boosters away from the still-acceleration orbiter and external tank. Also a part of the system is a device for separating electrical interconnection with the external tank.

The facility for providing the method to control the boosters' final descent velocity and altitude after separation is the recovery subsystem. This system in the forward section of each booster and within the nose cap consists of parachutes and location aids to help in the search and retrieval operations for each expended booster and its parachutes.

Following separation and entry into the lower atmosphere at about 4700 meters (15,420 feet), each booster is slowed by a pilot and drogue parachute and finally by three main parachutes, each 31.7 meters (104 feet) in diameter, to impact the water about 257.5 kilometers (160 miles) downrange, at a speed of about 95.5 kilometers (60 miles) per hour, aft and first. By entering the water in this way, the air in the hollow boosters is trapped and compressed, causing the boosters to float with the forward end out of the water. At booster impact, the main parachutes are disconnected and the direction finding beacons and lights are actuated to guide recovery craft to the boosters and parachutes. The parachutes are picked up by the recovery craft, and the boosters are towed to shore where they are disassembled and refurbished. The motor segments are shipped to the manufacturer by rail for refurbishment and reloading for a subsequent shuttle flight. The other systems are refurbished either at the launch site, or at the respective manufacturer's location.

The thrust vector control subsystem, structural subsystem, and the electrical subsystems are planned for 20 flights and the recovery subsystem for 10 flights. The separation system is not planned for reuse.

At the launch site, the two boosters are assembled vertically on the launch platform. Following assembly of the boosters, the external tank is attached to the boosters and finally the orbiter and payload to the external tank, thus making the shuttle ready for checkout and another flight.

External Tank. This tank has two principal roles in the space shuttle program: (1) To contain and deliver quality propellants, liquid hydrogen and liquid oxygen, to the engines; and (2) to serve as the structural backbone of the shuttle during launch operations.

The external tank is composed essentially of two tanks—a large hydrogen tank and a smaller oxygen tank, joined together to form one large propellant storage container that is 47 meters (154.2 feet) in length and 8.4 meters (27.5 feet) in diameter.

The oxygen tank is the forward portion of the external tank and, when loaded, contains 603,983 kilograms (1,331,783 pounds) of liquid oxygen. By comparison, the oxygen tank has considerably more volume than that of a house with a floor area of 186 square meters (2000 square feet). See Fig. 17.

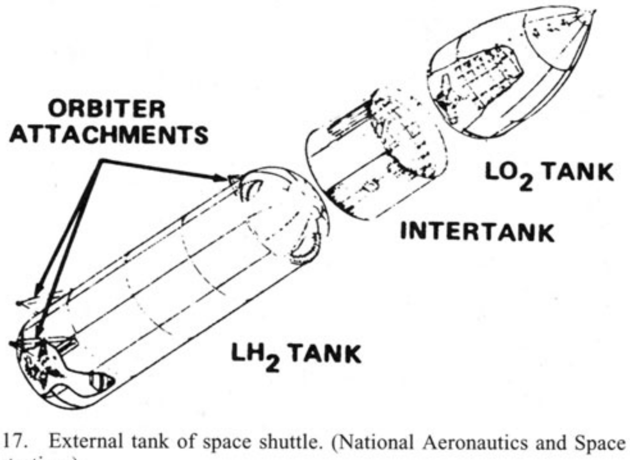


Fig. 17. External tank of space shuttle. (National Aeronautics and Space Administration.)

The forward end of the oxygen tank curves to a point to reduce aerodynamic drag and also to provide lightning protection for the shuttle vehicle once the shuttle has cleared the launch pad. Prior to launch, lightning protection is provided by the launch tower.

The liquid hydrogen tank, aft of the oxygen tank is about 2.5 times larger than the oxygen tank. In this tank is stored 101,503 kilograms (223,814 pounds) of cold liquid hydrogen (approximately −251°C; −240°F).

The intertank joins the two tanks and provides a protective compartment to house some of the instrumentation components in the space between the two propellant tanks.

For launch, the external tank supports the orbiter and solid rocket boosters at attach points on the tank. Since the thrust is generated by the orbiter main engines and the solid rocket boosters, the external tank must absorb the thrust loads for the shuttle during launch. The intertank takes the major thrust loads from the solid rocket boosters; the orbiter main engine thrust loads are transferred through other attach fittings on the tank.

Much of the outer surface of the tank is protected thermally. Spray-on foam insulation is applied over the forward portion of the oxygen tank, the intertank, and the sides of the hydrogen tank. The foam insulation is required for reducing ice or frost formation on the tank during launch preparation, thus protecting the orbiter insulation from free-falling ice during flight, to minimize heat leaks into the tank which would cause excessive boiling of the liquid propellants, and to prevent condensation and solidification of the air next to the tank.

An ablating material (substance that chars away) is applied to the external tank bulges and projections to protect them from aerodynamic heating during flight through the atmosphere. Sometimes a combination foam and ablator is used where both heating and insulation protection are needed. Protection is also required in areas where exhausts of launch engines provide high radiant energy to the tank and where separation motor exhaust plumes may strike the tank.

The external tank, having no electrical power, obtains required electrical energy from the orbiter fuel cells. It does, however, provide the needed cabling to carry power and signals to the external tank electronics and instrumentation components and to the two solid rocket boosters.

Fluid controls and valves, except the vent valves, for operation of the engines, are located in the orbiter. This is done to minimize throwaway costs, inasmuch as the external tank is not reused.

During flight, the two tanks are pressurized by gases supplied from the three engines. Pressurization is required for structural support of the tank and for operating pressure requirements of the engine pumps.

Near the end of the launch phase of a shuttle mission, when the orbiter is just short of orbital velocity, the main engines are cut off. About 10 seconds later, the external tank is severed from its attachment to the orbiter, playing a totally passive role in the separation sequence. Just prior to separation, the external tank tumbling system is activated, opening a valve and venting the oxygen tank through the nose cap. This causes the external tank to pitch away from the orbiter and begin to tumble at a rate that will assure that the tank will break up upon reentry and fall within the designated ocean impact area.

A variety of uses for the space shuttle has been proposed during the next few years, ranging from scientific experiments (in a weightless environment) and placement and retrieval of satellites to programs of military and political significance.

Russian Space Shuttle. The Russians did not acknowledge officially a shuttle program until April 1987, even though rumors of such a program had persisted for several years. The U.S. shuttle lands as a pure glider and must land on the first pass. The Russian shuttle was designed with a set of air-breathing jet engines installed near the tail that would allow cosmonauts to align the craft with the runway or fly to another airstrip. The Russians were reported to observe: "The notion by Americans that the use of a shuttle craft is more economical than other means is not quite so. We plan to use our shuttle in parallel with other craft in our stable of vehicles." For obvious reasons as of the end of 1993, many of the Russian space programs are best described as temporarily indeterminate.

Numerous other satellites are described in other articles in this encyclopedia. Consult alphabetical index.

Additional Reading

Abelson, P. H.: "Earth Observations from Space," *Science*, 901 (May 26, 1989).
Bjerklie, D.: "The Electronic Transformation of Maps," *Technology Review (MIT)*, 54 (April 1989).
Burrows, W. E.: "Deep Black Space Espionage and National Security," Random House, New York, 1986.
Canby, T. Y., and J. Schneeberger: "Satellite Rescue (LDEF)," *National Geographic*, 106 (November 1991).
Collins, M., and C. Pattiaratchi: "The Seas Viewed from Sensors in Space," *University of Wales Review*, 6 (March 1987).
Corcoran, E.: "The U.K., U.S., and France May Join Forces in Remote Sensing," *Sci. Amer.*, 72 (March 1989).
Drury, S. A.: "Guide to Remote Sensing: Interpreting Images of the Earth," Oxford University Press, New York, 1990.
Ford, J. P., et al.: "Faults in the Mojave Desert, California, as Revealed on Enhanced *Landsat* Images," *Science*, 1000 (May 25, 1990).
Golden, F.: "A Catbird's Seat on Amazon Destruction," *Science*, 201 (October 13, 1989).
Hamilton, D. P.: "Hard Start for UARS," *Science*, 19 (July 5, 1991).
Holden, C.: "UARS Launches Earth Mission," *Science*, 1352 (September 20, 1991).
Johnson, N. I.: "The Soviet Year in Space, 1989," Teledyne-Brown Engineering, Colorado Springs, Colorado, 1990.
Kaufman, W.: "Radar Imaging: Forest X-Ray," *Amer. Forests*, 46 (September–October 1990).
Kerr, R. A.: "Why Bigger Isn't Better in Earth Observation," *Science*, 1481 (September 27, 1991).
Krepon, M., et al.: "Commercial Observation Satellites and International Security," St. Martin's Press, New York, 1990.
Limaye, S. S., et al.: "Satellite Observations of Smoke from Oil Fires in Kuwait," *Science*, 1536 (June 14, 1991).
Mack, P. E.: "Viewing the Earth: The Social Construction of the Landsat Satellite System," MIT Press, Cambridge, Massachusetts, 1990.
Marshall, E.: "Space Cameras and Security Risks (SPOT French Satellite)," *Science*, 472 (January 27, 1989).
Marshall, E.: "Landsats: Drifting Toward Oblivion?" *Science*, 999 (February 24, 1989).
Marshall, E.: "Bringing NASA Down to Earth," *Science*, 1248 (June 16, 1989).
Marshall, E.: "A Military Landsat?" *Science*, 907 (May 17, 1991).
Mims, F. M., III: "A Remote-Control Camera that Catches the Wind and Captures the Landscape," *Sci. Amer.*, 126 (October 1990).
Murr, L. E., et al.: "Electron Optical Tools Aid Studies of LDEF Specimens," *Advanced Materials & Processes*, 45 (November 1991).
Platt, T., and S. Sathyendranath: "Oceanic Primary Production: Estimation by Remote Sensing at Local and Regional Scales," *Science*, 1613 (September 23, 1988).
Richelson, J.: "U.S. Keyhole Spy Satellite Program," Harper, New York, 1990.
Richelson, J. T.: "The Future of Space Reconnaissance," *Sci. Amer.*, 38 (January 1991).
Siuru, B.: "Mission to Planet Earth," *Sensors*, 48 (April 1992).
Spencer, R. W., and J. R. Christy: "Precise Monitoring of Global Temperature Trends from Satellites," *Science*, 1558 (March 30, 1990).
Thelin, G. P., and R. J. Pike: "Landforms of the Counterminous United States," U.S. Geological Survey, Distribution Branch, Federal Center, Denver, Colorado (*Box 25286*), 1992.
Van Sant, T. et al.: "The Earth—From Space: A Satellite View of the World," Spaceshots, Inc., Manhattan Beach, California, 1990.
Vasyutin, V. V., and A. A. Tishchenko: "Space Coloristics," *Sci. Amer.*, 84 (July 1989).
Zimmerman, P. D.: "Photos from Space," *Technology Review (MIT)*, 47 (May 1988).

Note: For the scholar of space satellite history, reference to the list given in the 7th Edition of this encyclopedia also is suggested.

SATURATED EDIBLE OILS. See **Vegetable Oils (Edible).**

SATURATED VAPOR. A vapor whose temperature corresponds to the boiling temperature at the pressure existing on it. A vapor is saturated when its temperature is a function of its pressure alone. A saturated vapor may be wet or dry; the term does not necessarily imply a wet vapor. A vapor of 100 percent quality, having no superheat, is said to be dry and saturated.

The physical attributes of saturated steam are pressure, temperature, volume, enthalpy, and entropy. These are always given for steam that is dry and saturated, leaving the reader to apply the quality factor when it occurs. The increase of volume on vaporization and the latent heat of evaporation are present in wet steam to the extent of the percent dryness of the steam. One of the important entries in the saturated steam table is that for atmospheric pressure. At 14.7 psi (1 atmosphere) absolute pressure, the saturation temperature of steam is 212°F (100°C). The heat contained in it as a boiling liquid is 180 Btu (45.4 kilogram-calories) (above 32°F (0°C)), and its latent heat of evaporation is 970.2 Btu/pound (539.1 kilogram-calories/kilogram).

SATURATION. The state of being satisfied, or replete, or the action of bringing about that state. Following are specific uses of this term, some applying to single substance, entity, or region, and others to relations between more than one:

1. The condition in which the partial pressure (i.e., the pressure of a single component of a gaseous mixture, according to Dalton's Law) of any fluid constituent is equal to its maximum possible partial pressure under the existing environmental conditions, such that any increase in the amount of that constituent will initiate within it a change to a more condensed state. In molecular-kinetic terms, saturation is attained when the rate of return of molecules of a substance from the dissolved liquid or vapor phase to the more condensed parent phase is exactly equal to the rate of escape of molecules from the parent phase. In meteorology, the concept of saturation is applied, almost exclusively, to water vapor as a constituent of the atmosphere.
2. The term applied to the maximum current that will pass through a gas under definite conditions of ionization.
3. The attribute of any color perception possessing a hue, that determines the degree of its difference from the achromatic color perception most resembling it.
4. In a nuclear reactor, the maximum activity obtainable by activation in a definite flux.
5. The maximum magnetization (or the maximum permanent magnetization) of which a body or substance is capable.
6. The process or condition of dissolving in a solvent all of a solute that the solvent can absorb, under equilibrium conditions at a given temperature.
7. The complete neutralization of an acid or base.

SATURATION (Atmospheric). See **Precipitation and Hydrometeors.**

SATURATION CURRENT. The ionization current that results when the applied potential is sufficient to collect all ions; the maximum current that will pass through a gas under definite conditions of ionization. The saturation current is a measure of the charge carried by the ions produced in each second, and hence may be used as a measure of the radioactivity of a substance.

SATURN. Sixth planet from the Sun, Saturn is unique in the solar system in that it is the only planet lighter than water, with a density of about 0.7 gram per cubic centimeter. Saturn is the second largest planet in the solar system, with a volume 815 times that of Earth, but with a mass only 95.2 times greater. Like Jupiter, Saturn's rapid rotation has caused the planet to be flattened at its poles. The equatorial radius is 60,330 kilometers (37,490 miles), while the polar radius is smaller—54,000 kilometers (33,554 miles). The surface gravity of Saturn is 1.14 (Earth = 1.0). Saturn requires 29.46 Earth-years to complete one orbit around the Sun. Although a Saturnian year is long, its days are short, lasting only 10 hours, 39 minutes, 24 seconds, as first determined by *Voyager 1* and later reconfirmed by *Voyager 2*. See also **Voyager Missions to Jupiter and Saturn.**

In its slow orbit around the Sun, Saturn is perturbed by the other planets, notably Jupiter, so that its orbital path is not strictly elliptical. The planet wavers in its distance from the Sun in a region of between 9.0 AU and 10.1 AU. (One AU or astronomical unit equals distance from Sun to Earth, i.e., 149,597,860 kilometers; 92,955,806.8 miles.)

Saturn receives only about one-hundredth as much sunlight as is received by Earth. Like the three other gas giants (Jupiter, Neptune, and Uranus), Saturn does not have a solid surface comparable to that found on the terrestrial planets, but is a huge, multilayered globe of gases, notably a mixture of about 11% (mass) of helium, with nearly all the rest being hydrogen. There may be a small core predominantly of iron and rocky material. Gravity field analyses and temperature-profile measurements suggest that Saturn's core may extend out from the center by about 13,800 kilometers (8,575 miles), making it twice the size of Earth, but extremely small in relation to the huge size of the planet. It has been estimated that the core is probably so compressed that it may contain from 15 to 20 times the mass of Earth. It has been postulated that a layer of electrically conductive metallic hydrogen may surround the core. This form of hydrogen has not been observed on Earth because of the immense pressure required to produce it. The interaction between this inner atmosphere and the outer layers (hydrogen and helium) may explain Saturn's emission of heat. Both Saturn and Jupiter release about twice the amount of energy they receive from the Sun. But scientists observe that the two planets probably produce their energy in different ways. Whereas it is postulated that Jupiter emits energy left over from the gravitational contraction that occurred when the planet was formed (estimated about 4.6 billion years ago), it is suggested that Saturn's heat production may be the result of the separation of hydrogen and helium in the outer layer, with heavier helium sinking through the planet's liquid hydrogen interior. Other statistics are given in the entry on **Planets and the Solar System.** See Figs. 1 and 2.

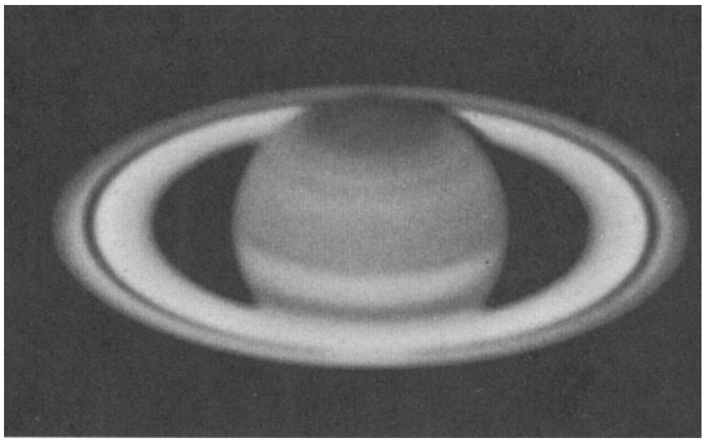


Fig. 1. View of Saturn as seen with an Earth-based telescope (100-inch; 254 cm). (*Hale Observatories.*)

Mission to Saturn. To date, the environs of Saturn have been explored by three U.S. spacecraft, as described in the next several paragraphs. Most of the progress made during the past several years has been the analysis of massive amounts of data sent back to Earth from these former missions. The Hubble space telescope has made comparatively few and fuzzy images of the planet during the early 1990s, largely pertaining to the more recently discovered "Great White Spot" that appears to be representative of periodic disturbances in the planet's atmosphere. This is described toward the end of this article. Considerable research from Earth also has gone forward with microwave radar imaging and is discussed later.

Plans to launch the CRAF/Cassini mission to Saturn are under way. From the outset, this project has faced budgetary competition from the proposed space station. CRAF/Cassini has been proposed as a joint venture with the European Space Agency (ESA). Initially estimated cost was $260 million, part of which will be sponsored by the ESA. The craft is designed to descend to the surface of Titan. The instrumentation, which was selected by the Europeans and NASA in October 1990, will include sensors for measuring the temperature and pressure of Titan's atmosphere, for measuring its turbulence, winds, and electricity, and for analyzing the atmosphere during descent. Cloud pictures are scheduled. Although the probe is not designed to survive a hard impact, it would survive for a few minutes if it impacts a liquid body of methane and ethane. If the project is on time, the orbiter-launched spacecraft is scheduled (originally) to reach Titan by the year 2002.

Past U.S. missions to Saturn include:

(1) *Pioneer 11,* in 1979, flew past the planet and on September 5 of that year, scientists reported the discovery of Saturn's 11th moon orbiting the planet and two more rings of debris encircling the planet. The space probe swept within 13,000 miles (20,917 kilometers) of the planet and then flew by Titan, Saturn's largest moon. Scientists had feared that the spacecraft would be damaged by some of the debris around the planet, but *Pioneer 11* passed through several potential trouble spots unharmed and beamed back a wealth of data to Earth.

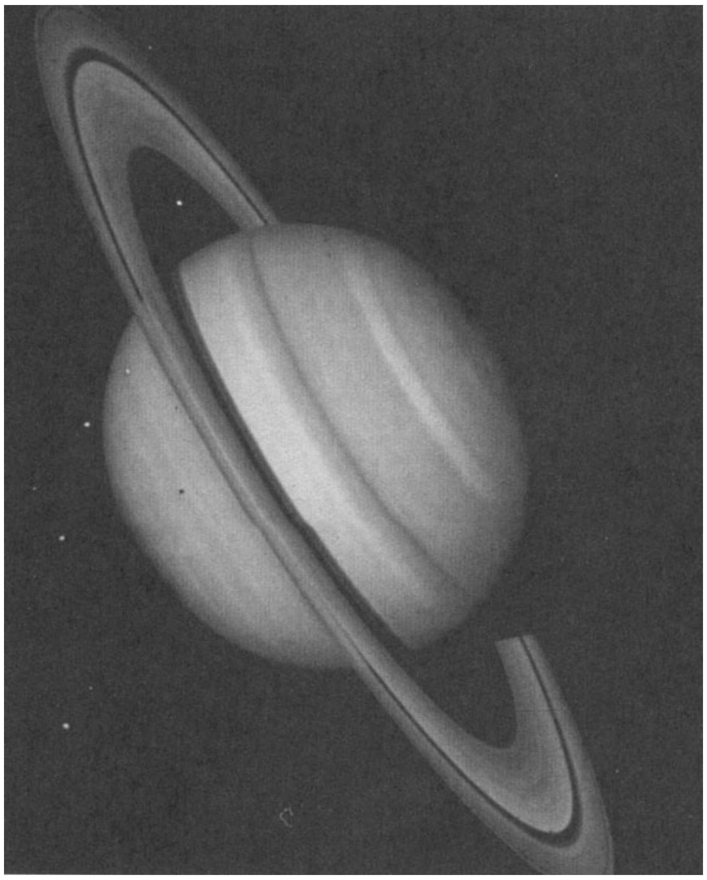


Fig. 2(a). Computer-enhanced photograph of Saturn taken by *Voyager 2* on August 4, 1981 from a distance of 21 million kilometers (13 million miles) on the spacecraft's approach trajectory. Three of Saturn's icy moons are evident at left. They are, in order of distance from the plant: Tethys (1050 kilometers; 652 miles in diameter); Dione (1120 kilometers; 696 miles in diameter); and Rhea (1530 kilometers; 951 miles in diameter). The shadow of Tethys appears on Saturn's southern hemisphere. A fourth satellite, Mimas, is less evident. It is located just a short distance from the planet's limb and above Tethys in the view. The shadow of Mimas appears directly above that of Tethys. In the original color version of this view, the bright and darker bands in both hemispheres of Saturn's weather system appear in pastel yellow hues. (*Jet Propulsion Laboratory*.)

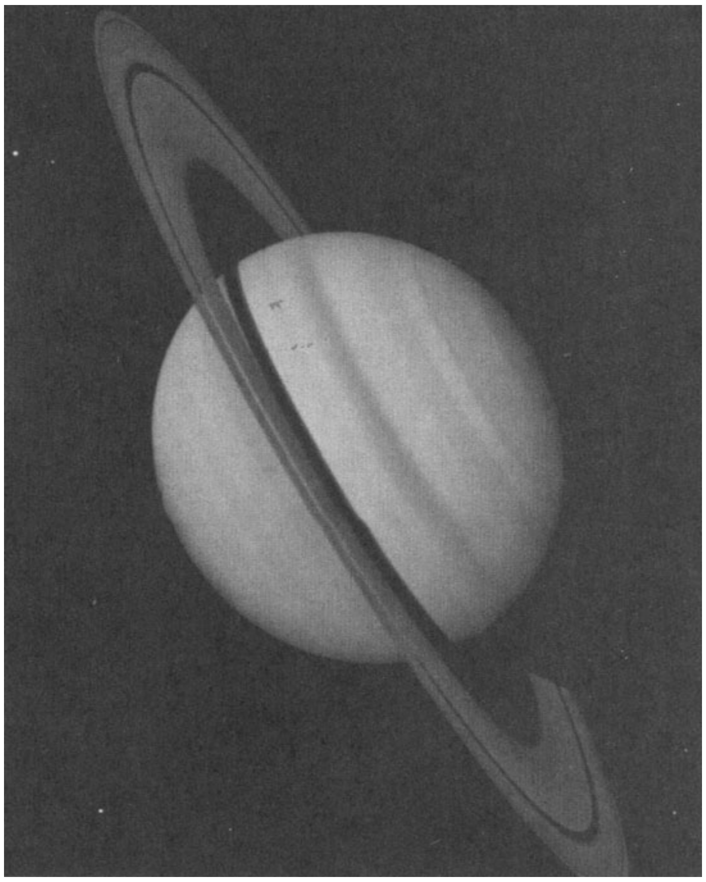


Fig. 2(b). Saturn and its satellites Tethys (upper left), Enceladus (inner left) and Mimas (right of rings) are seen in this mosaic of images taken by *Voyager 1* on October 30, 1980 from a distance of 18 million kilometers (11 million miles). The soft, velvety appearance of the low-contrast banded structure and increased reflection of blue light near the perimeter of the Saturn disk are due to scattering by a haze layer above the planet's cloud deck. Features larger than 350 kilometers (220 miles) are visible. The projected width of the rings at the center of the disk is 10,000 kilometers (6000 miles), which provides a scale for estimating feature sizes on the image. (*Jet Propulsion Laboratory*.)

Much of the information supported the image of Saturn as had been previously determined by Earth-based observations.

(2) *Voyager 1*, in the fall of 1980, viewed the planet after the spacecraft had visited Jupiter early in the spring of 1979. Early pictures of the planet taken during the summer of 1980, as the spacecraft approached Saturn, were almost featureless with exception of the three classic rings that had been studied for years from Earth. As *Voyager 1* approached more closely, it appeared that there were not just three rings, but scores, then hundreds, and finally thousands of thin ringlets. It would turn out that they were not individual rings separated by gaps; some of the variations were caused by the gravitational attraction of nearby satellites pulling millions of particles into motion, spiraling outward across the rings like waves in an ocean. The photos from *Voyager 1* also revealed other puzzling phenomena—dark features that resembled spokes in the bright B-ring. As *Voyager 1* zeroed in on Saturn, more satellites were seen, until a total of six had been found—three from Earth observations and three by *Voyager*. Two that were discovered in *Voyager* images appeared to shepherd the narrow F-ring. Two more, discovered from Earth, had appeared to share the same orbit. Inspection of *Voyager* photos, however, showed that the satellites' orbits are about 31 miles (50 kilometers) apart. A little more calculation yielded the astonishing prediction that, as the two satellites approached each other in January 1982, they would trade orbits and continue on their way, to resume their game of "musical chairs" the next time they approached. A day before *Voyager 1* swept by Saturn, it flew within 2500 miles (4000 km) of the huge satellite Titan and passed directly behind it, making what scientists had predicted would be extremely important observations. Titan was shrouded by a thick, opaque haze that completely obscured its surface from the cameras. But the infrared instrument and the spacecraft radio probed the atmosphere to measure the diameter of the satellite and the thickness, temperature, and composition of its atmosphere. Once beyond Titan, *Voyager 1* flew past Saturn and briefly disappeared behind it. En route to the Earth, the radio signals penetrated Saturn's atmosphere and passed through the rings. Measurements of the way the atmosphere altered the signals, and the rings scattered them, would tell much about the atmosphere and help determine the sizes of particles that make up the rings. *Voyager 1* swept past its targets and took a new course upward from the plane in which the planets orbit the Sun, outward toward the edge of the solar system. Its cameras and its UV and IR instruments were turned off, but other instruments still probe for galactic cosmic rays, the edge of the solar system, and the beginning of interstellar space. The trek of *Voyager 1* is depicted in Fig. 3.

(3) *Voyager 2*, in late summer 1981, made it possible to view Saturn for a second time. Having had an opportunity to study the images taken by *Voyager 1*, scientists decided to arrange the second encounter for a closer study of the planet's rings and its satellites, other than Titan. As the summer of 1981 approached, scientists, having gained experience from *Voyager 1*, were able to set their camera exposures at more exact levels, to cope with the low light levels and the general blandness of Saturn. On this approach, Saturn presented alternating dark and bright bands of clouds and high-speed jet streams. Swirling cloud patterns, which were smaller versions of the large and intense storms seen on Jupiter, were also visible through Saturn's haze layer. *Voyager 2's* cameras zeroed in on the rings, and scientists searched for small satellites in the rings that might cause the multiringed appearance. Those moon-

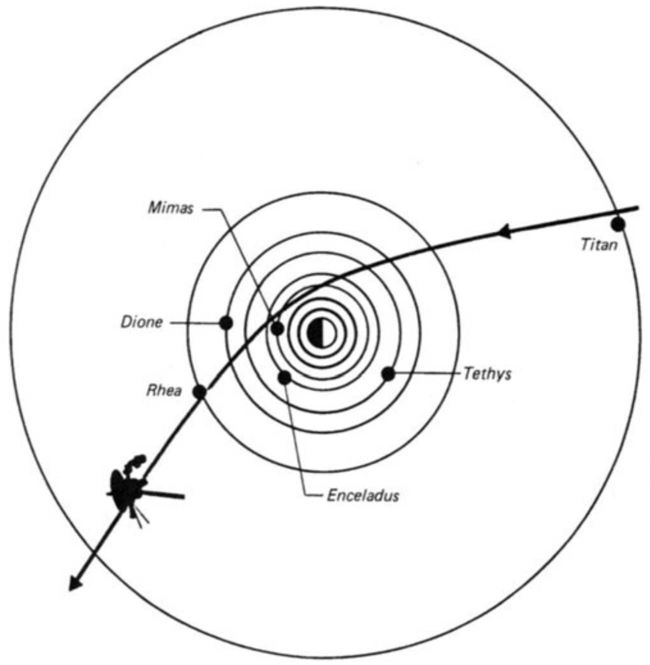


Fig. 3. Trek of *Voyager 1* as it swept past Saturn in the fall of 1980. (*Jet Propulsion Laboratory*.)

lets, some scientists believed, might sweep up material in the rings, creating gaps. *Voyager 2* would soar closer to the rings, and the improved resolution of the pictures should show structures as small as 0.6 mile (1 km) in diameter. One of *Voyager's* most important experiments involved an instrument called a photopolarimeter, which measures light intensity. As *Voyager 2* passed above Saturn, the photopolarimeter detected changes in the starlight's intensity as it was altered by changes in the thickness of the rings. Quick analysis of data showed that the rings' structure was far different from what it appeared to be in the photos.

No region was totally empty of ring particles. The members of the photopolarimeter team have 800,000 samples, each one a 330-foot (100-meter) slice of the rings. Years will be required to complete this analysis. *Voyager 2* also photographed and measured all of the planet's satellites that were then known, a number that had expanded to 17. Upon completion of the sweep by Saturn, *Voyager 2* proceeded on toward Uranus and Neptune. See separate articles on these planets. The trek of *Voyager 2* past Saturn is depicted in Fig. 4.

Highlights of *Voyager's* Scientific Findings. After studying the volumes of data returned by the two *Voyager* spacecraft, although analyses are still in progress, scientists at NASA (National Aeronautics and Space Administration), and Jet Propulsion Laboratory (JPL), with over 30,000 photos of the Saturn system, have selected the following highlights:

1. There is less helium in the top of Saturn's atmosphere than in Jupiter's.
2. Subdued contrasts and color differences (Jupiter and Saturn) are primarily a result of either more horizontal mixing or less production of localized colors on Saturn than on Jupiter.
3. Winds blow at extremely high speeds on Saturn. Near the equator, the *Voyagers* measured winds of about 500 meters per second (1100 miles per hour). The winds blow primarily in an eastward direction.
4. The *Voyagers* found auroralike ultraviolet emissions at midlatitudes on Saturn, and auroras at higher latitudes.
5. The *Voyagers* discovered radio emissions from the planet with which they determined the length of Saturn's day to be 10 hours, 39 minutes, 24 seconds.
6. The complicated structure in Saturn's rings appears to be caused, at least in part, by density waves, which are created by gravitational interactions with several of the inner satellites. Few clear gaps exist anywhere in the rings.
7. Radial, spokelike features were discovered in the rings and remain poorly understood.
8. Titan, Saturn's largest satellite, has an atmosphere composed of nitrogen, methane, and several organic compounds, including hydrogen cyanide.
9. Titan's surface atmospheric pressure is 1.6 bars (60% greater than the surface pressure on Earth).
10. The temperature at the surface of Titan is 95 K ($-288°F$; $\sim -172°C$). Methane, therefore, possibly plays much the same role on Titan as water does on Earth.
11. Saturn's regular satellites appear to be composed primarily of ice. Phoebe, the outermost, is believed to be a captured asteroid.
12. Many new satellites have been discovered at Saturn, some from Earth-based observations, others by the two *Voyagers*. There may be over twenty. Scientists expect to find more. As indicated by Fig. 5, Saturn's satellites, with exception of the huge Titan, are smaller than Earth's moon.
13. The size of Saturn's magnetosphere, like Jupiter's, is controlled by external pressure of the solar wind.

More details of some of the foregoing observations are developed in this article.

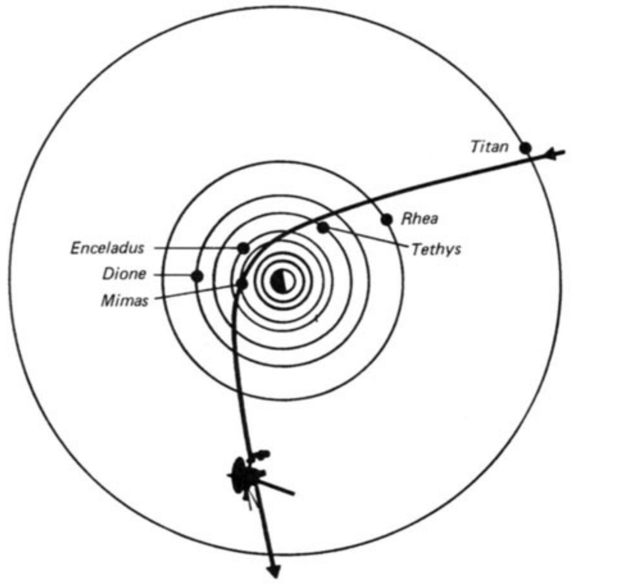


Fig. 4. Trek of *Voyager 2* as it swept past Saturn in late-summer of 1981. (*Jet Propulsion Laboratory*.)

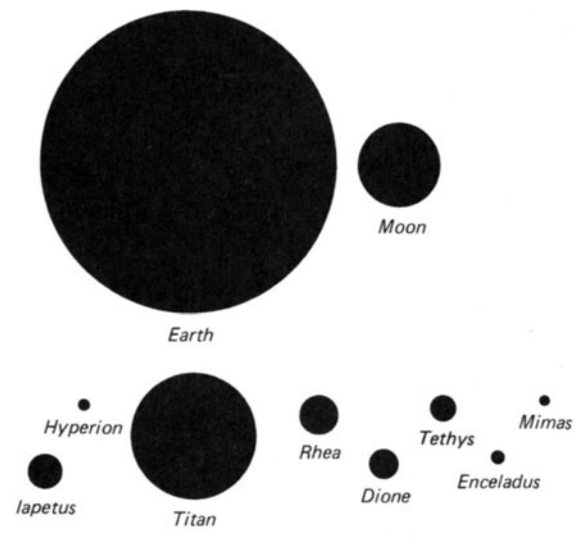


Fig. 5. Size comparison (diameter) of Saturn's major satellites with Earth and its moon. (*Jet Propulsion Laboratory*.)

Observations of Saturn by Early Astronomers. To the naked eye, Saturn appears like one of the brighter stars, but with the absence of twinkling. In a telescope, the planet itself has a belted appearance quite similar to that of Jupiter, but without as many distinctive surface features as are seen on the larger planet. Although Galileo realized that the planet had some unusual features, he did not live to know that what he saw was a ring system. Galileo's sketch, made in 1610, is shown in Fig. 6. Huygens first described the rings with reasonable accuracy for this period in 1650. Cassini, by 1675, identified a ring system that was divided into two parts by a dark line (now known as the *Cassini division*). Cassini also noted some belt or zone demarcations, as indicated by his sketch shown in Fig. 7. The first successful photo of Saturn, clearly showing the ring system, was not made until 1883. In 1895, Keller suggested that the rings were a swarm of particles in nearly independent orbits. Bond, in the late 1800s, found a second division of the rings and a third or "crepe" ring close to the planet itself.

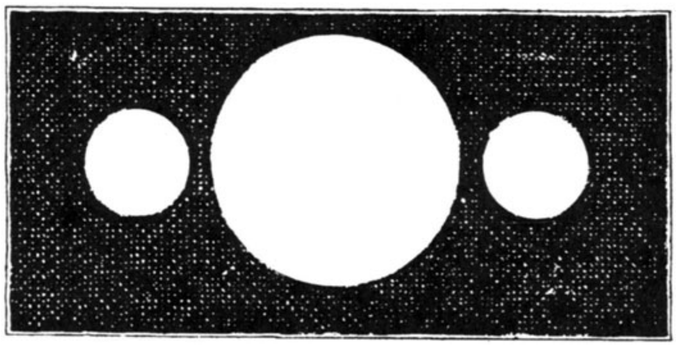


Fig. 6. Sketch of Saturn by Galileo (1610). When Galileo first focused his telescope on Saturn, he realized that the appearance of the planet was unusual, but he did not ascertain its real character because the power of his telescope was far too low. He believed he was looking at three globes, one large and two small, which seemed to change slowly in appearance. In 1655, Huygens, after years of observing the planet, finally realized that these projections were actually a flat ring slightly separated from the main globe.

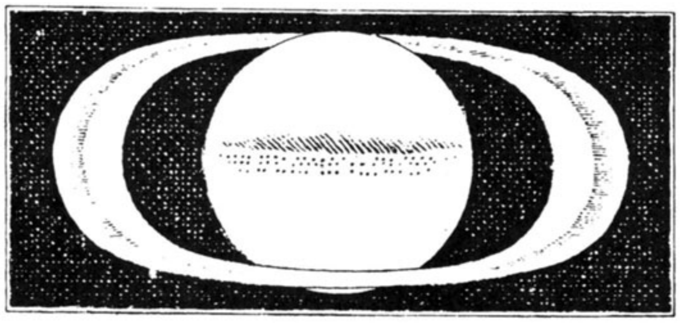


Fig. 7. Sketch of Saturn by Cassini (1675). Cassini found the first breach in the supposedly solid, rigid, and opaque ring when he discovered that it was divided into two parts by a dark line, now known as Cassini's division. In later years, Cassini also detected some of Saturn's moons. The earliest successful photograph of Saturn was taken in 1883 by Andrew Common. In 1895, James Keeler suggested that the rings are in fact a swarm of particles in near-independent orbits.

Atmosphere of Saturn

Saturn has cloud bands similar to those of Jupiter, although they are more difficult to see and contrast less with the planetary disk. Images of Saturn confirm its rather bland appearance. The blandness may be a result of lower temperatures and reduced chemical and meteorological activity compared with Jupiter; or the presence of a relatively permanent and uniform high-altitude haze. As shown by Fig. 8, the principal features of Saturn's visible surface are stripes that parallel the equator. There are dark belts and light zones which have been seen continuously over two centuries of observation. Although *Voyager 1* saw only a few markings, *Voyager 2* saw many, including long-lived ovals, tilted features in east-west shear zones and others similar to, but smaller than, the features on Jupiter. One white oval on Saturn was noted to be 7000 by 5000 kilometers (4000 by 3000 miles), with a 100-meter-per-second (200-mile-per-hour) circumferential wind. See Figs. 9 and 10. Winds appear to blow at extremely high speeds on Saturn. Near the equator, the *Voyagers* measured winds about 500 meters per second (1100 miles per hour). The winds blow primarily in an easterly direction. The strongest winds are found near the equator, and velocity falls off uniformly at higher latitudes. At latitudes greater than 35°, the winds alternate eastward and westward as the latitude increases. The marked dominance of eastward jet streams indicates that winds are not confined to the cloud layer, but must extend inward at least 2000 kilometers (1240 miles). Furthermore, measurements by *Voyager 2* showing a striking north-south symmetry has led some scientists to suggest that the winds may extend from north to south clear through the interior of the planet.

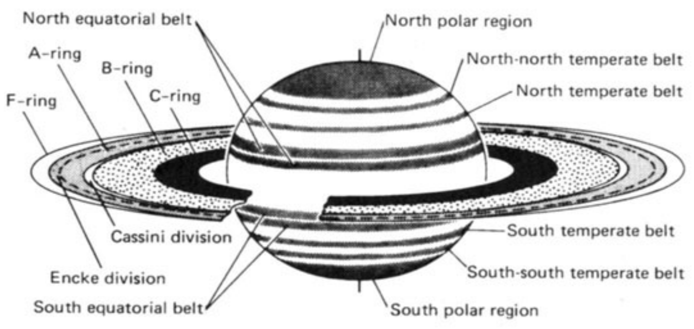


Fig. 8. Principal visual features of Saturn.

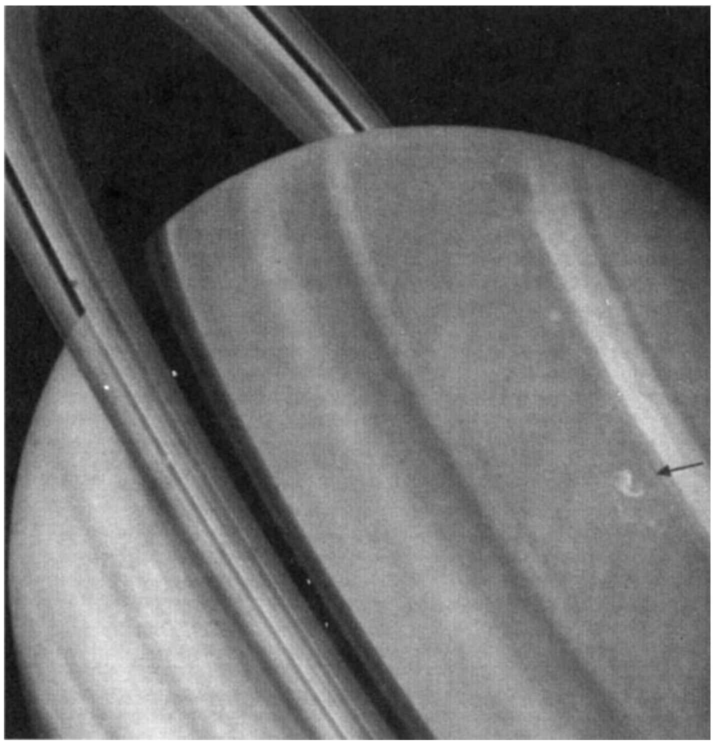


Fig. 9. View of Saturn and its ring system returned by *Voyager 2* on August 11, 1981 when the spacecraft was 13.9 million kilometers (8.6 million miles) away from the planet and approaching it at about 1 million kilometers (620,000 miles) per day. The ring system's shadow is clearly cast in the equatorial region. Storm clouds and small-scale spots in the mid-latitudes are apparent. The so-called "ribbonlike" feature in the white cloud band marks a high-velocity jet at about 47° N. At this location, the westerly wind speeds are about 150 meters per second (330 miles per hour). The banding on Saturn extends toward both poles. (*Jet Propulsion Laboratory.*)

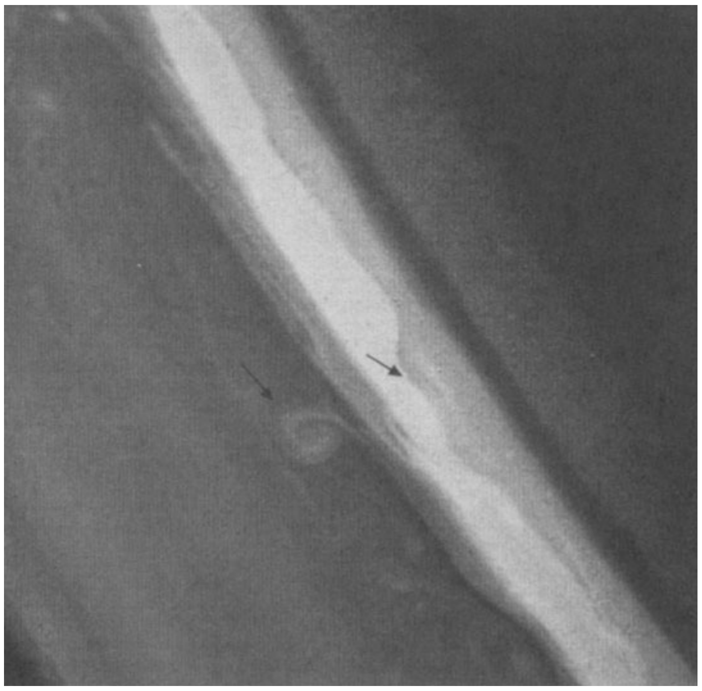


Fig. 10. In this *Voyager 2* observation of Saturn's northern mid-latitudes is seen a strangely curled cloud attached by a thin ribbon to the bright white cloud region to the north. The cloud was monitored for several rotations around the planet. It appears to be forming a closed loop. Other discrete clouds are seen to the east. Also evident is the ribbonlike structure in the white cloud region. The spacecraft took this image on August 16, 1981 from a distance of 9.3 million kilometers (5.8 million miles), when the smallest feature seen was about 90 kilometers (56 miles) across. (*Jet Propulsion Laboratory.*)

When *Voyager 2* flew behind Saturn, its radio beam penetrated Saturn's atmosphere, measuring the upper-atmosphere temperature and density. Minimum temperatures of about 82 Kelvin (−312°F; −191°C) were measured at the 70-millibar level (surface pressure on Earth is about 1000 millibars). The temperature increased to 143 K (−202°F; −130°C) at the deepest levels probed—about 1200 millibars. Temperatures near the north pole were about 10°C (18°F) colder at the 100-millibar level than temperatures at mid-latitudes. Scientists believe the difference may be a seasonal effect.

The *Voyagers* found auroralike ultraviolet emissions of hydrogen at mid-latitudes (above 65°). Scientists have suggested the high-latitude auroral activity leads to formation of complex hydrocarbon molecules that are carried toward the equator. The mid-latitude auroras, which normally occur only in sunlit regions, remain a puzzle; bombardment by electrons and ions, known to cause auroras on Earth, occurs primarily at high latitudes.

Saturn's Ring System

The alphabetical letter designations of Saturn's rings (Fig. 8) are based upon the chronology of discovery. The letters do not consistently relate to their positions relative to the planet. The A ring is the outermost ring visible with small telescopes. The B ring, the brightest ring, lies inside the A ring and is separated from it by Cassini's division. The C ring, barely visible with small telescopes, lies inside the B ring. The Cassini division between the B and A rings is the most prominent gap between the rings and is easily visible with Earth-based telescopes. A thin space within the outermost edge of the A ring is the Encke division. This was first identified by Pioneer spacecraft. The average thickness of the rings is estimated at no greater than 16 kilometers (10 miles).

Perhaps the greatest surprises and the most perplexing puzzles encountered by the Voyagers are the rings. *Voyager 1* found a great deal of unexpected structure in the classical A, B, and C rings. One suggestion was that the structure might be unresolved ringlets and gaps. Ring photos by *Voyager 1* were of lower resolution than those of *Voyager 2*, and scientists at first suggested that gaps might be created by tiny satellites orbiting within the rings and sweeping out bands of particles. One such large gap was detected at the inner edge of the Cassini division.

Cameras and instruments of much greater resolution than used in the past will be required on future missions to glean greater information pertaining to the variety and characteristics of particles actually making up what appears to be hundreds of thousands of rings. As some scientists believed, *Voyager* observations no longer support the concept that the gravity of small moonlets within the rings channel the trillions of ice chunks that compose the rings into so many narrow traffic lanes. The only moonlets of this nature detected were the "shepherds" discovered by *Voyager 1* which appear to maintain the thin outer F ring in place. As pointed out by Loudon (1981), there is more matter associated with the rings than would be expected from data that are converted into photographs. As the *Voyager 2* streaked through the ring system's thin 500-foot (150-meter) thick outer extension at a relative speed of 8 miles (13 kilometers) per second, the craft was bombarded by thousands of dust-sized particles not for 500 feet, but for about 1000 miles (1600 kilometers).

As observed by the Stanford University radio science team when *Voyager 1* passed behind Saturn and a radio signal was beamed to earth through the rings, useful information was gained concerning the ringlets and the distribution of particle sizes within the ringlets. It was observed that in some ringlets, the fine particles are denser on the inner and outer edges than they are in the middle. The rings appear to consist of a tenuous background sheet of material, with empty gaps and denser ridges superimposed on it. It was noted that a typical ridge has a very sharp edge. The investigators, by comparing the attenuation of the micro-wave signal at two wavelengths, were able to estimate the size distribution of ring particles. It was found that there are about 120 to 200 particles per square kilometer in the size range of 9 to 11 meters. Very few larger particles were detected. The numbers as well as uncertainties increased rapidly with diminishing size. It was found in the 3- to 5-meter particle range, that there are from 1700 to 6000 particles per square kilometer. The investigators noted that there may be millions that are smaller than that; or there may be none; the radio data are consistent with either interpretation.

Voyager 2 measurements provided much of the data required for an understanding of structure. Higher-resolution photos of the inner edge of the Cassini division showed no sign of satellites down to about 5–9 kilometers (3–6 miles). No systematic searches were conducted in any of the other ring gaps. *Voyager 2's* photopolarimeter provided more surprises. The instrument measured changes in starlight from Delta Scorpii as the spacecraft flew above the rings when the starlight passed through the rings. The photopolarimeter was capable of resolving structure smaller than 300 meters (1000 feet). The star-occulation experiment showed that few clear gaps exist anywhere in the rings. The structure in the B ring instead appears to be variations in density of ring material, probably caused by traveling density waves or other, stationary forms of waves. Density waves are formed by the gravitational effects of Saturn's satellites. They propagate outward from positions where the ring particles orbit Saturn in harmony with the satellites. Resonant points are locations where a particle orbits Saturn in one-half or one-third the time required by a satellite, such as Mimas. For example, at the 2:1 resonant point with the satellite 1980 S1, a series of outward propagating density waves has characteristics that indicate there are about 60 grams of material per square centimeter of ring area and that the velocity of the particles relative to one another is about one millimeter per second. The small-scale structure of the rings may therefore be largely transitory, although larger-scale features, such as the Cassini and Encke divisions, appear to be more permanent. The edges of the rings where the few gaps exist are so sharp that the ring must be less than about 200 meters (650 feet) thick at such points. See Fig. 11–14.

In almost every case where clear gaps do appear in the rings, eccentric ringlets are found. These seem to show variations in brightness. In some cases, the differences are due to clumping or kinking; in others to nearly complete absence of ring material. Some scientists believe the only plausible explanation for the clear regions and kinky ringlets within them is the presence of nearby, yet undetected moonlets.

Two separate, discontinuous ringlets were found in the A-ring gap, about 73,000 kilometers (45,000 miles) from Saturn's cloud tops. At

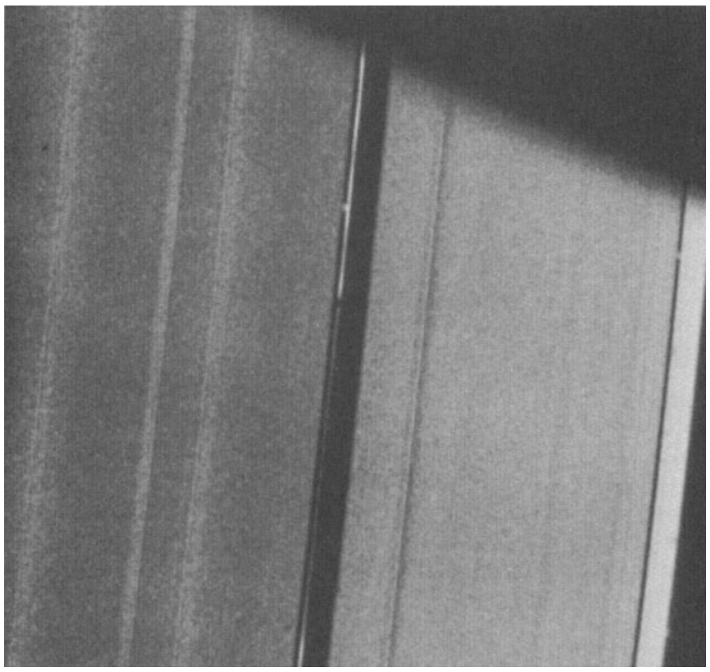


Fig. 11. *Voyager 2* cameras acquired this view of Saturn's A ring on August 26, 1981 from a distance of 227,800 kilometers (141,500 miles). This view of the ring's outer edge shows a small bright, clumpy ring within the Encke gap (center of the image) that exhibits kinks reminiscent of those observed in the F ring by *Voyager 1* in 1980, but not by *Voyager 2*. *Voyager 1* saw two similar clumpy rings in this region at much lower resolution. Also visible are a bright ringlet at the very outer edge of the A ring and several bright wave patterns in the Encke region. The small bright patch on the inner edge of the Encke gap near the ring is an artifact of processing. (*Jet Propulsion Laboratory*.)

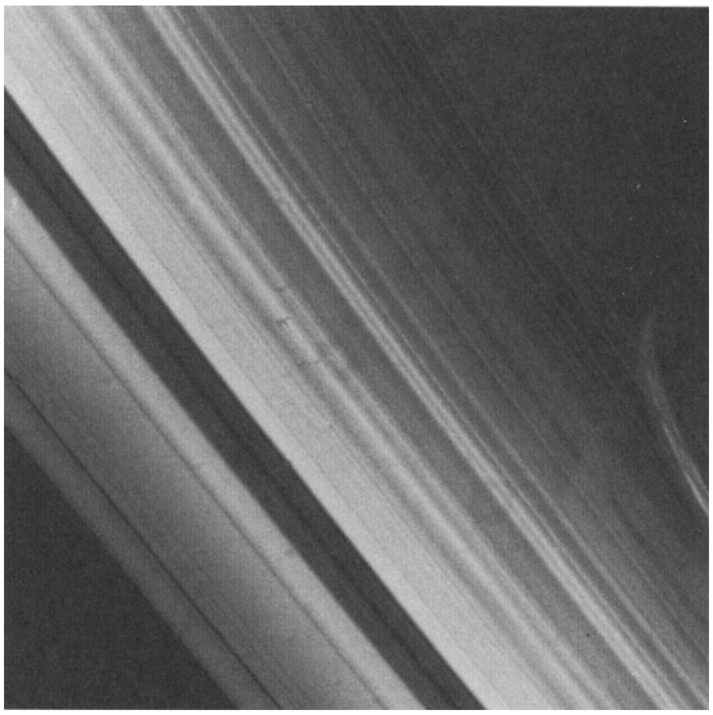


Fig. 13. *Voyager 2* returned this high-resolution view of Saturn's rings on August 23, 1981 at a range of 3.3 million kilometers (2 million miles). The planet's limb is visible through the C ring and the inner part of the B ring. The ring shadows have been obscured by the bright band of light, evident on Saturn's surface, that passed through the more transparent Cassini division. The Cassini division is the darker gap that extends from the lower center of the image to the upper left; it is about 5000 kilometers (3100 miles) wide. Many bright and dark ringlets are seen throughout the complex ring system. (*Jet Propulsion Laboratory*.)

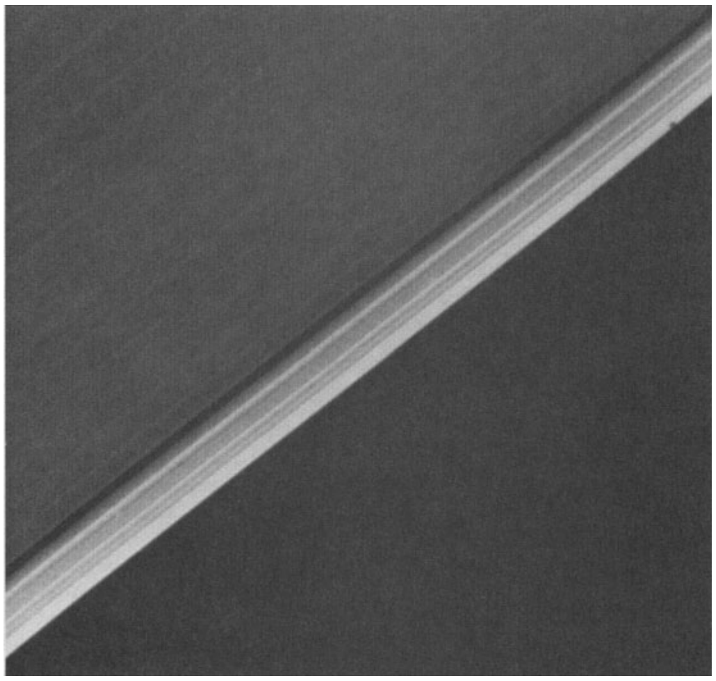


Fig. 12. The outer edge of Saturn's A ring is detailed in this image obtained by *Voyager 2* on August 26, 1981, just 30 minutes before closest approach, at a range of about 51,000 kilometers (31,700 miles). Resolution in this wide-angle view is about 4 kilometers (2.5 miles). The many sharp linear features parallel to the ring edge are most probably locations where ring particles are in resonance with one of the several small nearby satellites of about 100-kilometer (62-mile) radius discovered by *Voyager 1*. Such regions are likely characterized by more intense particle collisions and a greater density of small "chips." The entire outer band of the A ring has different scattering properties and therefore different particle characteristics than the main body of the rings to the upper left. (*Jet Propulsion Laboratory*.)

high resolution, at least one of these ringlets was noted to have multiple strands.

Saturn's F ring was discovered by *Pioneer 11* in 1979. Photos of that ring taken by *Voyager 1* showed three separated strands that appeared twisted or braided. *Voyager 2* found five separate strands in a region that had no apparent braiding, but did reveal apparent braiding in another region. The photopolarimeter found that the brightest of the F-ring strands was subdivided into at least ten strands. The twists in the F ring are believed to originate in gravitational perturbations caused by the two shepherding satellites, 1980 S26 and 1980 S27. See Fig. 15.

Clumps of material in the F ring appear fairly uniformly distributed around the ring every 9000 kilometers (5600 miles), a spacing that coincides with the relative motion of F-ring particles and the shepherding satellites in one orbital period. By analogy, scientists suggest that similar mechanisms may be operating for irregular ringlets that exist in gaps in the main ring system.

The spokes found in the B ring appear only at radial distances between 43,000 kilometers (27,000 miles) and 57,000 kilometers (35,000 miles) above Saturn's clouds. Some spokes, those that are narrow and have a radial alignment, may be recently formed. The broader, less radial spokes appear to have formed earlier than the narrow examples and seem to follow Keplerian orbits—individual areas rotate at speeds governed by distances from the center of the planet. In some cases, scientists suggest that they see evidence that new spokes are reprinted over older ones. Formation of the spokes is not restricted to regions near the planet's shadow. As both Voyagers approached Saturn, the spokes appeared dark against a bright ring background. As the spacecraft departed, the spokes appeared brighter than the surrounding ring areas, indicating they backscatter the reflected sunlight more efficiently. See Fig. 16.

Spokes are also visible at high phase angles in light reflected from Saturn on the unilluminated underside of the rings. This suggests, according to some scientists, that charging of the small particles by photoionization alone may not be responsible for levitating them above the bulk of the ring material.

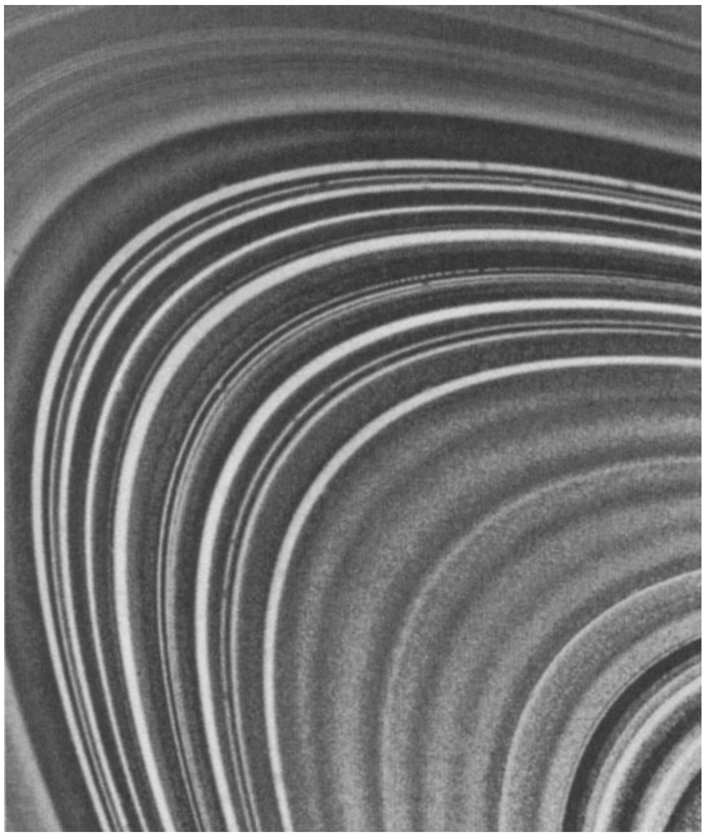


Fig. 14. This view focusing on Saturn's C ring (and to a lesser extent, the B ring at top and left) was compiled from three separate images taken through ultraviolet, clear, and green filters. The images were made on August 23, 1981 when *Voyager 2* was 2.7 million kilometers (1.7 million miles) from the planet. More than 60 bright and dark ringlets are evident; the small, bland squares are caused by the removal of reseau (reference) marks during processing. In general, C-ring material is very bland and gray, the color of dirty ice. Color differences between this ring and the B ring indicate different surface compositions for the material composing these complex structures. (*Jet Propulsion Laboratory.*)

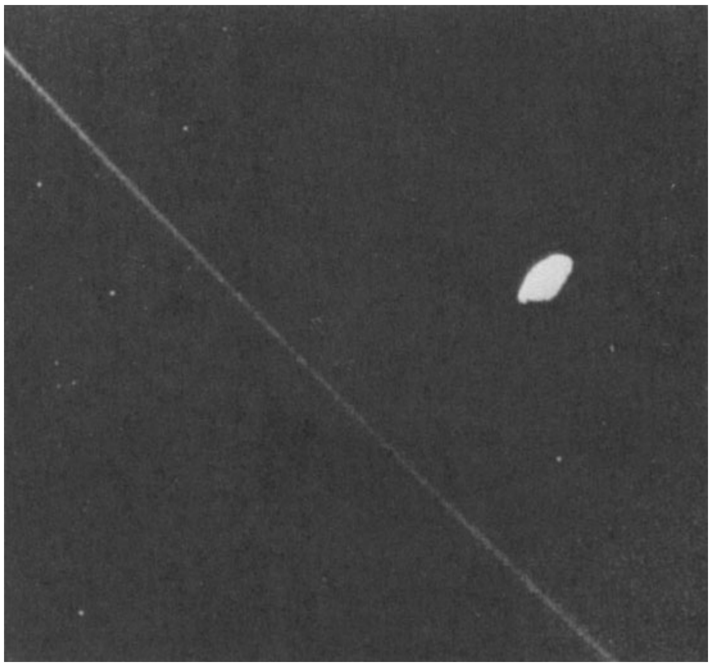


Fig. 15. Saturn's F ring and its inner shepherding satellite (1980S27) are pictured in this closeup *Voyager 2* image acquired on August 25, 1981 from a range of 365,000 kilometers (227,000 miles). Features as small as 6 kilometers (3.7 miles) are visible. The satellite is elongated and irregular, with its longest axis pointing toward the center of Saturn (toward upper right of view). As seen here, the F ring is thin and does not show the multiple, braided structure noted by *Voyager 1* in 1980. There is no indication of a band or kink in the ring at its closest point to the shepherd; such a feature would be consistent with some of the theories advanced on the formation of the braids. (*Jet Propulsion Laboratory.*)

Another challenge faced by investigators is an understanding of the observation that even general dimensions of the rings do not seem to remain true at all positions around the planet. The distance of the B ring's outer edge, near a 2:1 resonance with Mimas, varies by at least 140 kilometers (87 miles). Furthermore, the elliptical shape of the outer edge does not follow a Keplerian orbit, since Saturn is at the center of the ellipse, rather than at one focus. Although the gravitational effects of Mimas are most likely responsible for the elliptical shape, present theory predicts a somewhat smaller magnitude than that which was observed.

Voyager 1 measured radio waves that originate in sporadic electric discharges. The source of these discharges is still unknown. It is possible that they may originate in the rings. *Voyager 2* measured similar discharges, but at a rate only 10% that of the findings of *Voyager 1*, and with a different polarization.

Magnetosphere

The size of Saturn's magnetosphere is determined by external pressure of the solar wind. When *Voyager 2* entered the magnetosphere, the solar wind pressure was high and the magnetosphere extended only 19 Saturn radii (1.1 million kilometers; 712,000 miles) in the Sun's direction. Several hours later, however, the solar wind pressure dropped and Saturn's magnetosphere ballooned outward over a six-hour period. It apparently remained inflated for at least three days, since it was 70% larger than when *Voyager 2* crossed the magnetic boundary on the outbound leg.

Unlike all other planets whose magnetic fields have been measured, Saturn's field is tipped only about one degree relative to the rotation

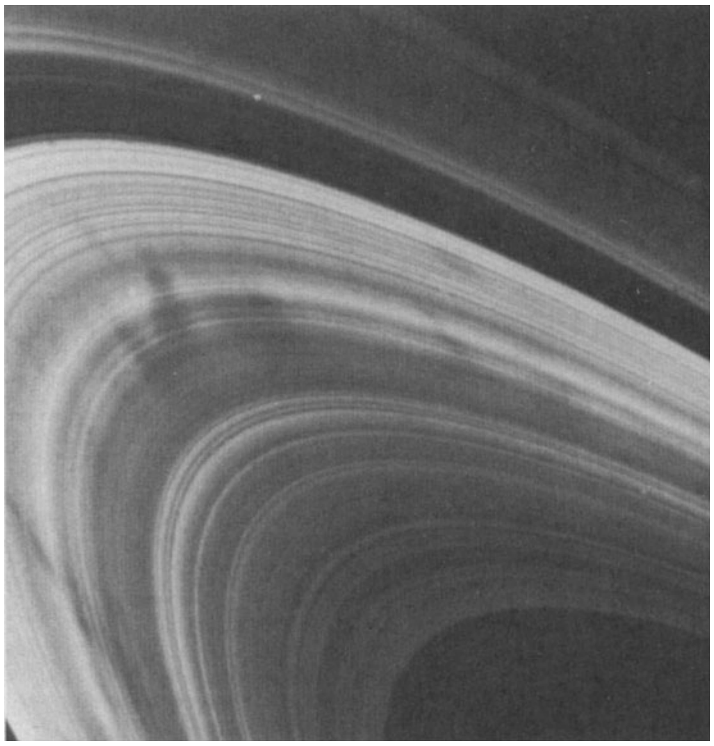


Fig. 16. *Voyager 2* obtained this high-resolution picture of Saturn's rings on August 22, 1981 when the spacecraft was 4 million kilometers (2.4 million miles) distant from the planet. Evident here are the numerous "spoke" features in the B ring; their very sharp, narrow appearance suggests short formation times. Some scientists believe that electromagnetic forces are responsible in some way for these features, but no detailed theory has been formulated. Pictures such as this and analyses of *Voyager 2*'s spoke motion pictures may reveal additional clues about the origins of these complex structures. (*Jet Propulsion Laboratory.*)

poles. That rare alignment was first measured by *Pioneer 11* in 1979 and was later confirmed by *Voyager 1*.

Several distinct regions have been identified within Saturn's magnetosphere. Inside about 400,000 kilometers (250,000 miles), there is a torus of H^+ and O^+ ions, probably originating from water ice sputtered from the surfaces of the satellites Dione and Tethys. These ions are positively charged atoms of hydrogen and oxygen that have lost one electron. Strong plasma-wave emissions appear to be associated with the inner torus.

At the outer regions of the inner torus, some ions have been accelerated to high velocities. In terms of temperatures, such velocities correspond to 400 to 500 million degrees Kelvin.

Outside the inner torus is a thick sheet of plasma that extends out to about 1 million kilometers (620,000 miles). The source for material in the outer plasma may be Saturn's ionosphere, Titan's atmosphere, and the neutral hydrogen torus that surrounds Titan between 500,000 kilometers (300,000 miles) and 1.5 million kilometers (1 million miles).

Radio emissions from Saturn had changed between the encounters of *Voyager 1* and *2*. *Voyager 2* detected Jupiter's magnetotail as it approached Saturn in the winter and early spring of 1981. Soon afterward, when Saturn was believed to be bathed in the Jovian magnetotail, the ringed planet's kilometric radio emissions were also undetectable.

During portions of *Voyager 2's* Saturn encounter, kilometric radio emissions again were not detected. The observations are consistent with effects caused by Jupiter's magnetotail, although *Voyager* scientists admit of no direct evidence that the shutdown of Saturn's natural radio signals was caused by Jupiter's magnetotail.

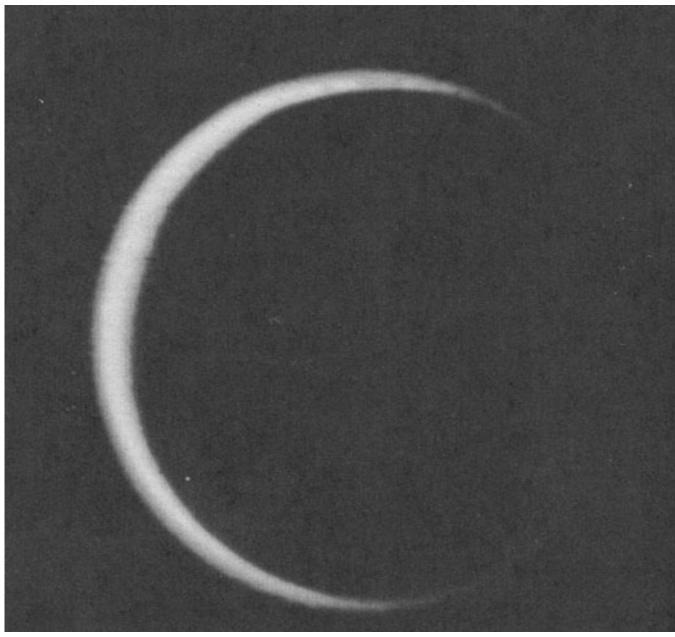


Fig. 17. Wide-angle view of the night side of Titan obtained by *Voyager 2* on August 25, 1981 at a high phase angle of 154° and a range of 907,000 kilometers (563,000 miles). Green and violet images were combined to make this photograph. The result is a view of the extended atmosphere of this satellite, the bright orangish ring being caused by the atmosphere's scattering of the incident sunlight. A bluish outer ring was further evidence of scattering by the submicron-size particles that extend several hundred kilometers above the main clouds. This view is a direct indication of Titan's extensive atmosphere. (*Jet Propulsion Laboratory*.)

Satellites of Saturn

Titan. The largest of Saturn's satellites is Titan and, in fact, is the second largest satellite in the solar system. It is the only satellite known to have a dense atmosphere. Some scientists believe that Titan may prove to be the most interesting body in the solar system, i.e., from a terrestrial perspective. For almost two decades, space scientists have searched for clues to the primeval Earth. The chemistry taking place in Titan's atmosphere may be similar to that which occurred in the Earth's atmosphere several billion years ago.

Because of its thick, opaque atmosphere, astronomers for some years believed that Titan was the largest satellite in the solar system. Their measurements were necessarily limited to measurements at the cloud tops. *Voyager 1's* close approach and diametric radio occultation showed, however, that Titan's surface diameter is only 5050 kilometers (3200 miles), which is slightly smaller than Ganymede, Jupiter's largest satellite. Both these satellites are larger than the planet Mercury. Titan's density appears to be about twice that of water ice. It has been suggested that it may be composed of nearly equal amounts of rock and ice.

Titan's surface was not seen in any of the *Voyager* photos. The surface is hidden by a dense, optically thick photochemical haze whose main layer is about 50 kilometers (30 miles) thicker in the southern hemisphere than in the northern hemisphere. Several distinct, detached haze layers can be seen above the visibly opaque haze layer. See Fig. 17. These haze layers merge with the main layer over the north pole of Titan, forming what scientists first thought was a darkened hood. The hood was found, under the better viewing conditions of *Voyager 2*, to be a dark ring around the pole. The southern hemisphere is slightly brighter than the northern hemisphere, possibly the result of seasonal effects. When the *Voyagers* flew past, the season on Titan was the equivalent of April on Earth; or early spring in the northern hemisphere and early autumn in the southern hemisphere.

The atmospheric pressure near Titan's surface is about 1.6 bars or 60% greater than that on Earth. The atmosphere is mostly nitrogen, also the major constituent of the Earth's atmosphere.

The surface temperature appears to be about 95 K (−288°F; −180°C), only 4 K above the triple-point temperature of methane. Methane, therefore, quite possibly plays the same role on Titan as water does on Earth—as rain, snow, and vapor. Rivers and lakes of methane may exist under a nitrogen sky. Clouds may drop liquid-methane precipitation. Titan's methane, through continuing photochemistry, may be converted to ethane, acetylene, an ethylene and, when combined with nitrogen, hydrogen cyanide. HCN is an especially important molecule, since it is a building block of amino acids. Titan's low temperature probably inhibits more complex organic chemistry.

Titan has no intrinsic magnetic field. Therefore, it has no electrically conducting and convecting liquid core. Its interaction with Saturn's magnetosphere creates a magnetic wake behind the satellite. Tital also serves as a source for both neutral and charged hydrogen atoms in Saturn's magnetosphere.

In later studies of Titan data (from *Voyager*), B. L. Lutz, C. de Bergh, and T. Owen reported discovery of carbon monoxide in the atmosphere of the satellite. The researchers reported that the 3-O rotation-vibration band of CO in the near-IR spectrum of Titan had been identified, and a reflecting layer model mixing ration of CO to molecular N of 6×10^{-5} had been determined. This result supported the earlier probable detection of CO_2 and strengthened possible analogies between the atmosphere of Titan and conditions on primitive Earth.

Eshelman, Lindal, and Tyler in post-*Voyager* data analysis, reported that Titan's dense and cold nitrogen atmosphere contains a small amount of methane under conditions at least approaching those at which one or both constituents would condense. The possibility of methane and nitrogen rain clouds and global methane oceans has been discussed widely, but the researchers report that from specific features of radio occultation and other *Voyager* data, it is concluded that nitrogen does not condense on Titan and that the satellite has neither global methane oceans nor a global cloud of liquid methane droplets. The investigators further observed that certain results indirectly support the conjecture that methane does not condense at any location. However, other considerations favor a methane ice haze in the troposphere, and liquid and solid methane may exist on the surface and as low clouds at polar latitudes.

F.M. Flaser (Goddard Space Flight Center), working on the same problem, observed that if global oceans of methane exist on Titan, the atmosphere above them must be within 2% of saturation. The two *Voyager* radio occultation soundings, made at low altitudes, probably occurred over land, since they imply a relative humidity $\lesssim 70\%$ near the surface. Oceans might exist at other low-altitude locations if the zonal wind velocities in the lowest 3 km are $\lesssim 4$ centimeters per second.

As a point of contrast, Lunine, Stevenson, and Yung (California Institute of Technology), from studies of *Voyager* data, suggested that Titan is covered by an ocean one to several kilometers deep consisting mainly of ethane. They report that if the ocean is in thermodynamic equilibrium with an atmosphere of 3% (mole fraction) methane, then its composition is roughly 70% ethane, 25% methane, and 5% nitrogen. Photochemical models predict that ethane is the dominant end product of methane photolysis so that the evolving ocean is both the source and the sink for continuing photolysis. The coexisting atmosphere is compatible with *Voyager* data.

As pointed out by D. O. Muhleman and colleagues, Division of Geological and Planetary Sciences, California Institute of Technology, "The present understanding of the atmosphere and surface conditions on Saturn's largest moon, Titan, including the stability of methane, and an application of thermodynamics leads to a strong prediction of liquid hydrocarbons in an ethane-methane mixture on the surface." These scientists utilized the Very Large Array radio telescope in New Mexico as the receiving instrument. Statistically significant echoes were obtained that show that Titan is not covered with a deep, global ocean of ethane, as previously thought. The researchers conclude their mid-1990 paper with, "The VLA/JPL (Very Large Array/Jet Propulsion Laboratory) radar measurements will be repeated as often as possible during the closest passages of Titan to the Earth. However, most of the questions concerning the Titan surface will not be answered until the proposed Cassini spacecraft reaches the Saturn system. It is quite likely that such a spacecraft would have an imaging radar and altimeter that will reveal much of the surface structure and will be very important for the study of Titan's geology, but will not be able to measure the surface reflectivity with an accuracy near to that of even current Earth-based radars."

Inner Satellites. Those Saturnian satellites which range in distance outward from the planet between 99,760 kilometers (62,000 miles) and 482,700 kilometers (300,000 miles) are known as *inner* satellites. These include Janus, Mimas, Enceladus, Tethys, Dione, and Rhea. These satellites are approximately spherical in shape and appear to be composed mostly of water ice. Enceladus reflects almost 100% of the sunlight that strikes it. These satellites represent a size of satellite not previously explored during pre-*Voyager* missions. Mimas, Tethys, Dione, and Rhea are all heavily cratered; Enceladus features fewer craters.

Enceladus. This satellite appears to have by far the most active surface of any satellite in the Saturnian system. At least five types of surface-terrain elements have been identified. Although craters can be seen across portions of its surface, the lack of craters in other areas implies an age less than a few hundred million years for the youngest region on the surface. See Fig. 18. It seems likely that portions of the surface of Enceladus are still undergoing change, since there are areas that are covered by ridged plains with no evidence of cratering down to the limit of resolution of *Voyager 2* cameras (2 kilometers; 1.1 miles). Other areas are criss-crossed by a pattern of linear faults. It is unlikely that a body so small could contain enough radioactive material for the modification to have been produced internally. A more likely source of heating appears to be tidal interaction with Dione—similar to the action between Jupiter and its satellite, Io. For Enceladus' present orbit, however, current theories of tidal heating do not predict generation of sufficient energy to explain all the heating that must have occurred. Because the satellite reflects so much sunlight, Enceladus' current surface temperature is only 72 K (−330°F; −201°C).

Tethys. Photos of Tethys taken by *Voyager 2* show an enormous impact crater nearly one-third the diameter of the satellite and larger than Mimas. The crater appears to have formed when Tethys was relatively fluid because nearly the original shape of the satellite was restored after impact. A gigantic fracture covers three-fourths of Tethys' circumference. See Fig. 19. Scientists suggest that the fracture can be explained if Tethys were once fluid and its crust hardened before the interior. Freezing and consequent expansion of the interior could have caused a surface fracture about the size of that observed. However, expansion would not be expected to cause only one large crack. The canyon has been named Ithaca Chasma. Tethys' current surface temperature is about 86 K (−305°F; −187°C).

Mimas. Photos of Mimas show a huge crater. The crater is about one-third the diameter of the satellite. This crater has greater surface relief, implying that Mimas was much more rigid at the time the cratering event occurred.

Dione. The terrain of this satellite is described as "wispy," probably consisting of numerous cracks rimed with ice. As shown in Fig. 20, a very large and wide fracture is noted in the southern hemisphere of the body. The wispy streaks stand out vividly against an already bright surface.

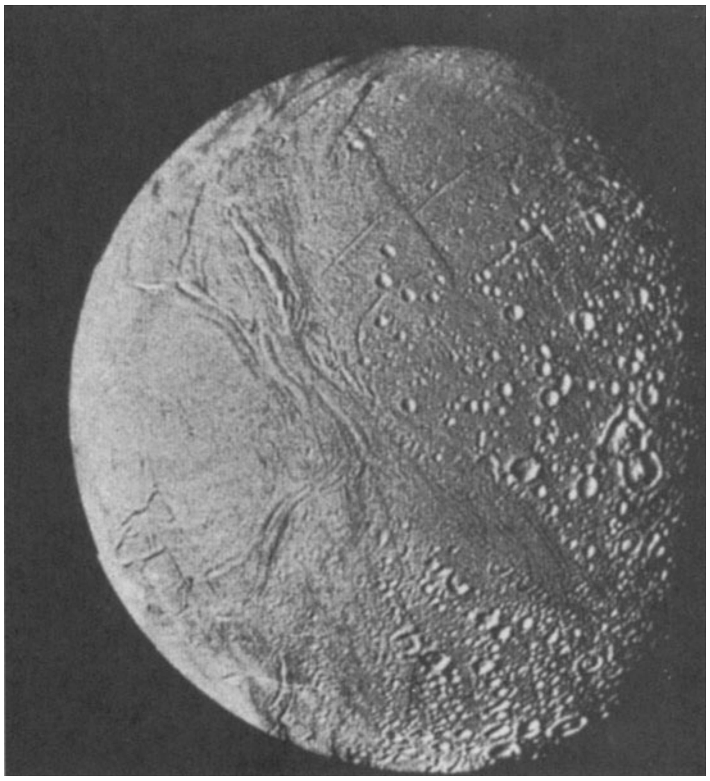


Fig. 18. High-resolution filtered image of Enceladus made from several images obtained on August 25, 1981 by *Voyager 2* from a range of 119,000 kilometers (74,000 miles). The view shows surface detail of the satellite which, in many ways, resembles Jupiter's Galilean satellite Ganymede. However, Enceladus is only one-tenth the size of Ganymede. Some regions of Enceladus show impact craters up to 35 kilometers (22 miles) in diameter, whereas other areas are smooth and uncratered. Linear sets of grooves tens of kilometers long traverse the surface and are probably faults resulting from deformation of the crust. The uncratered regions are geologically young and suggest that Enceladus has experienced a period of relatively recent internal melting. The rims of several craters near the lower center of the view have been flooded by the smooth terrain. The satellite is about 500 kilometers (310 miles) in diameter and has the brightest and whitest surface of any of the Saturnian satellites. Features as small as 2 kilometers (1.2 miles) are visible in this highest-resolution view made of Enceladus. (*Jet Propulsion Laboratory*.)

Rhea. This is a crater-saturated body, possibly impacted by many short-period comets. This is one of the better mapped satellites of Saturn.

Outer Satellites. In addition to Titan, previously described, the outer satellites of Saturn include Hyperion, Iapetus, and Phoebe. The outer satellites range in distance outward from the planet between about 1.15 million kilometers (720,000 miles) and 12.9 million kilometers (8 million miles).

Iapetus. This satellite has long been known to have large differences in surface brightness. Brightness of the surface material on the trailing side has been measured at 50%, while the leading-side material reflects only 5% of the sunlight that strikes it. Most of the dark material is distributed in a pattern directly centered on the leading surface, causing conjecture that the material was swept up as it spiralled inward, presumably from Phoebe. The trailing face of Iapetus, however, has several craters with dark floors. That implies that the dark material originated in the satellite's interior. It is possible that the dark material came both from Phoebe and from Iapetus' interior. See Fig. 21.

Hyperion. This satellite shows no evidence of internal activity. Its irregular shape and evidence of bombardment by meteoritic material makes it appear to be the oldest surface in the Saturnian system. Hyperion is shaped like a hamburger and is probably not in a gravitationally stable position. It is possible that one or more meteorites jostled Hype-

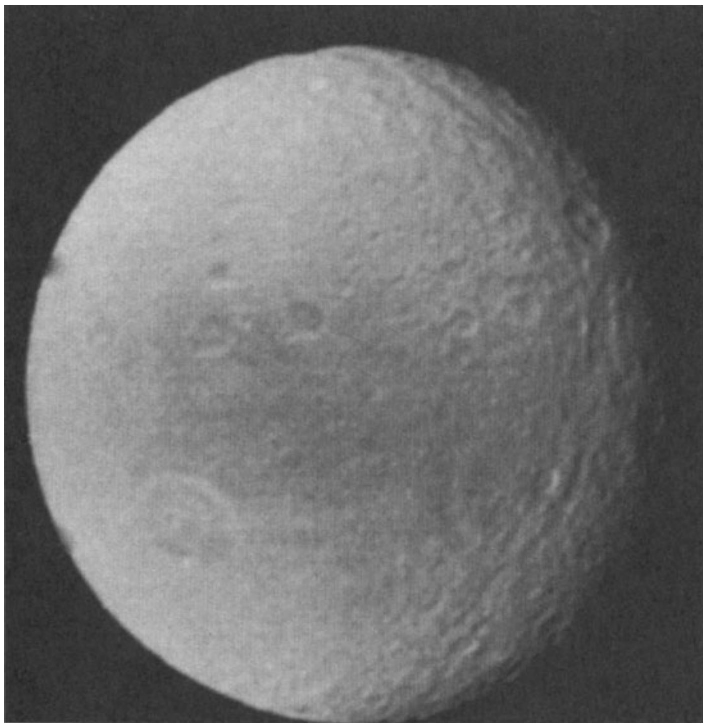


Fig. 19. Image of Tethys made by *Voyager 2* on August 25, 1981 when the spacecraft was 594,000 kilometers (368,000 miles) from the satellite. This view was compiled from images taken through the violet, clear, and green filters of Voyager's narrow-angle camera. Tethys shows two distinct types of terrain—bright, densely cratered regions; and relatively dark, lightly cratered plains that extend in a broad belt across the satellite. The densely cratered terrain is believed to be part of the ancient crust of the satellite; the lightly cratered plains are thought to have been formed later by internal processes. Also shown is a trough that runs parallel to the terminator (the day-night boundary) seen at the right). This trough is an extension of the huge canyon system noted by *Voyager 1* in 1980. This system extends nearly two-thirds the distance around Tethys. (*Jet Propulsion Laboratory*.)

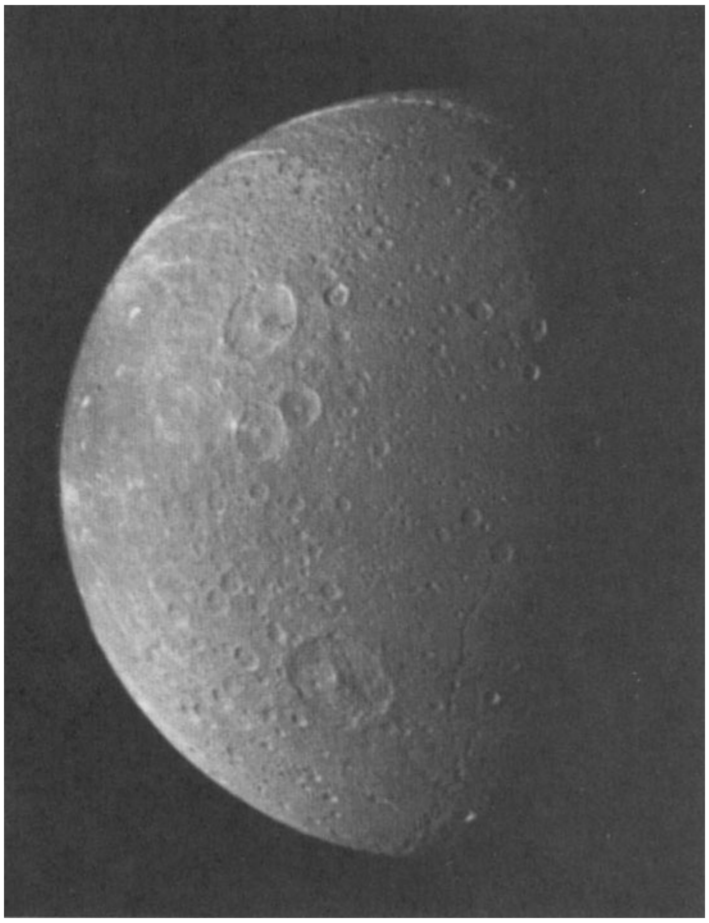


Fig. 20. A mosaic of Saturn's satellite Dione. Many impact craters display a historical record of collisions with cosmic debris. The largest crater is less than 100 kilometers (62 miles) in diameter and shows a well developed central peak. Bright rays represent material ejected from other impact craters. Sinous valleys probably formed by faults break the satellite's icy crust. Images in this mosaic were taken from a range of 162,000 kilometers (100,600 miles) by *Voyager 1* on November 12, 1980. (*Jet Propulsion Laboratory*.)

rion out of position and that the satellite will eventually swing back. See Fig. 22.

In studies of *Voyager* data, J. Wisdom and S. Peale (University of California, Santa Barbara) and F. Mignard (Research Center for the Study of Geodynamics and Astronomy, Grasse, France), reported that Hyperion tumbles chaotically, not just end over end, but first one way and then another, slowing down and then speeding up. These theorists believe that Saturn's gravity would synchronize Hyperion's rotation if the satellite were not of such an odd shape (115 × 145 × 190 km) and in an elongated, eccentric orbit. It is suggested that as Hyperion follows its eccentric orbit, Saturn tugs on different parts of the satellite with differing effects, sending it into chaotic tumbling. There is also the gravitational effect of Titan on the small satellite. Unfortunately, a difference of 18 months in the timing of observations is the basis of the current conclusions.

Further studies by Thomas (Cornell University) and colleagues of the brightness of Hyperion indicate a regular 13-day period of rotation during the 61 days of *Voyager 2*'s encounter with Saturn. The case of Hyperion's rotation has not been solved conclusively.

Phoebe. This satellite was photographed by *Voyager 2* after the spacecraft passed Saturn. Phoebe orbits in a retrograde direction in a plane much closer to the ecliptic than to Saturn's equatorial plane. *Voyager 2* found that Phoebe is roughly circular in shape and reflects about 6% of the sunlight. It is also quite red in color. Phoebe rotates on its axis about once in nine hours. Thus, unlike the other Saturnian satellites, it does not always show the same face to the planet. If, as scientists suggest, Phoebe is a captured asteroid with its composition unmodified since its formation in the outer solar system, it is the first such object that has been photographed.

Newly Discovered, Very Small Satellites. A total of seven very small satellites was observed by *Voyagers 1* and *2*. See Fig. 23. These satellites are irregularly shaped bodies that have been highly cratered by the impact of cosmic debris. The irregularity is probably due to fracturing by large impacts and is sustained by the rigidity of the bodies. These objects range from about ten to several hundred kilometers across. Included are the two F-ring shepherding satellites, first seen by *Voyager 1* in 1980.

Post-Mission Analyses

Based upon the masses of information collected from prior Saturn missions, numerous scientists establish continuing programs of data analysis to refine prior concepts of the characteristics of the planet. As one example, J. E. Klepeis and colleagues at the University of California, Lawrence Livermore National Laboratory, reported on their experiences by establishing models of Saturn and making comparisons with the conditions on Jupiter. The group has observed, "Models of Jupiter and Saturn postulate a central rock core surrounded by a fluid mixture of hydrogen and helium. These models suggest that the mixture is undergoing phase separation in Saturn, but not Jupiter. State-of-the-art total energy calculations of the enthalpy of mixing for ordered alloys of hydrogen and helium confirm that at least partial phase separation has occurred on Saturn and predict that this process also has begun on Jupiter."

Radar Imaging from Earth. The *Voyager* missions to Saturn did not end all Earth-based astronomical imaging of the planet. These would include imaging by the Hubble space telescope mentioned later. However, much greater progress has occurred in connection with radar imaging. For example, A. W. Grossman and colleagues at the Division of Geological and Planetary Science, California Institute of Technology, describe their use of high-resolution microwave imaging of the planet. In a 1989 paper, the researchers used this technology to study

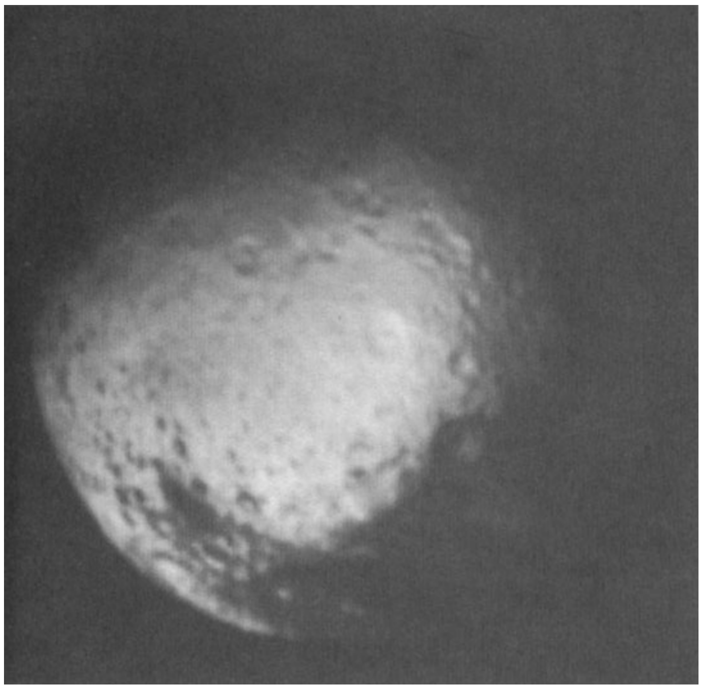


Fig. 21. View made by *Voyager 2* at time of its closest approach to lapetus, the outermost of Saturn's large satellites. Image was made on August 22, 1981 when the spacecraft was 1.1 million kilometers (680,000 miles) from the satellite. The camera resolution was about 21 kilometers (13 miles). This view, which is lit from above, primarily shows the heavily cratered northern hemisphere toward the bright trailing side of the satellite. The north pole itself is near the large central-peak crater seen partly in the shadow at the top of the image. Iapetus is noteworthy for the very dark material (lower and right-hand parts of this frame) that apparently covers the ice crust of the satellite primarily at its leading hemisphere. This dark material is red and reflects only about 5% of the incident sunlight; it may be of either external or internal origin. Study of its relationship with the underlying topography, for example, near the large crater at the border of the dark material, may help to resolve this mystery. (*Jet Propulsion Laboratory.*)

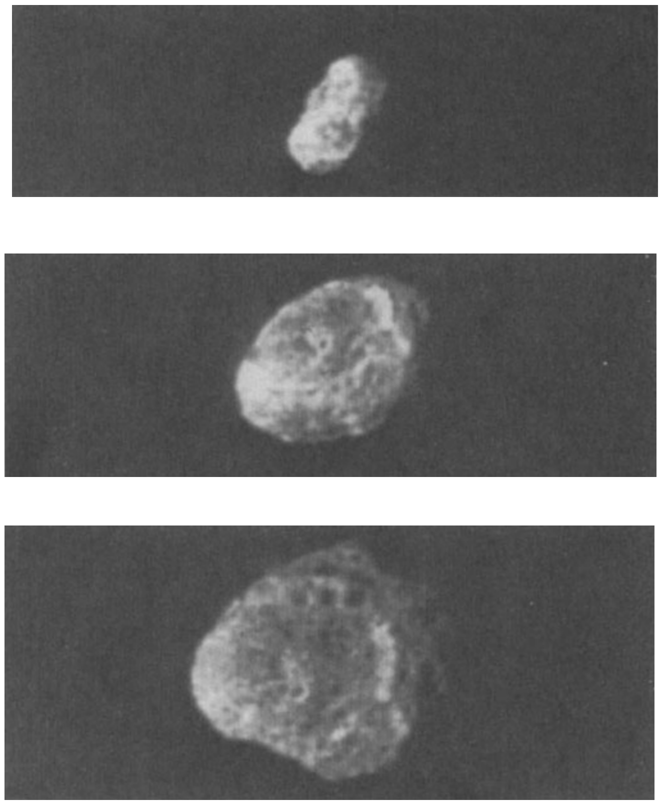


Fig. 22. Three views of Hyperion obtained as *Voyager 2* flew by this satellite. The views were taken: top view, on the morning of August 23, 1981 from a range of 1.2 million kilometers (740,000 miles); middle view, on the morning of August 24 from 700,000 kilometers (430,000 miles); lower view, at noon on August 24 from 500,000 kilometers (310,000 miles). Together, these views show the changing aspect of the satellite as *Voyager 2* moved in for closer views. Hyperion, roughly 360 by 210 kilometers (220 by 130 miles) and shaped like a hamburger, is probably not in a gravitationally stable position. Its surface is pock-marked with many meterorite-impact craters. The large indeptation at the bottom limb (lower view) is one such crater (about 100 kilometers; 60 miles across). The smallest visible crater pit is about 10–20 kilometers (6–12 miles)

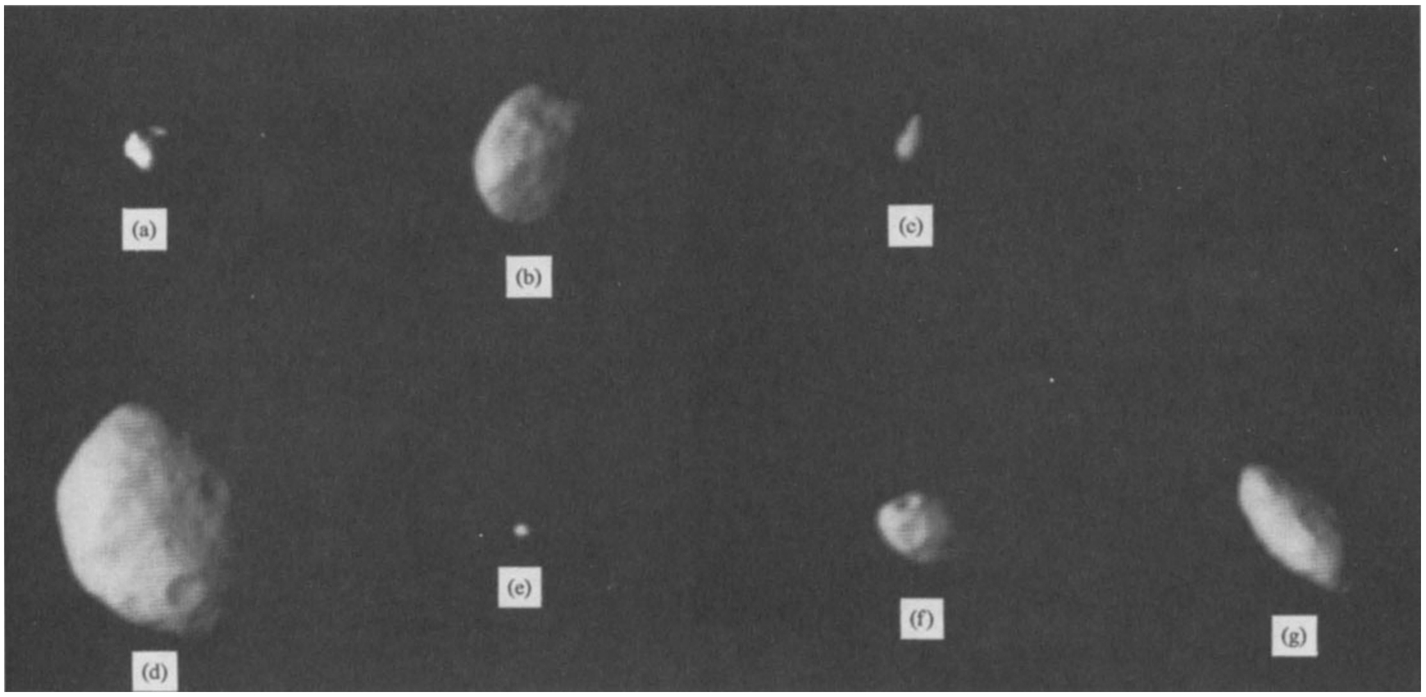


Fig. 23. Composite showing seven of the very small satellites of Saturn as photographed on August 25, 1981 by *Voyager 2*. (a) 1980S6 (Dione trojan); (b) 1980S3 (Trailing co-orbital); (c) 1980S25 (Trailing Tethys trojan); (d) 1980S1 (Leading co-orbital); (e) 1980S13 (Leading Tethys trojan); (f) 1980S26 (Outer F-ring shepherd); (g) 1980S27 (Inner F-ring shepherd). It should be noted that these views taken at differing distances from the various satellites, ranging from 248,000 kilometers (154,000 miles) to 667,000 kilometers (414,000 miles). (*Jet Propulsion Laboratory.*)

the ring systems of Saturn. The radio interferometric observations were made at the Very Large Array telescope in New Mexico at wavelengths of 2 and 6 centimeters. With this information, the investigators prepared maps that show an increase in brightness temperature of about 3 K from equator to pole at both wavelengths, while a map made with 6-meter radiation indicated a bright band at northern mid-latitudes. "These data are consistent with a radiative transfer model of the atmosphere that constrains the well-mixed, fully saturated, NH_3 mixing ratio to be 1.2×10^{-4} in a region just below the NH_3 clouds, while the observed bright band indicates a 25 percent relative decrease of NH_3 in northern mid-latitudes. Brightness temperatures for the classical rings also are presented in the paper. Ring brightness shows a variation with azimuth and is linearly polarized at an average value of about 5 percent. The variations in ring polarization suggest that at least 20 percent of the ring brightness is the result of a single scattering process."

The scientists conclude their paper, "The observations represent the ultimate in resolution and sensitivity obtainable from Earth-based radio telescopes. A vast improvement in our knowledge of Saturn's deep atmosphere and rings can be obtained from a Saturn-orbiting spacecraft instrument observing at radio wavelengths."

Great White Spot on Saturn. When the Hubble space telescope was pointed toward Saturn in late August 1990, the planet's image, after computer corrections to compensate for Hubble's disability, exhibited no surprises. The planet appeared calm and much as expected. However, in late September, a number of amateur astronomers reported the sudden development of a white spot reminiscent of a gigantic storm system. Larger, professional telescopes were turned to view Saturn, and these better images confirmed some unusual activity on the planet spread across Saturn's equatorial region. Scientists then interrupted Hubble's planned schedule for viewing Saturn. The white spot (Fig. 24) was reconfirmed. Some scientists liken the spot to the Great Red Spot of Jupiter. Because magma in the interior of Saturn had not been detected or even suspected, a volcanic eruption was quickly ruled out. There was a press conference on November 20, 1990, at which time a color image was released. The spot was described as a brick-colored hurricane. One scientist has observed that apparently massive amounts of energy, if not magmatic, well up from its interior (this energy remains from the period when the planet was formed) and that this heat, when it reaches the planet's thin atmosphere, periodically causes considerable turbulence.

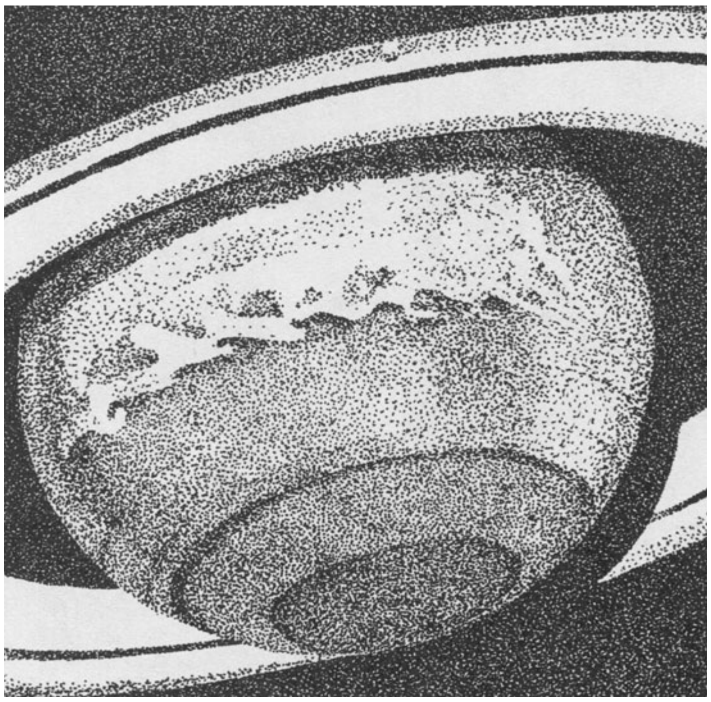


Fig. 24. Computer-enhanced fuzzy image of "Great White Spot" on Saturn first noted in August 1990 and still not fully explained. Image taken by the Hubble space telescope.

Additional Reading

Allison, M., Godfrey, D. A., and R. F. Bebbe: "A Wave Dynamical Interpretation of Saturn's Polar Hexagon," *Science*, 1061 (March 2, 1990).
Cherfas, J.: "Saturn Mission Backed, Europeans Relieved," *Science*, 628 (November 2, 1990).
Cowen, R.: "Spotting an Ephemeral Artifact on Saturn," *Science News*, 228 (October 13, 1990).
Dickson, D.: "Europeans Decide on a Trip to Saturn," *Science*, 1375 (December 9, 1988).
Eberhart, J.: "Saturn Ring Ripple Suggests 19th Moon," *Science News*, 31 (July 14, 1990).
Eberhart, J.: "Five-Year Hunt Locates Saturn's 18th Moon," *Science News*, 69 (August 4, 1990).
Godfrey, D. A.: "The Rotation Period of Saturn's Polar Hexagon," *Science*, 1206 (March 9, 1990).
Grossman, A. W., Muhelman, D. O., and G. L. Berge: "High-Resolution Microwave Images of Saturn," *Science*, 1211 (September 15, 1989).
Ingersoll, A. P.: "Atmospheric Dynamics of the Outer Planets," *Science*, 308 (April 20, 1990).
Kerr, R. A.: "A Passion for the Little Things Among the Planets," *Science*, 998 (November 24, 1989).
Klepeis, J. E., et al.: "Hydrogen-Helium Mixtures at Megabar Pressures: Implications for Jupiter and Saturn," *Science*, 986 (November 15, 1991).
Kunine, J. I.: "Origin and Evolution of Outer Solar System Atmospheres," *Science*, 141 (July 14, 1989).
McKay, C. P., Pollack, J. B., and R. Courtin: "The Greenhouse and Antigreenhouse Effects on Titan," *Science*, 1118 (September 6, 1991).
Muhleman, D. O., et al.: "Radar Reflectivity of Titan," *Science*, 975 (May 25, 1990).
Powell, C. S.: "Hubble Gags a Great White," *Sci. Amer.*, 26 (February 1991).
Rosen, P. A., and J. J. Lissauer: "The Titan Nodal Bending Wave in Saturn's Ring C," *Science*, 690 (August 5, 1988).
Waldrop, M. M.: "Titan: Continents in a Hydrocarbon Sea," *Science*, 129 (July 14, 1989).
Waldrop, M. M.: "Images of an Unquiet Planet (Saturn)," *Science*, 1201 (November 30, 1990).

Pre-1986 References

Bane, D.: "The *Voyager 1* and *2* Saturn Science Results," Jet Propulsion Laboratory, Pasadena, California, December 1981.
Davis, D. R., et al.: "Saturn Ring Particles as Dynamic Ephemeral Bodies," *Science*, 224, 744–747 (1984).
Dyer, J. W.: "*Pioneer* Saturn" (Contains 14 papers relating to *Pioneer 11* encounter with Saturn), *Science*, **207**, 400–453 (1980).
Eshleman, V. R., Lindal, G. F., and G. L. Tyler: "Is Titan Wet or Dry?" *Science*, **221,** 53–55 (1983).
Esposito, L. W., et al.: "Eccentric Ringlet in the Maxwell Gap at 1.45 Saturn Radii: Multi-Instrument Voyager Observations," *Science*, **222,** 57–59 (1983).
Flasar, F. M.: "Oceans on Titan?", *Science*, **221**, 55–57 (1983).
Gehrels, T., and M. S. Matthews, Eds.: "Saturn," Univ. of Arizona Press, Tucson, Arizona, 1984.
Ingersoll, A. P.: "Jupiter and Saturn," *Sci. Amer.*, 90–108 (December 1981).
Loudon, J.: "The Last Picture Show," *Technol. Rev. (MIT)*, **84(2)**, 19 (1981).
Lunine, J. I., Stevenson, D. J., and Y. L. Yung: "Ethane Ocean on Titan," *Science*, **222,** 1229–1230 (1983).
Lutz, B. L., De Bergh, C., and T. Owens: "Titan: Discovery of Carbon Monoxide in Its Atmosphere," *Science*, **220,** 1374–1375 (1983).
Muhleman, D. O., Berge, G. L., and R. T. Clancy: "Microwave Measurements of Carbon Monoxide on Titan," *Science*, **223**, 393–396 (1984).
News: "Frigid Oceans for Triton and Titan: Chaotic Rotation Predicted for Hyperion; Could Saturn's Rings Have Melted Enceladus?" *Science*, **221,** 448–449 (1983).
News: "The Rotation of Saturn's Hyperion Looks Chaotic," *Science*, **230,** 1027 (1985).
Owen, T.: "Titan," *Sci. Amer.*, 98–109 (February 1982).
Soderblom, L. A., and T. V. Johnson: "The Moons of Saturn," *Sci. Amer.*, 100–116 (January 1982).
Stone, E. C., and E. D. Miner: "Voyager 1 Encounter with the Saturnian System," (Followed by a series of detailed articles). *Science*, **212,** 159–243 (1981).
Stone, E. C., et al: "Voyager 2 Encounter with the Saturnian System," *Science*, **215,** 499–594 (1982).
Zebker, H. A., and G. L. Tyler: "Thickness of Saturn's Rings Inferred from *Voyager 1* Observations of Microwave Scatter," *Science*, **223**, 396–398 (1984).

Editor's Note: The staff of this encyclopedia is most appreciative of the cooperation extended by D. Bane and A. S. Woods and their associates at the Jet Propulsion Laboratory, Pasadena, California, for furnishing information for numerous portions of this article.

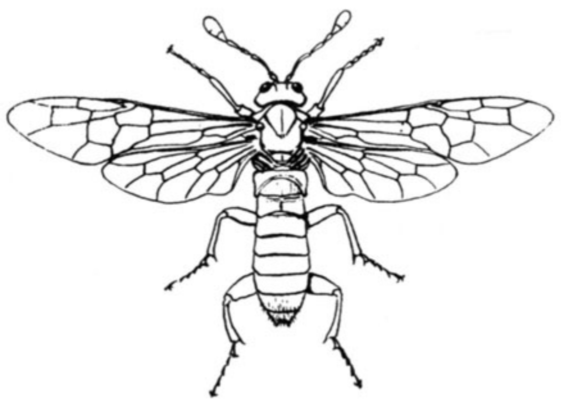
SAUERKRAUT. See **Brassica.**

SAUGER. See **Perches and Darters.**

SAVANNAH. A tropical or subtropical region of grassland and other drought-resistant (xerophilous) vegetation. This type of growth occurs in regions that have a long dry season but a heavy rainy season, and continuously warm temperatures. Africa has the most extensive areas of savannah, but they are also widespread in South America (the *campos*), and, to a lesser extent, in India and southeast Asia, Australia, and Central America. See also **Biome.**

SAWFISHES. See **Skates and Rays.**

SAWFLY (*Insecta, Hymenoptera*). A plant-feeding member of this order, whose more familiar species are the ants, bees, and wasps. The sawflies have four wings, somewhat like those of the wasps, but the abdomen is broadly connected with the thorax, in contrast with the thin-waisted bodies of the other forms. See accompanying illustration. They are named from the sawlike ovipositor with which slits are cut in the tissues of plants to receive the eggs. Some species are of economic importance. Since the larvae eat leaves they can be destroyed by a number of control chemicals. Cultural practices are also extremely important.


Sawfly. (*USDA*.)

A native of North America, the *wheatstem sawfly* (*Cephus cinctus*, Norton) attacks wheat, as well as barley, spring rye, spelt, timothy, and a number of native grasses. The *European wheatstem sawfly* (*Cephus pygmacus*, Linne) is a closely related species. These insects are most damaging in the wheat-growing stages of the north and west of the Mississippi, although the European species is found mostly in the eastern states of New York and Pennsylvania. Without effective control measures, records indicate that in some years these insects have destroyed more than 50% of some crops. An infested field displays fallen straw (similar to the damage wrought by the hessian fly and jointworm). This is because the sawfly larvae have been feeding there. With the brown-headed, legless larva ($\frac{1}{3}$–$\frac{1}{2}$) inch; 8–12 millimeters long) will be straw cuttings that look like sawdust. This is the result of the cutting-type action characteristic of this insect. As the season progresses, the larva works their way down the stem and finally makes a V-shaped groove on the inside of the stem, causing the stem to break. The insect drops to the base of the plant and plugs up the opening, where it hibernates and ultimately pupates.

Control practices include turning the stubble shortly after harvest. Burning is not practical because the insects live close to ground level. Early cutting of grain, when possible, also reduces the damage. Rotation of crops, planting corn (maize), flax, alfalfa or sweet clover, is effective. Where the sawfly is a persistent problem, wheat varieties that are solid-stemmed should be selected.

The *raspberry sawfly* (*Monophadnoides geniculatus* or *Priophorus rubivorus*, Rowler) is sometimes abundant on raspberry leaves at a time in spring when the plant has reached its full foliage. The larva is a spiny, multi-legged, pale-green worm that usually feeds along the edges of leaves. Unless controlled, the plant may be fully stripped of its leaves.

The *currant stem girdler* (*Janus integer*) is another economic species found in the United States.

Horntail flies are large sawflies (woodwasps) whose larvae bore in the trunks of trees. The adults have a cylindrical body; the female has a short, strong ovipositor which is the source of the name horn-tail. With this organ, holes are drilled into the wood of the tree for deposition of eggs. Among these is the species *Cimbex americana*, the *elm sawfly*.

The *pear slug* (*Caliroa cerasi*, Linne) is the olive to dark green or black larva of a shiny black sawfly. The larva resembles a snail or slug and ranges up to $\frac{1}{2}$ inch (12 millimeters) in length. The larva feeds on the upper surface of cherry, pear, and plum trees, retards growth and development of fruit, and generally weakens the tree. The pest is distributed throughout the United States. Chemical controls are the same as those for fruit curculios. See **Curculio.** In small orchard operations, handpicking the larvae from the leaves and placing them in a pail containing kerosene is effective. Also, a few applications of lime diluted with water and applied to the leaves is effective. The slugs can be washed from the foliage with a strong stream of water and then destroyed at groundlevel.

Knerer and Atwood have made an extensive study of the polymorphism and speciation of *Diprionid* sawflies, and this is reported in *Science*, **179** (4078), 1090–1099 (1973). The *Diprionidae* represent only a small fraction of all the sawflies known, but they share the habits of most other leaf feeders. The family is interesting mainly because of the diversity of distinct races or physiological strains that are adapted to specific host plants, and because of the social behavior exhibited by the larvae in aggregations. Both phenomena illustrate various evolutionary mechanisms at work, simultaneously providing examples of newly emerging biologic units and of the origin of some of the most primitive social behavior found in insects. These factors are well summarized by Knerer and Atwood.

SAYBOLT VISCOSIMETER. See **Viscosity.**

SCABIES. See **Dermatitis and Dermatosis.**

SCAD (*Osteichthyes*). Of the order *Percomorphi*, suborder *Percoidea*, family *Carangidae*, the scad is a species common in the region of the West Indies. The scad is also called cigar fish or round robin. The big-eyed scad, goggler, or chicharro is related and although common in warmer waters, does range as far north as Cape Cod. The term *scad* also has been applied to the related horse mackerel, a European species occasionally taken on the Atlantic coast of North America.

SCALAR FIELD. Consider the space reflection $x' \rightarrow; x$; $\mathbf{x}' = -\mathbf{x}$, $x'^{o} = x^{o}$. Define the transformation of a spinless field operator under space reflection such that

$$\begin{aligned}\phi'(x') &= P\phi(x)P^{-1}\\ &= \eta_p\phi(x')\end{aligned}$$

Where η_p is a complex constant and P the unitary operator which induces the transformation. If the operator of space reflection is performed twice, we must revert to the original field so that $\eta_p^2 = 1$ or $\eta_p = \pm 1$. A field which transforms with $\eta_p = +1$ is called a *scalar field*, while one that transforms with $\eta_p = -1$ is called a *pseudoscalar field*. The particles described by scalar (pseudoscalar) fields have an intrinsic parity $+1$ (-1).

SCALAR PRODUCT. See **Vector Multiplication.**

SCALAR QUANTITY. A number, either in the sense of the ratio between two quantities of the same kind, or as distinguished from other mathematical constructs, e.g., vectors, matrices, etc.

SCALD FISHES. See **Flatfishes.**

SCALE. Six common uses of this term are: 1. A balance used for weighing.

2. A series of markings at regular intervals which are used for measurement or computation.

3. A defined set of values in terms of which quantities of the same nature as those defined may be measured, e.g., temperature scale.

4. A musical scale consisting of a series of notes (symbols, sensations or stimuli), arranged from high to low by a specified scheme of intervals, suitable for musical purposes.

5. In metallurgy, scale is the oxide layer that forms on the surface of metals upon heating in air or other oxidizing gases. The heavy scale that forms on steel ingots or billets upon heating for rolling or forging breaks away as the metal is deformed in the mill or under the hammer; however a lighter, often very adherent scale always remains after hot-working operations and after annealing or other heat treatments unless a non-oxidizing protective atmosphere is provided. Scale is removed from steel by pickling, generally in warm dilute sulfuric or hydrochloric acid. The scale which forms on stainless steel is much more resistant and its removal requires strong acids such as mixtures of hot nitric and hydrofluoric. A tight adherent scale is often left on steel for its protective value; for example, steam-blued steel sheets are used for stovepipe.

6. In animal anatomy, a scale is a flat structure developed as a covering. A. The scales of fishes are in many cases arranged like shingles to form a complete armor at the surface of the body, and in other forms are small and scattered, merely adding to the resistant qualities of the skin. The elasmobranch fishes have placoid scales, whose form includes a broad base and a projecting enamel-covered point. This form of scale is much like the teeth of these fishes and is supposed to be ancestral to all vertebrate teeth. Ganoid scales, found in relatively few fishes, are regarded as a modification of the placoid type by the loss of the point and the addition of a hard outer layer known as ganoin. Some of the more primitive fishes of other groups have rounded scales, with smooth margins, known as cycloid, and others have ctenoid scales, with comblike edges. The two last forms are found in the higher fishes. They bear neither enamel nor ganoin. B. The scales found over the entire surface of the body in reptiles, on the legs of birds, and to a much more limited extent in mammals, as on the tails of rodents, are quite different from the scales of fishes. Each is a modified area of the skin, thickened and hardened by the development of the horny substance, keratin. Scales reach their highest development among the mammals in the pangolins and armadillos. In the latter they are underlaid by bony plates. C. The butterflies and moths, a few beetles, and some other insects have the surface of the body and wings more or less covered by flattened scales. These structures are modified setae. They often contain pigments and in many species are so formed that they produce iridescent, metallic, or glossy physical colors by breaking up the light rays which they reflect. D. Scale insects, often called scales, are highly specialized sucking insects which live on plants. They belong to the order *Hemiptera* and are related to the plant lice and phylloxerans. Most species are minute. The young and the female adults are simplified and remain closely attached to the plant, secreting over themselves a protective covering or scale which gives them their name. This covering is usually characteristic of the species.

Scale insects include many species of economic importance. Among them are the useful cochineal insect, a source of dye, and the lac insect whose scale is the raw material from which shellac is made. China wax is also a scale insect product. Among the harmful species the purple scale of citrus fruits, the San José scale, and the oystershell scale are important. Because of the protective scale these species are not easily destroyed by the usual contact sprays. They are controlled by spraying but the concentration of poison must be high and spraying during the dormant stage of the plant is often necessary as a result. Fumigation of citrus trees with cyanide is practiced extensively. This method requires special equipment since the tree must be enclosed in a tent. See **Scale Insects.**

SCALE FACTOR (Computer System). In analog computing, a proportionality factor which relates the magnitude of a variable to its representation within a computer. In digital computing, the arbitrary factor which may be associated with numbers in a computer to adjust the position of the radix so that the significant digits occupy specified columns.

SCALE FISH. See **Flatfishes.**

SCALE HEIGHT. See **Atmospheric Pressure.**

SCALE INSECTS (*Insecta, Hemiptera*). These insects include numerous species of the family *Coccidae*, of which the mealy bugs also are members. However, there is a marked contrast: Whereas the mealy bugs can move about freely, the scale insects, within a few hours after birth or hatching, remain fixed throughout the remainder of their life. The typical scale insect may be described as wingless and grublike in appearance, but masked by a covering of a cottony, powdery, or waxy secretion. In some species, this coating hardens into a scale-like shell or skin—hence the term *scale insects*. See Fig. 1. Generally, the scale insects tend to become specialists, thriving on a few select trees and fruits as their source of food and habitat. For certain food crops, such as apple, citrus, and other deciduous fruits, the scale insects are regarded as very damaging pests. The insects are found throughout the world wherever these crops are cultivated.

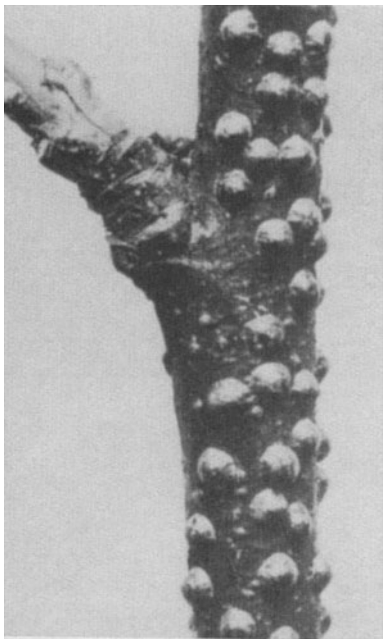


Fig. 1. Terrapin scale attached to twig of peach tree.

The scalelike coverings make it difficult to kill these insects. They are very small and thus easy to transport as stock may be moved from a glasshouse or nursery for transplanting. They multiply very rapidly. For example, researchers have found that a single female of the *San Jose species* can rear from 100 to 600 young at a rate of 4 to 5 generations per year. It has been established that one *Lecanium scale insect* can lay over 2000 eggs. Fortunately, the scale insects have a number of natural enemies which most likely control the highest percentage of these pests, even though chemical intervention is often required to make up for any deficiencies of the natural enemies. A number of years ago, the *cottony cushion scale* (*Icerya purchasi*) was accidentally introduced into California. Originally an Australian insect, this scale multiplied very fast once free of its natural enemies and spread rapidly over the West Coast citrus crop. A very large measure of control was accomplished biologically by the purposeful introduction of lady beetles which have a great preference for this scale insect as a food source. Parasitic wasps and mites also are effective in naturally keeping the scale populations under control.

Because the scale insects have a considerable amount of built-in protection, however, including an ability to tightly attach themselves to the bark of trees, which makes them difficult to remove by washing, for years the most effective means of controlling really bad infestations has been the use of various fumigants, notably cyanogen gas. Gastight buildings for use in treating plants prior to transplanting have been used, as well as large tentlike enclosures over trees, glasshouses, and other buildings. For mild populations of the insects, however, various sprays, such as oil and kerosene emulsions, sometimes lime-sulfur formulations, and even soap solutions (under high pressure), are effective. In some cases, mechanical scraping of bark or twigs can be effective.

The principal damage caused by scale insects is their sucking out plant juices, causing discolored (sometimes red) spots on leaves, stems, and fruit. Wounds make excellent sites for invasion by fungus spores.

Following is an abridged list of various scale insects that are damaging to crops.

Barnacle Scale (*Ceroplastes cirripediformis*). Found on citrus and *Eupatorium*. Insect is about $\frac{1}{5}$ inch (5 millimeters) in length. Scales are dark brown and surface of scale sculptured to appear like miniature barnacles.

Black Scale (*Saissetia olex; Lecanium oleae*). Found on citrus and deciduous fruits; olive. Insect is oval-shape, about $\frac{1}{4}$ inch (6 millimeters) in diameter. Scale covering is black. On back of female, there is an H marking. Scales secrete honeydew, providing a locale for fungus to grow, thus smutting the fruit.

Black Parlatoria Scale (*Parlatoria zizyphus*). A pest to citrus in many parts of the world. The insect spends most of its life beneath a dull-black scale. It gives off a whitish or brownish secretion that sometimes extends beyond the scale. The female is about $\frac{1}{16}$ inch (1.5 millimeters) long; the male is somewhat smaller. The insects attach themselves to leaves and fruit of citrus. When abundant, they form a black crust. A heavy infestation causes the leaves to turn yellow and drop and prevents full development of the fruit.

Coconut Scale (*Aspidiotus destructor*). Found mainly on coconut. Insect is white to cream in color and has transparent scales. The insects frequently are found in large numbers on underside of leaves and fruit.

Cottony Cushion Scale (*Icerya purchasi*). Found on almond, apricot, citrus, and fig. Insect is red to yellowish in color and covered with comparatively large, white, fluted cottony masses for covering eggs. Insect is from $\frac{1}{4}$ to $\frac{1}{2}$ inch (6 to 12 millimeters) in length. Control is mainly by natural enemies, such as *Vedalia* lady beetle. See Fig. 2.

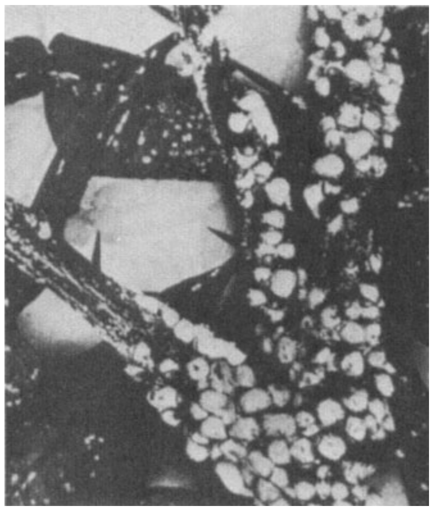


Fig. 2. Cotton cushion scale insects on orange tree. (*USDA*.)

Date Palm Scale (*Parlatoria blanchardii*). Found mainly on date palm. A small, elongate insect having gray or black scales with white edges. The males are always white.

European Fruit Scale (*Lecanium corni*). Found mainly on plum. A rather large, circular-shaped insect which sometimes can be very destructive.

Euonymous Scale (*Chionaspis evonymi*). Found on small trees and shrubs of genus *Euonymous*. Insect is about $\frac{1}{12}$ inch (2 millimeters) in diameter with dark brown, convex scales on female and pure white scales on males.

Fig Scale (*Lepidosaphes ficus*, Signoret). This insect is a pest in fig orchards and traces of its presence may persist on the fruit after harvest. When infestation on the leaves becomes heavy, crawlers go to the developing fruit, settle, and feed on the green fig. Heavy infestations result in dried figs that are somewhat spotted and deformed.

Florida Red Scale (*Chrysomphalus aonidum*). Found on aspidistra (Asian herb and sometimes houseplant), *Betula* (alder and birch trees), citrus, coconut, and *Hedera* (woody vines; true ivy). Insect is circular with flat brown scales. Size ranges from $\frac{1}{16}$ to $\frac{1}{8}$ inch (1.5 to 3 millimeters) in diameter. Infestations require fumigation. See Fig. 3.

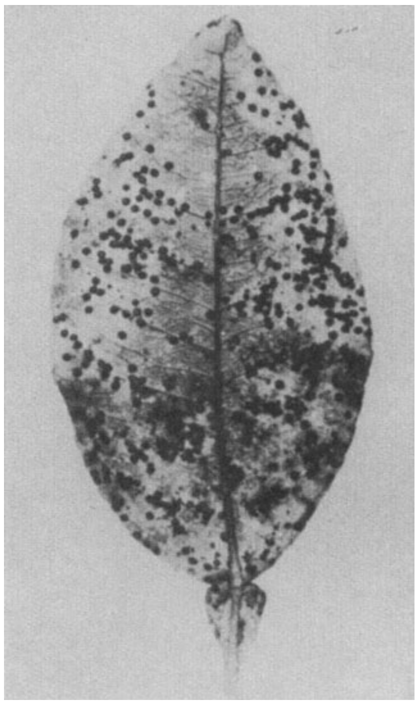


Fig. 3. Infestation of Florida red scale insects on tree leaf. (*USDA*.)

Frosted Scale (*Eulecanium pruinosum*). Found on apricot, *Betula* (particularly birch in California), and *Laurus*. Insect is comparatively large (about $\frac{1}{2}$ inch; 12 millimeters long), is hemispherical in shape, and has a frosty-appearing wax covering.

Hemispherical Scale (*Saissetia hemisphaerica*). Found on *Betula*, citrus, coffee, *Cycas*, ferns, guava, *Hedera*. Insect is covered with a smooth, soft scale without markings. Very commonly found in glasshouses.

Japanese Wax Scale (*Ceroplaste ceriferus*). Found mainly on gardenia, but can spread to other plants. Insect is from $\frac{1}{4}$ to $\frac{1}{3}$ inch (6 to 8 millimeters) in diameter. It is covered with white-to-creamy waxy masses.

Juniper Scale (*Diaspis carueli*). Found mainly on juniper, but can spread to other plants. Insect is snow white with circular scales and with yellow central exuviae.

Oyster Scale (*Lepidosaphes ulmi*). Found on apple, *Betula*, *Ceanothus*, *Crataegus*, currant, and poplar. Sometimes also called the bark louse, this scale is about $\frac{1}{8}$ inch (3 millimeters) long and resembles an oyster shell in shape. The insect hibernates as small white eggs under old scales. The young hatch in late spring and appear as whitish lice crawling on the bark.

Pineapple Scale (*Diaspis bromeliae*). Found on *Billbergia*, olive, and pineapple. The insect is thin and circular, usually with very white scales and yellow exuviae infesting the leaves and fruit.

Pine Leaf Scale (*Chionaspis pinifoliae*). Found mainly on pine trees. The insect is quite small with white scales.

Pit-Making Oak Scale (*Asterolecanium variolosum*). Found mainly on oak. The insect is circular in shape and has glossy greenish-yellow scales. The scale is particularly damaging to the golden oak.

Purple Scale (*Lepidosaphes beckii*). Found on citrus, *Codiaeum*, fig, and olive. The insect ranges from a reddish-brown to a rich purple in color and has oyster-shell shaped scales. The insect is from $\frac{1}{16}$ to $\frac{1}{8}$ inch (1.5 to 3 millimeters) in length.

Rose Scale (*Aulacaspis rose*). Found mainly on rose. The insect is circular, with small whitish scales.

San Jose Scale (*Aspidiotus perniciosus*). One of the better known and more damaging of the scales. Found on almond, apple, cherry, *Cornus*, *Crataegus*, currant, pear, plum, *Sorbus*, etc. The insect is nearly circular and about as large as a pinhead. See Fig. 4. When in large numbers, the insects form a crust on branches, causing small red spots on the fruit. This scale multiplies with amazing rapidity. Three to four broods per year are possible. The young are born alive. Breeding goes on until quite late in the autumn, at which time all stages of the insect are killed, with exception of the small, half-grown black scales which often hibernate through the winter safely. The scale also sometimes attacks rose, gooseberry, elm, chestnut, oak, and walnut.

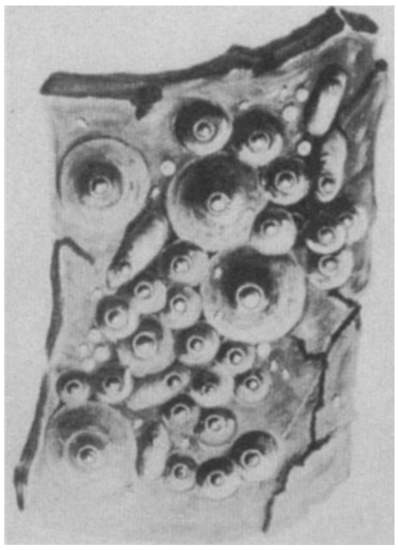


Fig. 4. Magnified view of San Jose scale insects fastened to branch of fruit tree. (*USDA*.)

Scurfy Scale (*Chionaspis furfurus*). Found on apple, *Crataegus*, and *Sorbus*. The insect is pearl-shaped, white to off-white in color, and about $\frac{1}{8}$ inch (3 millimeters) long. The insect frequently becomes intimately encrusted to the bark, with a resulting scurfy appearance of the bark. Hibernation is in the form of purple eggs located under old scales.

Soft Brown Scale (*Coccus hesperidum*). Found on citrus, *Cycas*, fig, *Hedera*, *Ipomoea*, jasmin, and *Laurus*. The insect is oval-shaped, flat, about $\frac{1}{4}$ inch (6 millimeters) long, and with flat, soft scales.

White Peach Scale (*Diaspis pentagonia*). Found on *Cycas*, *Diospyros*, *Laurus*, and peach. The insect is circular in shape, with gray scales and with exuviae at one side of center.

Several closely related scale insects have had economic significance outside the realm of agriculture over the years. Some have been used as the source of dyes, such as cochineal (*Coccus cacti*), a species very common in Mexico, and the scarlet grains (*Coccus poionicus*), which occur in Poland on the roots of the knawel plant and once were used as a source of dye for fabrics. One scale, known as the wax insect of China (*Coccus sinensis*) has been used for centuries as a source of wax used in candles, lacquers, sealers, and varnishes. The shell-lac insects (from which shellac is derived) are closely related.

The iridescent covering secreted by some of the scale insects, known as *ground pearl*, is useful in making ornaments.

More detail on scale insects can be found in "Foods and Food Production Encyclopedia," (D. M. and G. D. Considine, Eds.), Van Nostrand Reinhold, New York, 1982.

SCALE (Weighing). See **Weighing.**

SCALES (Fish). See **Fishes.**

SCALING (Boiler). See **Feedwater (Boiler).**

SCALING CIRCUIT. A circuit that produces an output pulse whenever a prescribed number of input pulses have been received. A binary scaler produces an output pulse whenever two input pulses have been received. By putting binary scalers in sequences, scales of two, four, eight, sixteen, etc., are obtained. A decade scalar produces an output pulse whenever ten input pulses have been received. By putting decade scalers in sequence, scales of ten, one hundred, one thousand, etc., are obtained.

SCALLOP. See **Mollusks.**

SCANDIUM. Chemical element symbol Sc, at. no. 21, at. wt. 44.956, periodic table group 3, mp 1540°C, bp 2850°C, density 2.985 g/cm^3 (alpha form), 3.19 g/cm^3 (beta form). The alpha form is close-packed hexagonal. The face-centered cubic allotrope, although generally accepted, has not received full recognition. Scandium is a relatively soft metal with a silvery luster. The metal oxidizes rapidly in air. Scandium combines readily with water, oxygen, acids, halogens, and chalcogenides. ^{45}Sc occurs in nature and is not radioactive. Nine radioactive isotopes have been identified ^{40}Sc through ^{44}Sc and ^{46}Sc through ^{49}Sc, all with relatively short half-lives, ranging from a fraction of a second up to 84 days. Scandium occurs widely throughout nature, but in reasonably concentrated forms only in a few uncommon minerals. Abundance in the earth's crust is estimated at approximately 5–6 $\times 10^{-4}$ %, ranking it ahead of such elements as antimony, bismuth, silver, and gold. It is estimated that a cubic mile of seawater contains about 375 pounds of the element. Scandium was predicted by Mendeleev in 1869, at which time he called it *ekaboron* and foretold accurately a number of its properties. A small amount of scandium oxide was extracted from euxenite and gadolinite by Nilson in 1879, a material that Nilson called *scandia*. In the same year, Cleve isolated a greater quantity of the oxide, from which several compounds were prepared and favorably compared with Mendeleev's predictions for ekaboron. First ionization potential, 6.56 eV; second, 12.8 eV; third, 24.64 eV. Other physical properties are given under **Chemical Elements.**

Scandium occurs in some ores with the Lanthanum Series elements. It is easily separated from the Lanthanides, as well as yttrium, by taking advantage of the greater solubility of its thiocyanate in ether. The three recognized scandium minerals are thortveilite, a silicate; and sterrettite and kolbeckite, both phosphates. Wiikite and bazzite, complex niobates and silicates, are known to contain more than 1% scandium. Davidite, with a concentration of 0.02% Sc_2O_3, also is a major source of the element. Scandium has not been found without the Lanthanide elements and an association with yttrium. The element usually is separated from ore extracts and concentrates by precipitation as the oxalate. Scandium with a purity of 99.99% has been produced.

In water solutions, the scandium ion has a triple positive charge. Studies show, however, that the simple Sc^{3+} ion seldom exists. Rather, the form is highly polymerized and hydrolyzed—with hydroxy-bonded structures. In forming compounds, scandium parallels aluminum, yttrium, gallium, indium, and tellurium. Several carbides of scandium have been reported, the most stable being ScC.

Like the hydroxides of the Lanthanides, scandium hydroxide, $Sc(OH)_3$, is precipitated by addition of alkalies to solutions of scandium salts; however, the latter is precipitated at pH 4.9, while the former require pH 6.3 or more, a property which is utilized in one method of separation. Upon heating the hydroxide (or certain oxyacid salts), scandium oxide, Sc_2O_3 is produced. Scandium hydroxide is less acidic than aluminum hydroxide, requiring boiling KOH solution to form the complex potassium compound, $K_2[Sc(OH)_5 \cdot H_2O] \cdot 3H_2O$.

All four trihalides of scandium are known. The trifluoride is very slightly soluble in H_2O, and is precipitated from scandium nitrate, $Sc(NO_3)_3$, solutions by hydrofluoric acid. It dissolves in alkali fluorides to yield the complex ion $[ScF_6]^{3-}$. The chloride is formed in solution by treating the hydroxide or oxide with HCl, yielding hydrated crystals on concentration, which give hydroxychlorides on heating. The bromide is also prepared from the oxide or hydroxide and hydrobromic acid, or in anhydrous form from the oxide, carbon, and bromine, on heating. The iodide is also prepared by the latter method.

The thiocyanate is prepared in solution by adding ammonium thiocyanate, NH_4SCN to HCl solutions of the chloride. Both basic and double carbonates are known. The former is precipitated from Sc^{3+} solutions by adding carbonate solutions, and is probably $Sc(OH)CO_3 \cdot H_2O$. The latter are obtained by the use of an excess of the soluble carbonate. Normal, basic, and double sulfates are known. The first exists in several degrees of hydration; the second is obtained as $Sc(OH)SO_4 \cdot 2H_2O$, by treating the normal sulfate tetrahydrate with the hydroxide. The alkali double sulfates and alums are obtained by treating the sulfate solution with an excess of the alkali (or ammonium) sulfate solution.

The nitrate is readily obtained by action of dilute HNO_3 on the hydroxide. In aqueous solution, the anhydrous nitrate yields a monobasic nitrate on heating.

To date, the applications for scandium and its compounds have been extremely limited, mainly because of its high reactivity and high cost. In exotic light sources, scandium iodide enhances luminosity. Minor inclusions of scandium are made in substituted yttrium garnets for electronic applications. At one time, it was felt that scandium might serve in a substitute fashion for aluminum, particularly for aircraft applications.

Scandium Clusters in Fullerene Cages. In mid-1992, C. S. Yannoni (IBM Research Division, San Jose, California) and associates reported that the research team had produced fullerenes containing multiple scandium atoms and obtained evidence that suggests that the atoms form molecules within the confines of the fullerene cage. Electron paramagnetic reonance (EPR) spectra have been obtained for metallofullerenes containing single and multiple (3) scandium atoms. The Sc_2C_n species detected by mass spectrometry appear to be diamagnetic. Because the metal molecules are protected from reaction by the surrounding fullerene shell, it is possible to examine them spectroscopically at ambient temperature rather than at the cryogenic temperatures required with the rare-gas matrix isolation technique. The researchers state: "An implication of this work is that production of macroscopic quantities of cluster-containing fullerenes may make possible the fabrication of exotic new structures with regular arrays of metal-atom clusters isolated in fullerene molecules, resulting in a new type of host/guest nano-structured material."

R. C. Vickery, M.D., D.Sc., Ph.D., Blanton/Dade City, Florida.

Additional Reading

Carter, G. F., and D. E. Paul: "Materials Science and Engineering," ASM International, Materials Park, Ohio, 1991.

Sax, N. R., and R. J. Lewis, Sr.: "Dangerous Properties of Industrial Materials," 8th Edition, Van Nostrand Reinhold, New York, 1992.

Staff: "ASM Handbook—Properties and Selection: Nonferrous Alloys and Pure Metals," ASM International, Materials Park, Ohio, 1990.

Staff: "Handbook of Chemistry and Physics," 73rd Edition, CRC Press, Boca Raton, Florida, 1992–1993.

Vickery, R. C.: "Chemistry of Yttrium and Scandium," Pergamon, London, 1960.

Vickery, R. C.: "Analytical Chemistry of the Rare Earths," Pergamon, London, 1961.

Vickery, R. C.: *Scandium, Yttrium and Lanthanum* in A. F. Trotman-Dickenson (editor): "Comprehensive Inorganic Chemistry," Pergamon, Oxford, 1971.

Yannoni, C. S., et al.: "Scandium Clusters in Fullerene Cages," *Science*, 1191 (May 22, 1992).

SCANNING TUNNELING MICROSCOPE. This microscopic technique (STM), invented by Gerd K. Binnig and Heinrich Rohrer (IBM Research Laboratory, Zurich, Switzerland) in 1981, generates topographic images of surfaces with atomic resolution. With the STM, scientists have obtained previously unseen images of gold, silicon, nickel, oxygen, and carbon atoms. STM views can reveal flaws and contaminants in atomic surface structure. Detailed views of three-dimensional atomic "landscapes" are improving and will continue to improve the knowledge of surface physics and chemistry and thus be of exceptional value in technical fields as varied as metallurgy, magnetism, semiconductor technology, and biology. Before development of the STM technique, scientists had been puzzled about the exact surface structure of silicon, the basic material from which computer chips are made. Many models of the silicon surface have been constructed. The STM has enabled researchers to sort out prior assumptions and hypotheses. In connection with gold, the STM has revealed a surface structure created by the spontaneous formation of ribbonlike facets, features that previously had not been distinguished with such clarity. The renowned physicist, Wolfgang Pauli, many years ago observed, "The surface was invented by the devil," with reference to the fact that the surface of a solid is its boundary or interface with the environment outside and beyond the solid. Whereas atoms within a solid interact with other atoms within the solid, those atoms on the surface can interact only with those atoms directly underneath and those atoms in the environment beyond the surface. Thus, surface atoms behave with different rules.

Instrumental means for investigating the atomic pattern of surfaces have included electron microscopy and later (1950s), the field-ion microscope (invented by Edwin W. Müller). These techniques continue, but are generously abetted by the STM. In 1984, the inventors of the STM received the Hewlett-Packard Europhysics Prize for outstanding achievements in solid state physics and the King Faisal International Prize in Science. In 1986, Binnig and Rohrer shared in the Nobel Prize in Physics.

Expanded Applications of STM

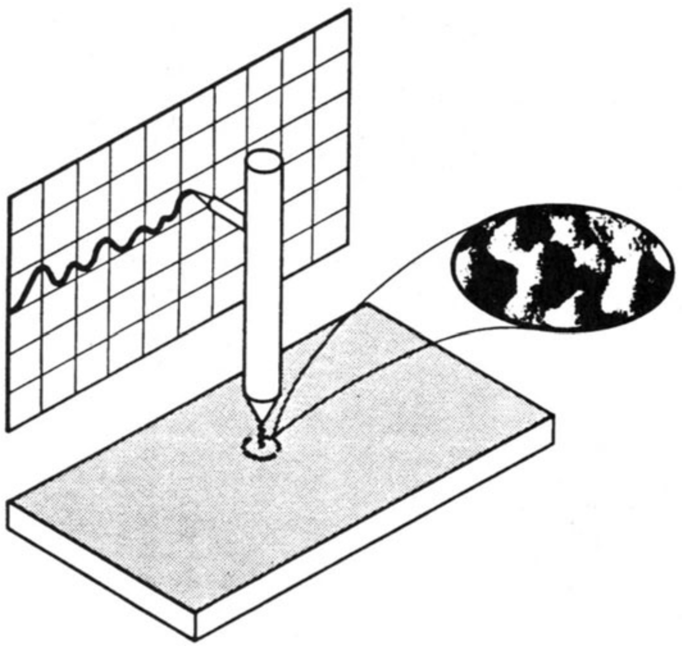
Since its introduction for practical uses in the early 1980s, the scanning tunneling microscope has found scores of applications that were not contemplated at the outset of STM technology. These applications include the manufacture of optical grating masters, the manufacture of recording thin-film magnetic recording heads, the manufacture of compact disk stampers, and the repair of costly masks used for integrated circuit manufacture—just to mention a few examples of practical usage in the electronics field. STM can be operated under water and other fluids, permitting the examination of biological materials in a more natural setting. The versatility of the STM is demonstrated by the appearance of hundreds of technical papers in the literature over a 1-year period. The STM can achieve lateral resolution of 1 to 2 angstroms and, in the vertical dimension, better than 0.05 angstrom. To overcome the relatively few limitations of STM, the technology has spawned several other microscopic techniques, including the atomic force microscope, the friction force microscope, the magnetic force microscope, the electrostatic force microscope, the attractive mode force microscope, the scanning thermal microscope, the optical absorption miscroscope, the scanning ion-conductance microscope, the scanning near-field optical microscope, the scanning acoustic microscope, and the molecular dipstick microscope—all resulting from the character of innovative thinking that scientists have come to apply to the application of new forms of energy to microscopy, once wholly dependent upon materials interactions with visible light. STM has played an invaluable role, not only in the development of new materials, but in understanding how surface atoms differ from the permanently embedded atoms of a material.

STM Fundamentals

As put forth in their excellent 1985 paper (reference listed), Binnig and Rohrer credit *electron tunneling* as the phenomenon that underlies the operation of the STM. As indicated by Fig. 1, an electron cloud occupies the space between the surface of the sample and the needle tip used. The cloud is a consequence of the *indeterminacy*[1] of the electron's location (a result of its wavelike properties). Because the electron is "smeared out" so to speak, there is a probability that it can lie beyond the surface boundary of a conductor. The density of the electron cloud decreases exponentially with distance. A voltage-induced flow of electrons through the cloud thus is extremely sensitive to the distance between the surface and the tip. As the STM tip is swept across the surface, a feedback mechanism senses the flow (the *tunneling current*) and holds constant the height of the tip above the surface atoms. In this way, the tip follows the contours of the surface. The motion of the tip is read and processed by computer and displayed on a screen or plotter. By sweeping the tip through a pattern of parallel lines, a high-resolution, three-dimensional image of the surface is obtained.

Views of the STM are shown in Figs. 2 and 3. Binnig and Rohrer describe the STM as having two stages, suspended from springs, that operate within a cylindrical steel frame. The microscope mechanism is contained by the innermost stage. Vibration is a severe problem, requir-

[1]The roots of atomic structure extend back over 50 years to the development of the concept of quantum mechanics and experiments in 1927 by Davisson and Germer, the researchers who confirmed the wave nature of the electron. The first experimental verification of tunneling was made about three decades ago by Ivar Giaever.



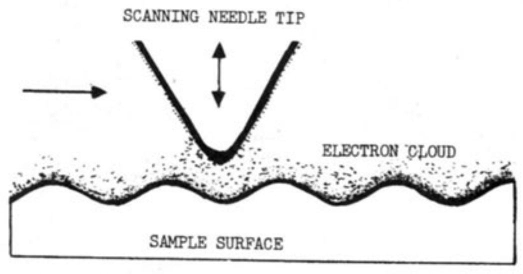


Fig. 1. As the probe tip of the STM is scanned across the microscopic "hills and valleys" of a surface, the vertical position of the probe is precisely adjusted to maintain a constant tip-to-surface distance. This is accomplished by keeping the tunneling current constant. Consequently, the probe follows the surface contour as it moves, yielding a 2-dimensional, enlarged representation of the surface contour for each such scan. A full 3-dimensional image is obtained by assembling an entire sequence of scans. (*IBM Corporation.*)

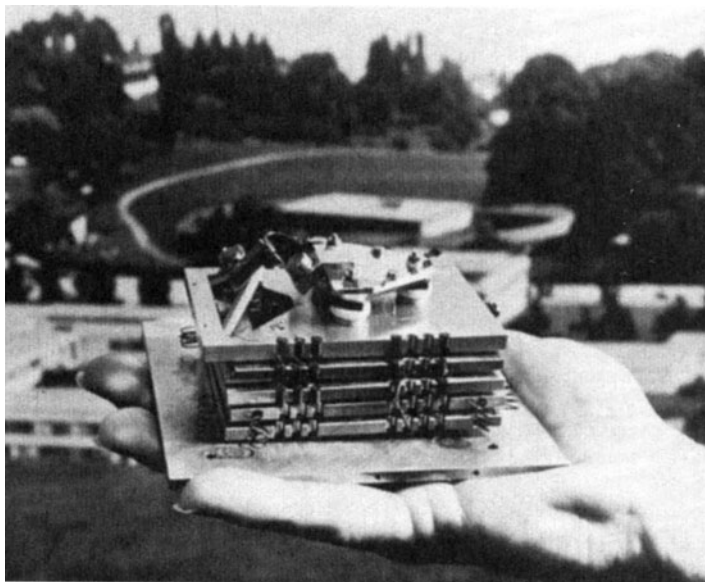


Fig. 3. Miniaturized version of the scanning tunneling microscope. The STM has become a very important tool for materials research and, in particular, to gaining new knowledge of computer chips. STMs are being used or are under construction by some fifty research groups worldwide in a broad range of physical, chemical, biological, and materials studies. (*IBM Corporation.*)

ing special preventive measures. To obtain images with high resolution (See Fig. 4) of surface structures, obviously the microscope must be protected from even very small vibrations as may be caused by sound and footsteps and other disturbances within the laboratory. To subdue vibration, copper plates are attached to the bottom of the stainless-steel frame and magnets are attached to the bottom of the inner and outer stages. Any disturbance causes the copper plates to move up and down in the field generated by the magnets. The movement induces eddy currents in the plates. Interaction of the eddy currents with the magnetic field retards the motion of the plates and thus the motion of the microscope stages. Where investigations require vacuum conditions, a steel cover is placed over the outer frame of the microscope.

The microscopic device incorporates a sample and a scanning needle. Piezoelectric materials (expand or contract with applied voltage) make it possible for the device to resolve features that are about a hundredth

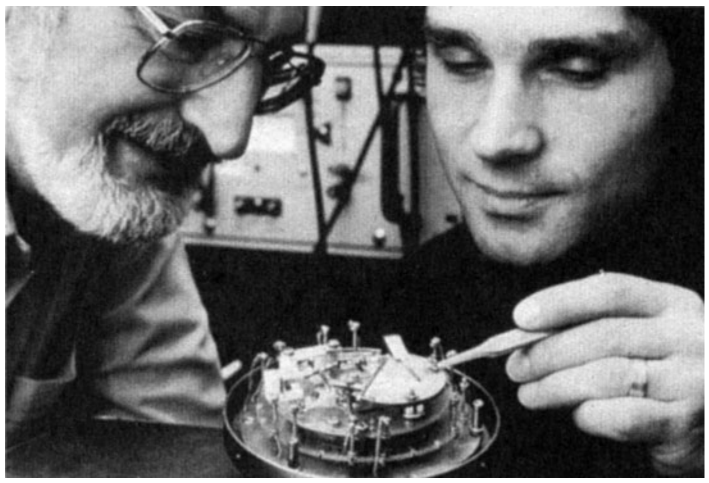


Fig. 2. Inventors Rohrer (left) and Binnig (right) shown adjusting the sample in the chamber of an early (1981) version of the scanning tunneling microscope. Later (1986), these scientists shared in the Nobel Prize in Physics for developing the STM instrument and technique. (*IBM Corporation.*)

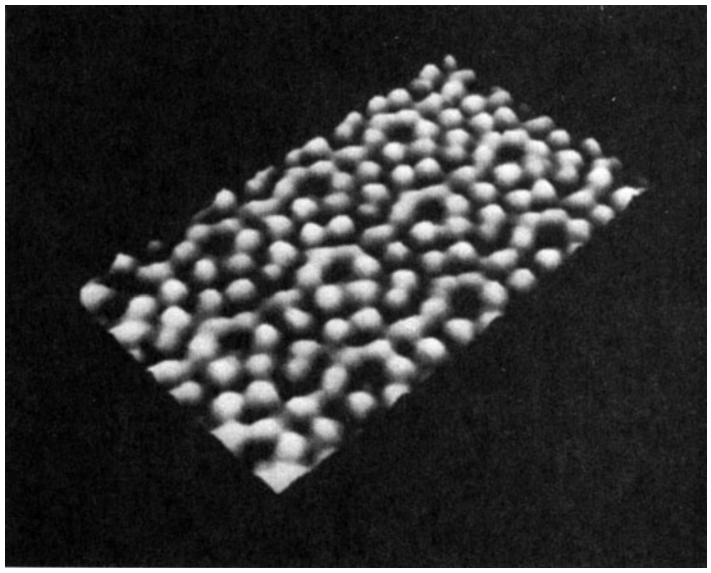


Fig. 4. Magnified millions of times are surface atoms of silicon. The image is computer-generated from data produced by a scanning tunneling microscope. (*IBM Corporation.*)

the size of an atom. A piezoelectric drive positions the sample on a horizontal metal plate and a piezoelectric tripod then sweeps the scanning needle over the surface of the sample, simultaneously achieving high stability and precision. See also Figs. 5 through 8.

Not only does the STM portray atomic topography, but it also reveals atomic composition. The inventors observe that the tunneling current depends both on the tunnel distance and the electronic structure of the surface and on the fact that each atomic element has an electronic structure unique to itself. The ability of the STM to resolve both topography and electronic structure assures wide use of the technique in numerous fields. Unlike other imaging techniques, the STM does not alter or partially destroy the sample. The STM already has demonstrated its utility in biology even though lateral resolutions of only 10 angstroms can be achieved. In this instance, the relatively poor resolving power of the microscope is more than compensated for by its ability to provide a direct and nondestructive method of viewing biological samples. Researchers at the IBM Zurich Research Laboratory and the Swiss Federal Institute of Technology have scanned the surface of DNA and observed a series of zigzags which correspond to its helical structure. Researchers at the Autonomous University of Madrid have made detailed examinations of viruses, notably phi 29 which measures 400 × 300 × 200 angstroms. It is currently visualized that the STM also will be very useful for testing electronic circuits, particularly as further reductions in size are achieved.

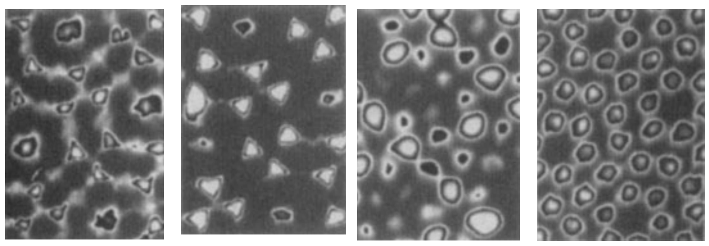


Fig. 5. Looking successively deeper into the atomic structure of silicon. Sequence shows surface magnified some ten million times, to a depth of about nine angstroms (36 billionths of an inch). Shown from left to right are the geometric positions of the atoms and three different classes of electronic bonds. (a) shows the position of the top atoms; (b) the dangling bonds that reach up from those atoms; (c) the dangling bonds that reach up from other atoms in the second layer in the surface; and (d) bonds (called "back bonds") that reach out sideways from the atoms in the second layer in the surface. (*IBM Corporation.*)

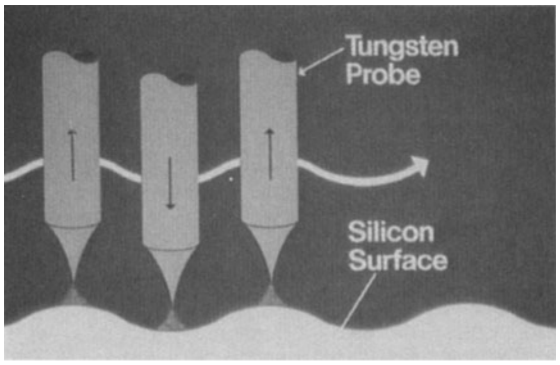


Fig. 6. As probe is scanned across silicon surface, its height above the surface is adjusted to keep the tunneling current constant. The monitoring of those height changes provides the desired topographic information. (*IBM Corporation.*)

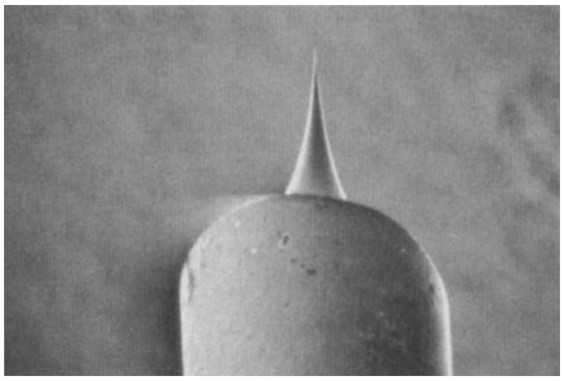


Fig. 7. Tungsten probe tip for scanning tunneling microscope as seen by a scanning electron microscope. Such probes, produced by field-ion microscopy, can be constructed with tips only one atom in width. (*IBM Corporation.*)

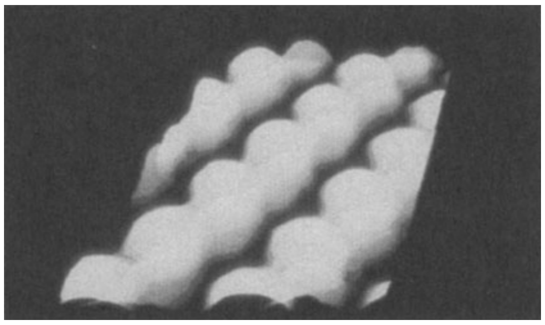


Fig. 8. Scanning tunneling microscope image of gallium arsenide (GaAs). Atoms to the left of each row are gallium; others are arsenic. (*IBM Corporation.*)

Additional Reading

Abelson, P. H.: "Phenomena at Interfaces," *Science*, 1357 (March 18, 1988).

Bard, A. J., et al.: "Chemical Imaging of Surfaces with the Scanning Electrochemical Microscope," *Science*, 68 (October 4, 1991).

Baro, A. M., Binnig, G., Rohrer, H., Gerber, C., et al.: "Real-Space Observation of 2 × 1 Structure of Chemisorbed Oxygen on NI(110) by Scanning Tunneling Microscopy," *Physical Review Letters*, **52**(15), 1304–1307 (April 9, 1984).

Beebe, T. P., Jr., et al.: "Direct Observation of Native DNA Structures with the Scanning Tunneling Microscope," *Science*, 370 (January 20, 1989).

Bindell, J. B.: "Elements of Scanning Electron Microscopy," *Advanced Materials & Processes*, 20 (March 1993).

Binnig, G., Rohrer, H., Gerber, C., and E. Weibel: "Facets as the Origin of Reconstructed Au(110) Surfaces," *Surface Science*, **131**, L379–L384 (1983).

Binnig, G., and H. Rohrer: "The Scanning Tunneling Microscope," *Sci. Amer.*, 50–56 (August 1985).

Binnig, G., and H. Rohrer: "The Scanning Tunneling Microscope," in *The Laurates' Anthology,* 72, Scientific American, Inc., New York, 1990.

Clemmer, C. R., and T. P. Beebe, Jr.: "Graphite: A Mimic for DNA and Other Biomolecules in Scanning Tunneling Microscope Studies," *Science*, 640 (February 8, 1991).

Epstein, A. W.: "A Tunneling Microscope Can Cleave Molecules," *Sci. Amer.*, 26 (April 1988).

Flam, F.: "Scopes with a Light," *Science*, 30 (April 5, 1991).

Giaever, I.: "Energy Gap in Superconductors Measured by Electron Tunneling," *Physical Review Letters*, **5**(4), 147–148 (August 15, 1960).

Hansma, P. K., et al.: "Scanning Tunneling Microscopy and Atomic Force Microscopy: Application to Biology and Technology," *Science*, 209 (1988).

Hansma, P. K., et al.: "The Scanning Ion-Conductance Microscope," *Science*, 641 (February 3, 1989).

Kinoshita, J.: "Scanning Tunneling Microscope Spawns Diverse Applications," *Sci. Amer.*, 33 (July 1988).

Pomerantz, M., et al.: "Rectification of STM Current to Graphite Covered with Phthalocyanine Molecules," *Science*, 1115 (February 28, 1992).

Pool, R.: "The Children of the STM," *Science*, 634 (February 9, 1990).

Pool, R.: "A New Role for the STM," *Science*, 130 (December 7, 1990).

Schardt, B. C., Yau, Shueh-Lin, and F. Rinaldi: "Atomic Resolution Imaging of Adsorbates on Metal Surfaces in Air: Iodine Adsorption on Pt(III)," *Science*, 1050 (February 24, 1989).

Smith, D. P. E., et al.: "Smectic Liquid Crystal Monolayers on Graphite Observed by Scanning Tunneling Microscopy," *Science*, 43 (July 7, 1989).

Takayanagi, K., et al.: "Structure Analysis of the Silicon(111) 7 x 7 Reconstructed Surface by Transmission Electron Diffraction," *Surface Science*, **164**, 367 (1985).

Tromp, R. M., and E. J. van Loenen: "Ion-Beam Crystallography on Silicon Surfaces III. Si(111)," *Surface Science*, **155**, 441 (1985).

Whitman, L. J., et al.: "Manipulation of Adsorbed Atoms and Creation of New Structures on Room-Temperature Surfaces with the Scanning Tunneling Microscope," *Science*, 1206 (March 8, 1991).

SCAPIE. See **Virus.**

SCAPOLITE. The mineral scapolite is a silicate of calcium and aluminum which contains also some potassium, sodium, and chlorine. The name identifies all intermediate members of a series with the following end members: Marialite $3Na(AlSi_3)O_8{\cdot}NaCl$, Meinoite $3Ca(Al_2Si_2)O_8{\cdot}CaCO_3$. Its tetragonal crystals are coarse and thick, often very large. It occurs also in massive forms. It has a distinct prismatic cleavage; subconchoidal fracture; is brittle; hardness, 5.5–6; specific gravity, marialite, 2.5–2.62, meionite, 2.72–2.78; luster, vitreous to rather dull, color, white to gray, red, green, blue, or yellow; translucent to opaque, rarely transparent.

Scapolite is found in the metamorphic rocks, particularly those rich in calcium; also in contact metamorphic deposits in limestones. It has been found in basic igneous rocks, probably as a secondary mineral. Notable localities are Lake Baikal, Siberia; Arendal, Norway; and Madagascar. In the United States it is found in Massachusetts, New York, and New Jersey. Grenville, in the Province of Quebec, Canada is an important locality. Superb transparent yellow gem crystals have recently been found in Brazil and Tanzania. Wernerite (scapolite) was named in honor of A. O. Werner, a famous German mineralogist (1749–1817).

SCARAB (*Insecta, Coleoptera*). A species of beetle which was regarded as sacred by the ancient Egyptians. The term is applied especially to the sculptured likenesses of the beetle. From this name the family *Scarabeidae* has arisen, containing, among the many North American species, the June bugs or May beetles, the tumble bugs, and the rose chafer.

SCARLET FEVER. Although staphylococci can induce a similar infection (scalded skin syndrome), Group A streptococci (*S. pyrogenes*) are essentially responsible for this disease, which is characteristically one of school children. There is no doubt that scarlet fever is caused by an erythrogenic toxin because its symptoms are reproduced by injection of filtrates of cultures of *S. pyrogenes* (the so-called Dick test). The disease is now regarded as a combination of the direct toxicity of the erythrogenic toxin together with a secondary toxicity due to hypersensitivity to a heat stable component of streptococcal toxin.

In the latter half of the 19th Century, scarlet fever was the most common cause of death in children over one year old and was often simply referred to as "the fever." Although mild infections still are quite common, most instances escape attention and the disease has now virtually disappeared in the developed countries.

The disease is usually easy to recognize. An abrupt fever (100–103°F; 37.8–39.5°C) is accompanied by a disproportionate tachycardia. Coughing is notably absent. As well as an initial tonsillitis, an exanthem develops on the palate and the tongue is covered by a white fur through which red-tipped papillae project (*strawberry tongue*). The fur later peels off, leaving a raw red surface (*raspberry tongue*).

The rash, which is attributed to capillary damage, usually appears on the chest, neck, and arms within 24 hours of onset and lasts from 2 to 20 days before peeling begins with desquamation of fine flakes or large areas of skin.

All group A streptococci are still highly susceptible to benzyl penicillin and this, with closely related antibiotics, is the first choice for therapy. When penicillin allergy is suspected, erythromycin may be prescribed. In untreated patients, the possibility of subsequent development of rheumatic fever cannot be excluded. Otherwise, prognosis is excellent.

R. C. V.

SCARLET TANAGER. See **Tanager.**

SCATTER DIAGRAM. Also called scattergram. A plot representing corresponding observed values of two variables x and y as points in rectangular coordinates. If the two variables are functionally related, the points will be bunched, but if they are not functionally related, the points will be scattered uniformly over the plane.

Scatter diagrams are used to explore the influence of one variable upon another, strong relationships being revealed as a concentration around a definite curve. The Hertzsprung-Russel diagram showing individual stars is an example of a scatter diagram. See also **Coordinate System.**

SCATTERING. In its general sense, this term refers to the redistribution of a group of entities, or bringing about a less orderly arrangement, either in position or direction. Thus when visible light enters a body of matter, however transparent, part of it is diffusely reflected or "scattered" in all directions. This is due to the interposition in the light stream of particles of varying size, from microscopic specks down to electrons, and the deflection of light quanta resulting from their encounters with these small obstacles. Similar effects are produced upon infrared, ultraviolet, x-rays, and other forms of electromagnetic radiation, and upon streams of particles such as cathode rays or alpha rays. More specifically, the term denotes the change in direction of particles or photons owing to collision with other particles or systems; it may also be regarded as the diffusion of a beam of sound or light (or other electromagnetic radiation) due to the anisotropy of the transmitting medium.

Coherent or elastic scattering, either of particles or photons, is scattering in which there are definite phase relations between incoming and scattered waves. Ordinary or Rayleigh scattering (defined below) is of this nature. In coherent scattering, interference occurs between the waves scattered by two or more centers. This type of scattering is exemplified by the Bragg scattering of x-rays and the scattering of neutrons by crystals, which gives constructive interference only at certain angles, called Bragg angles.

Rayleigh scattering is coherent scattering in which the intensity of the light of wavelength λ, scattered in any direction making an angle θ with the incident direction, is directly proportional to $1 + \cos^2 \theta$ and inversely proportional to λ^4. The latter point is noteworthy in that it shows how much greater is the scattering of the short wavelengths. These relations apply when the scattering particles are much smaller than the wavelength of the radiation. Thus the sky is blue, and tobacco smoke appears blue, because blue light is scattered more than red. The unscattered light is of course complementary to blue, i.e., orange or yellow—which explains the "warm" hues of the sunset.

Mie scattering is any scattering produced by spherical particles without special regard to comparative size of radiation wavelength and particle diameter; to be contrasted, therefore, with Rayleigh scattering.

Rutherford scattering is a general term for the process in which moving charged particles are scattered at various angles by interaction with the nuclei of atoms of a solid material. In Rutherford's original work, high speed α-particles from radon were focused in a narrow beam to strike a thin gold foil. Most of them pass through, but some are scattered.

On the other hand, *incoherent or inelastic scattering* is scattering in which the scattering elements act independently so that there are no definite phase relations between different parts of the scattered beam. The intensity of the scattered radiation at any point is determined by adding the intensities of the scattered radiation reaching the point from the independent scattering elements.

Thomson scattering is the scattering of electromagnetic radiation by free charged particles, computed either classically or according to nonrelativistic quantum theory. Scattering by electrons is interpreted classically as a process in which some of the energy of the primary radiation is reduced because electrons radiate when accelerated in the transverse electric field of the radiation. The scattering cross section is given by

$$\sigma_r = \frac{8}{3}\pi\left(\frac{e^2}{mc^2}\right)^2$$

which is 0.657 barn for an electron, and is called the *Thomson cross section*, or the *classical scattering cross section*.

Compton scattering is inelastic scattering of photons by electrons in the Compton effect; because the total energy and total momentum are conserved in the collisions, the wavelength of the scattered radiation undergoes a change that depends in amount on the scattering angle. If the scattering electron is assumed to be at rest initially, the Compton shift is given by the equation:

$$\lambda' - \lambda = \lambda_0(1 - \cos\theta) = (h/m_ec)(1 - \cos\theta)$$

where λ' is the wavelength associated with the scattered photons, λ is the wavelength of the incident photons, λ_0 is the (Compton) wavelength of the electron, and θ is the angle between the paths of incident and scattered photons.

Delbrück scattering is the scattering of light by a Coulomb field, a process which according to quantum electrodynamics occurs as a scattering of the light by the virtual electron-pairs produced by the Coulomb field. The total cross section is approximately 6 millibarns for uranium.

Single scattering is the deflection of a particle from its original path owing to one encounter with a single scattering center in the material traversed. This is to be distinguished from plural scattering and multiple scattering, which involve successive encounters with scattering centers.

Multiple scattering is any scattering of a particle or photon in which the final displacement is the sum of many displacements, usually small. A type of scattering intermediate between single and multiple scattering is called *plural scattering*.

Potential scattering is that portion of scattering by the nucleus of the atom that has its origin in reflection from the nuclear surface, thus leaving the interior of the nucleus undisturbed. The term usually is used in contradistinction to *resonance scattering*, which is the scattering arising from the part of the incident wave that penetrates the surface and interacts with the interior of the nucleus. In general, the scattered wave may have components arising from both kinds of scattering processes. The term potential scattering is also used to denote scattering of an incident wave by reflection at a change or discontinuity in the potential field.

Acoustic scattering is the irregular and diffuse reflection or diffraction of sound in many directions. Scattering frequently occurs when the reflecting surfaces or bodies are small compared with the wavelength of sound; in certain cases the reflecting bodies may be small inhomogeneities in the medium.

SCAUP. See **Waterfowl.**

SCAVENGING. 1. The use of an unspecific precipitate to remove from solution by adsorption or coprecipitation a large fraction of one or more undesirable radionuclides. Voluminous gelatinous precipitates are usually used as scavengers, e.g., $Fe(OH)_3$. 2. The removal of impurities from molten metal by addition of substances to form slags, or other compounds that can readily be removed. 3. The removal of unwanted gases from systems, e.g., of products of combustion from an internal combustion engine, or residual gases from an evacuated tube.

SCHEELITE. The mineral scheelite is calcium tungstate, $CaWO_4$, with molybdenum substituting for tungsten up to 25% in the molybdian scheelite variety. It is a tetragonal with an octahedral habit although also at times tabular, and may occur massive. It displays an octahedral cleavage; is brittle; hardness, 4.5–5; specific gravity, 6.1; luster, vitreous; color, white to yellowish, reddish, greenish and brownish; white streak; transparent to translucent. Scheelite is found in pegmatite and ore veins associated with granites, also as a contact metamorphic mineral. It is known from the Czech Republic and Slovakia, Saxony, Italy, Alsace, Finland; Cumberland and Cornwall in England; and Mexico. Crystals of exceptional length (6–10 inches; 15–25 centimeters) are found at various localities in Korea and Japan; and in the United States, in Connecticut, Colorado, South Dakota, Arizona, Nevada, and California. The Swedish chemist, Karl Wilhelm Scheele, discovered tungsten in this mineral, which later was named for him.

The mineral fluoresces vivid bluish white to white; or yellowish white with increasing molybdenum content under exposure to short-wave ultra-violet light.

SCHERING BRIDGE. See **Bridge Circuits (Electrical).**

SCHILBEIDS. See **Catfishes.**

SCHINUS TREE. A shrubby tree capable of reaching a height of about 30 feet (9.1 meters) for some years has posed a wildlife and vegetation threat in the ecosystem of Everglades National Park, Florida. The schinus tree (*Schinus terebinthifolius*) is one of several exotic and tropical specimens introduced into Florida in the early 1890s as an ornamental shrub. It became known as Brazilian holly, Florida holly, or Brazilian pepper. In the early 1900s, the plant was studied in some detail at the U.S. Department of Agriculture's plant-introduction station in Miami and its remarkable growth rate was noted. It was found that seedlings would reach a height of 15 feet (4.6 meters) within a period of three years, producing fruit in the second year. Dwellers in south Florida, notably in the Homestead area, were pleased with the manner in which great numbers of birds were attracted to the red berries of the schinus. It was soon noted, however, that the tree was taking root in locations where it was not planted or wanted. The schinus was found extremely difficult to control and it spread rapidly and profusely. It was also found that the schinus irritated the skin much as poison ivy does (not surprising in retrospect, because the schinus is closely related to the poison ivy and poison sumac family). It was inevitable that the schinus would spread into parts of Everglades National Park, crowding out native species. The schinus also has spread into stands of native slash pine. It tolerates flooding and drought and sprouts prolifically after fires. Other imported trees (all with a positive intent on the part of the original introducers) which have caused somewhat similar problems include the *casuarina* and *melaleuca* trees/shrubs. The latter two plants are described in separate entries in this volume.

SCHIST. The schists form a great group of metamorphic rocks chiefly notable for the preponderance of the lamellar minerals such as the micas, chlorite, talc, hornblende, graphite, etc. Quartz often occurs in drawn out grains to such an extent that a quartz schist is produced. Most schists have in all probability been derived from clays and muds which have passed through a series of metamorphic processes involving the production of shales, slates and phyllites as intermediate steps. Certain schists have been derived from fine-grained igneous rocks such as lavas and tuffs. Most schists are mica schists, but graphite and chlorite schists are common. Schists are named for the prominent or perhaps unusual mineral constituent, as garnet schist, tourmaline schist, glaucophane schist, etc. The word schist is derived from the Greek meaning to split, with reference to the easy separation of these rocks in a direction parallel to that in which the platy minerals lie.

SCHISTOSOMIASIS. Once known as bilharziasis, this is an invasion of the body by a genus of trematode parasites or flukes, the *schistosomae*. The disease reaches far back into history and ancient Egyptians even had a hieroglyphic symbol for it—a penis dripping blood. The disease was indeed so common that blood in the urine was considered a puberty symbol for males. Approximately 5% of the world's population is affected, with most cases appearing in Africa, the Middle East, Central and South America, China, the Philippines, and Malaysia. Many cases are also found in the Caribbean countries.

Infected persons excrete the microscopic schistosome eggs in their feces and urine. If these reach fresh water, tiny embryos hatch and invade an intermediary host, usually a snail. Therein, they develop and multiply and 6 to 8 weeks later the snail releases cercariae, or schistosome larvae, in numbers amounting to some tens of thousands per day.

Humans are infected by swimming in the water, the infective larvae penetrating the skin very rapidly to the blood stream where they travel to various organs. Species of schistosomes select different target sites in which to reside. S. *mansoni* targets the mesenteric veins of the upper intestine; S. *japonicum*, the mesenteric veins of the lower intestine; *S. hematobium*, the bladder veins. Once settled, the schistosomes produce a steady output of eggs (1000 or more per day) and live for years in their selected sites.

About half the eggs produced migrate through the wall of the intestine or bladder and the remainder are swept to the liver by the blood stream where they cause a severe inflammatory reaction. The eggs that do exit to the intestine or bladder are excreted and able to enter a new host.

Schistosomiasis patients frequently suffer vomiting and diarrhea. An infection by *S. mansoni* or *S. japonicum* develops a hard, cirrhosislike hepatomegaly; *S. hematobium* patients have intense bladder damage and heavy bleeding into the urine.

Diagnosis depends upon finding schistosome eggs in urine or feces or by serological tests, such as complement-fixation or ELISA (enzyme-linked immunosorbent assay).

Some progress has been made toward development of an anti-schistosome vaccine but, despite optimism, the body's immune system is continuously evaded by the organism's ability to camouflage itself with major histocompatibility antigens so that it resembles the host's own body. Meanwhile, praziquantel is the most effective drug against the parasite although metrifonate and niridazole are also of value.

R. C. Vickery, M.D., D.Sc., Ph.D., Blanton/Dade City, Florida.

SCHIZOPHRENIA. A major medical illness. It has been estimated that a child living at the age of 15 years statistically has a 2 to 3% risk of being diagnosed as a schizophrenic sometime during its lifetime. Persons with the disease in the United States exceed the population of New York City. No real medical breakthroughs in diagnosing and curing the disease have occurred. Inasmuch as the disease is not fatal, lifetime costs of care are astonishing.

If one were to condense all that is known concerning the disease, three elements would predominate: (1) the disease appears to run in families; (2) neuroleptics (drugs that act in some manner with the brain's dopamine system) help to improve the symptoms of the disease; and (3) there may be something structurally awry in the schizophrenic brain.

Fundamentals still are argued among the experts—does schizophrenia represent a discrete mental disorder, or is it just one component of a spectrum of mental illness. Ming Tsuang (Harvard Medical School) has expressed the opinion that present data seem to favor the traditional view that schizophrenia is made up of two discrete, major psychoses: (1)schizophrenia that starts during adolescence and usually worsens; and (2) affective (mood) disorders that occur later in episodes and are less likely to incapacitate the patient. Researchers are now looking to molecular genetics for evidence to prove the two stages. Still other researchers, for example Timothy Crow (Northwick Park Hospital, Harrow, England), views affective illnesses and schizoprenia at opposite ends of a continuum of psychotic disorders. Depression is at one end of the spectrum, followed by manic depression, mixed schizoprenia and affective disorder, and schizophrenia sans affective disorder. Crow observes that there are more patients in the middle of the spectrum than at the extremes.

Numerous studies of identical twins have been made to get a handle on the familial or hereditary characteristics of the disease. The evidence of a concentration is strong. If one identical twin has the disease, the chances of the other identical twin having the disease is greater than 50%. Another research finding reveals the less obvious, i.e., the children of schizophrenics who are adopted by nonschizophrenic parents surprisingly show a higher incidence of schizophrenia than a control population. These and other similar studies have not convinced all specialists, however, that there is a strong hereditary factor. Researcher Kendler has observed that it is not fair to call schizophrenia a genetic disease; it is not a classic Mendelian trait inherited like the gene for Huntington's chorea. But, one cannot deny evidence that like coronary heart disease and early-onset hearing loss, schizophrenia tends to run in families. Presently, there are two principal hypotheses regarding the inheritance of schizophrenia: (1) a model based on a single major gene locus (supported by Philip Holzman, Harvard University); and (2) a model that suggests multiple genetic causes, of which Tsuang is an advocate. Some researchers consider that poor eye tracking may be a biological marker for schizophrenia. Others claim that poor eye tracking stems from other causes as well, including drug-induced effects, Parkinson's disease, multiple sclerosis, and some brain lesions.

Crow has suggested that a retrovirus may be involved. Perhaps the virus may incorporate into the human genome and thus be capable of passage from one generation to the next. There have been other viral postulates, supported by a survey which indicates over half of the schizophrenic population were born in winter or spring (virus season). Other researchers have associated the disease with the immune system. Although very generalized, Sedvall (Karolinska Institute, Stockholm) observes that schizophrenia probably represents a complex interplay of genetic and environmental factors.

More recent research targets include direct studies of the brain—to observe blood flow, brain metabolism, and receptor mapping to monitor brain structure. The effects of various drugs can be observed with key instrumental tools in connection with human and experimental animals. These tools include computerized tomography (CT) scans, magnetic resonance imaging, and photon imaging.

Symptoms and Treatment

Several decades ago, schizophrenia was called *dementia praecox.* This terms means "a precocious demented state." In 1911, Blueler advocated the substitution of *schizophrenia—schizo* standing for splitting and *phrenia* for the mind, thus indicating a "breaking away" of the patient's mind from a normal evaluation of reality. There are several forms in which the illness manifests itself. However, denial of reality and inappropriate emotional responses are the most common symptoms. The distorted psyche is revealed by the patient's bahavior. During some spans of time, the patient may be wild in behavior, as illustrated by the breaking up of furniture and the throwing of portable objects. During some periods, a patient may rip off clothing and go naked; or, during other periods, dress in a fantastic manner. The patient may laugh or cry without apparent cause. The patient may use a language consisting of jumbled fractions of words and phrases that are incomprehensible.

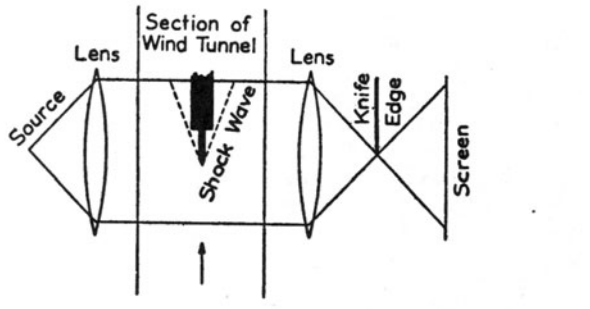
The patient may be confused as to identity and make fantastic claims as being someone of high repute. In all, the actions and mannerisms of the schizophrenic appear bizarre and unintelligible when viewed in the light of the real world. The actions are more easily understood when one realizes that this behavior is the product of a dream world, erected because the patient does not have the ability to perceive reality in a normal way.

When a schizophrenic becomes unable to find any solution which will enable the acceptance of a painful situation, the patient's normal defenses break down entirely and "reality" is imagined in terms of desires. Daydreams offset the poverty of true relationships and thus become more gratifying than reality.

Traditionally, patients who present the symptoms of acute schizophrenia are referred to a psychiatrist. A wide variety of antipsychotic drugs is available. Generally, these drugs can be used safely with most patients because of the wide difference between therapeutic and lethal doses. Further, most of these drugs are not addictive and, in most situations, the patient does not develop a tolerance (thus decreasing the effectiveness) to the antipsychotic effects of these drugs. Among the antipsychotic agents commonly used are the phenothiazines, thioxanthenes, dibenzoxazepines, butyrophenones, and indolones. Certain generalizations can be made concerning these drugs—rapid absorption $\frac{1}{2}$ to 1 hour with oral doses; 10 minutes with intramuscular administration); oxidative metabolism, usually in the liver, with byproducts excreted in the urinary tract. There are side effects of these drugs, particularly in elderly patients.

Often a considerable period of time will be required to find the most satisfactory antipsychotic agent for a given patient. With some of these drugs, however, a minimum period of three months is required to determine the full clinical response. Proper dosage of a given drug also varies considerably from one patient to the next. Acute schizophrenia may not require indefinite maintenance. In chronic schizophrenia, the drug maintenance program is long-term.

SCHLIEREN. Refraction anomalies produced in transmitted light by differences in density or other anisotropy in parcels or strata of air (or other fluids). All the natural scintillation phenomena in the atmosphere result from density schlieren developed by turbulent processes. Schlieren optical systems are frequently used for observing and photographing colloidal refractive index gradients which arise in the study of electrophoresis, diffusion, sedimentation velocity, chromatography, and airflow in wind tunnels. The figure shows a diagram of schlieren measurements in a wind tunnel, in which an interferometer is used to determine the interference effects produced by turbulence. Steady flow of air at constant density does not show on the interference pattern, but turbulent flow, shock waves and similar density variations do appear in



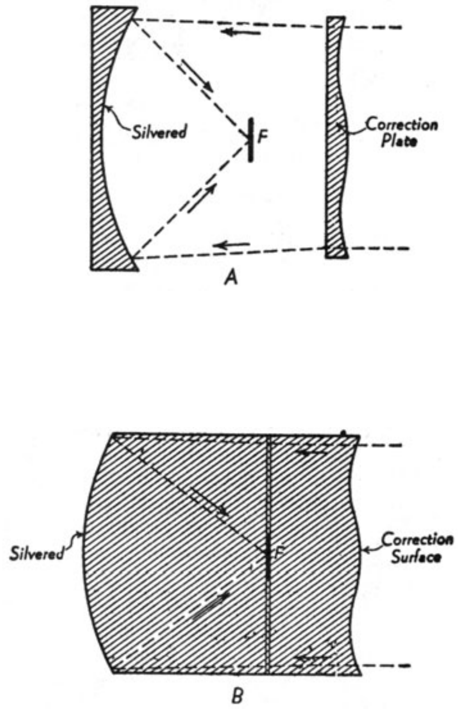
Idealized Schlieren apparatus for observing shock wave in supersonic tunnel.

the interference pattern. The word *schlieren* is used to describe the whole phenomenon. The interferometer plates are called *schlieren plates*, the interference pattern is called a *schlieren picture*, the whole instrument is called a *schlieren interferometer* or simply *schlieren apparatus*.

SCHMIDT OBJECTIVE. An objective for reflecting telescopes, designed to correct the aberration of the spherical mirror without introducing the coma (blurring) to which even a parabolic reflector is subject for wide fields. The results are obtained in somewhat the same way that spectacles correct for defects in vision. A slight chromatic aberration is introduced.

The Schmidt objective as originally designed consists of a concave spherical mirror, functioning in the same way as the objective of any reflecting telescope, but with a plate of glass interposed in front of it perpendicular to its axis at its center of curvature. This glass plate is not plane, but has one surface "figured" in such a way that, as the rays pass through it on their way to the mirror, it so modifies their course as to effect almost perfect correction for the spherical aberration and coma which the mirror would otherwise produce. In a later design, the objective consists of two coaxial cylinders of glass, in contact along a plane perpendicular to the axis. The rear surface of the rear piece is spherically convex and is silvered on the outside, thus presenting a concave spherical mirror to the interior of the glass cylinder. The front surface of the front piece, passing through the center of curvature of the mirror, is the correction surface, serving the same purpose as the glass plate in the older design. The reason for using two pieces is that the final real image F is of course produced between the mirror and the correction surface; and it is here that the plane of separation is located, so that the small photographic film used may be introduced. The general nature of these arrangements will be clear from the figures. There are numerous other variations of this idea, all referred to as Schmidt telescopes.

See also **Telescope (Astronomical-Optical).**



Schmidt objective: (A) Original design; (B) the "solid" Schmidt (diagrammatic).

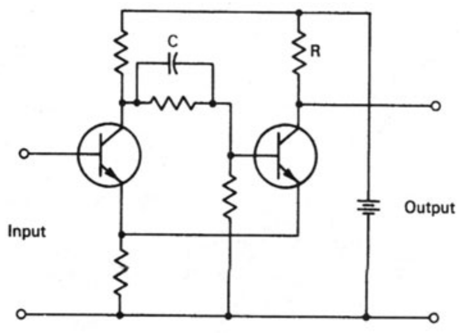
SCHMIDT PROCESS. A method for converting a given set of vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ into an orthonormal set $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$. If the length of $\mathbf{v}_i$ is l_i and $\mathbf{e}_i = \mathbf{v}_i/l_i$ is a unit vector, the calculation may be performed with the recursion formula

$$\mathbf{v}_{i+1} = \mathbf{u}_{i+1} - \sum_{k=1}^{n} (\tilde{\mathbf{e}}_k \, \mathbf{u}_{i+1})\mathbf{e}_k$$

where $\tilde{\mathbf{e}}_k$ is the transpose of $\mathbf{e}_k$. The procedure may also be used for functions with the appropriate integrations in place of summations. If, for example, the original functions are $1, x, x^2, \ldots$, defined over the range $x = \pm 1$, the orthonormal set obtained from them by the Schmidt process is a set of Legendre polynomials.

See also **Legendre Differential Equation.**

SCHMITT TRIGGER. A form of bistable multivibrator which furnishes a fast-acting on-off switch action, is capable of generating rectangular pulses from sinusoidal or other nonrectangular waveforms, and can determine when a signal reaches a specified dc level. See accompanying diagram.



Schmitt trigger.

SCHOTTKY DEFECT. A lattice vacancy created by removing an ion from its site and placing it on the surface of the crystal. For electric neutrality, the number of cation Schottky defects must equal the number of anion Schottky defects. The number, n, of Schottky defects is given by

$$\frac{n}{N-n} = C_s e^{-W/kT}$$

where there are N lattice points, and W is the energy required to remove an ion from a lattice point, and then add it to the surface. C_s is a numerical factor of the order of $10^3 - 10^4$. This relation may be derived from thermodynamic arguments. It can be shown that the factor C_s includes a vibrational entropy term.

SCHRÖDINGER EQUATION. The basic equation of wave mechanics. It is developed by using the de Broglie wavelength in the description of a particle and then associating with the measurement of the energy E or of the x-component of momentum p_x of the particle a differential operator

$$E = i\hbar\frac{\partial}{\partial t} \quad \text{or} \quad p_x = -i\hbar\frac{\partial}{\partial x}$$

where is the Dirac **h.** The Hamiltonian function can be expressed either in terms of total energy or in terms of potential energy and momentum. Expressing it in both ways, one obtains:

$$-\frac{\hbar^2}{2m}\nabla^2\psi + V(\mathbf{r})\psi = i\hbar\frac{\partial\psi}{\partial t} = E\psi$$

where ∇^2 is the Laplacian, m the mass of the particle, E, its total energy, and $V(\mathbf{r})$ its potential energy (usually a function of position). This is the time dependent Schrödinger equation for ψ.

In many instances, we are interested in the allowed values of E in stationary states of the system. Using the Planck law we may set $E = h\nu = 2\pi\hbar\nu$ and write

$$\psi = \phi(\mathbf{r})e^{2\pi i\nu t} = \phi(\mathbf{r})e^{iEt/h}$$

where $\phi(\mathbf{r})$ is a function of position only. We then obtain the time independent equation:

$$\left[\nabla^2 + \frac{2m}{\hbar}\{E - V(\mathbf{r})\}\right]\phi = 0$$

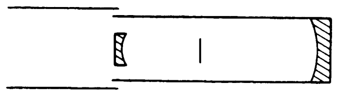
It is often found that solutions of the equation exist only for specific eigenvalues of E. To each eigenvalue E_n there corresponds an eigenfunction of the coordinates ϕ_n. The probability of finding the particle in a region of volume dV is

$$\int |\phi|^2\, dV = \int \phi\phi^*\, dV$$

assuming that ϕ has been normalized so that the integral over all space is unity.

See also **Quantum Mechanics.**

SCHWARZSCHILD TELESCOPE. A reflecting telescope using two mirrors designed to be free of spherical aberration and coma. It requires a large secondary mirror which causes a loss of light and a long foretube to act as a light baffle. See also **Telescope.**



Schwartzschild telescope.

SCINTILLATION (Astronomical). Any scintillation phenomena, such as irregular oscillatory motion, variation of intensity, and color fluctuation observed in the light emanating from an extraterrestrial source; to be distinguished from terrestrial scintillation primarily in that the light source for the latter lies somewhere within the earth's atmosphere. Also called *stellar scintillation*. See also **Seeing (Astronomy).**

SCINTILLATION (Atmospheric). Generic term for rapid variations in apparent position, brightness, or color of a distant luminous object viewed through the atmosphere. If the object lies outside the earth's atmosphere, as in the case of stars and planets, the phenomenon is termed *astronomical scintillation*; if the luminous source lies within the earth's atmosphere, the phenomenon is termed *terrestrial scintillation*. As one of three principal factors governing astronomical "seeing" (i.e., the disturbing effects produced by the atmosphere upon the image quality of an observed astronomical body), scintillation is defined as variations in luminance, only.

It is clearly established that almost all scintillation effects are caused by anomalous refraction occurring in rather small parcels or strata of air (schlieren), whose temperatures and, hence, densities differ slightly from those of their surroundings. Normal wind motions transporting such schlieren across the observer's line of sight produce the irregular fluctuations characteristic of scintillation. Scintillation effects are always much more pronounced near the horizon than near the zenith. Parcels of the order of only centimeters to decimeters are believed to produce most of the scintillatory irregularities in the atmosphere.

See also **Schlieren.**

SCINTILLATION COUNTER. Radiation detectors that respond by emitting a flash of visible light are classified as scintillators or fluors. Typical scintillators used for counting gamma radiation are thallium doped crystals of sodium iodide, NaI(TI). Plastic (organic) scintillators are commonly used for counting beta radiation, and for counting alpha particles; zinc sulfide, ZnS, is a good detecting crystalline substance.

SCLERENCHYMA. A tissue composed of thick-walled cells of various forms. These cells have small pits and walls so thick that in many cases the cavity of the cell is nearly obliterated. Mature sclerenchyma cells are dead, containing no protoplasm. They serve to strengthen the part of the plant in which they are found or to protect the more delicate structures within. Sclerenchyma cells are of two kinds, stone cells and fibers. Stone cells are small, irregular in shape, and only slightly if at all elongated. They may be found in the cortex of the stem or elsewhere in the plant, but are particularly abundant in the endocarp of certain fruits and in seeds. The flesh of pears and blueberries contains many grit particles, which are groups of stone cells. Fibers are very much elongated cells, generally with long pointed ends and with simple pits in the walls. Hemp and flax are bundles of fibers of great value to people.

SCLERODERMA. This may be a localized disease (*scleroderma*) or a diffuse disease (*progressive systemic sclerosis*) which is caused by the deposition of fibrous connective tissue in the skin. Scleroderma is classified as one of the rheumatic diseases. The skin becomes thick and there is moderate swelling of affected tissue, followed by a deadening and wasting away of skin with accompanying loss of hair follicles, sebaceous and sweat glands. The skin tightens, loses pliability, and may be described as hide-like. In the course of the disease, there may be involvement of the lungs, heart, kidneys and other organs. Infrequently, systemic complications can be fatal. Where only the skin is affected, the disease may not be the primary determinant of life span. Sclerodermatous skin is seen most frequently in the hands and fingers and somewhat less frequently in the facial area, this latter involvement frequently reducing the size of the mouth. However, in some patients, the disease may be diffusive and involve all of the skin of the body.

The systemic aspects (sclerosis) of scleroderma are usually more serious than simple scleroderma. When the *musculoskeletal system* is involved, there may be a mild form of arthritis that is similar to rheumatoid arthritis. Sometimes *acral osteolysis* (shortening of fingers and toes) may develop. Also, there may be weaking of skeletal muscles. Scleroderma may involve the *gastrointestinal tract*, leading to diverticula. Where smooth muscle of the esophageal mucosa is replaced by fibrous tissue, *dysphagia* (difficulty in swallowing) may result—as well as gastric reflux and peptic esophagitis. *Lung involvement*, in which fibrosis occurs in the walls of the pulmonary arteries, may cause pulmonary hypertension, a condition that can lead to progressive respiratory failure. In *heart involvement*, fibrous tissue replaces extensive areas of cardiac muscle, precipitating arrhythmias, conduction disturbances, and congestive heart failure. In *kidney involvement*, scleroderma may cause chronic mild proteinuria and mild hypertension. The syndrome known as *malignant hypertension* may precipitate oliguric renal failure.

The etiology of scleroderma is unknown. The abnormal amount of collagen present in this disease appears to be normal rather than altered collagen.

SCLEROMETER. An apparatus for determining the hardness of a material by measuring the pressure on a standard point that is required to scratch the material. A scleroscope is a similar apparatus which measures hardness by determining the rebound of a standard ball dropped on the subject material from a fixed height.

SCOLECITE. This mineral is a zeolite, a hydrous calcium-aluminum silicate, $CaAl_2Si_3O_{10} \cdot 3H_2O$. It occurs in slender monoclinic prisms and in fibrous and nodular masses. Hardness is 5; specific gravity, 2.27; luster vitreous to silky; transparent to translucent. When heated, some specimens of scolecite curl up like worms; hence its name, derived from the Greek meaning a worm. This mineral occurs with other zeolites, at Baden, Switzerland; Iceland; Greenland; the Deccan region of India; and in the United States at Golden, Colorado, and Paterson, New Jersey. Single crystals up to 12 inches (30 centimeters) in length have recently been found in a single large cavity in the basaltic trap rocks near Nasik, India.

SCOLIOSIS. A deformity, usually of the spine, in which there is abnormal lateral displacement of the spine (with at least one other compensatory curve in the opposite direction), which in extreme conditions causes what is commonly termed "hunchback." In *Friedreich's ataxia*, a hereditary spinocerebellar degradative disease, scoliosis is present. In some cases (uncommon), scoliosis can cause pulmonary or cardiac insufficiency, or both. Thoracic scoliosis is the result of a developmental anomaly and is precipitated by weakness in the paravertebral muscles.

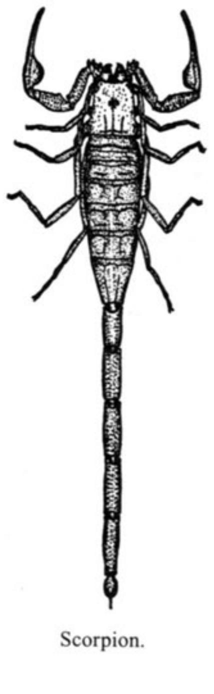
SCOMBROID POISONING. See **Foodborne Diseases.**

SCORIA. The term applied to lava which is highly vesicular and slaggy in appearance, due to the escape of the volcanic gases while the lava is still viscous. Scoria may be considered as a very coarse variety of pumice, the vesicles occupying approximately the same amount of space as the solid material, and extremely variable in size and shape.

SCORODITE. This hydrated arsenate of ferric iron and aluminum $(Fe^{3+}, Mg^{3+})AsO_4{\cdot}2H_2O$, crystallizing in the orthorhombic system, is the iron-rich isomorphous end member of a complete series extending to the aluminum-rich mineral Mansfieldite. Crystals usually occur as drusy crusts. Also occurs as massive, compact, and earthy material. Hardness of 3.5–4, with specific gravity of 3.278. Vitreous to subadamantine luster, of pale green to liver-brown color.

Scorodite occurs as a secondary alteration mineral in the oxidized zone of metallic arsenic-containing veins. The mineral also may be a product of deposition from certain hot springs. World localities of note include Siberia; Laurium, Greece; Carinthia; Cornwall, England; and Nevada, Utah in the United States. Currently being deposited by hot springs at Yellowstone National Park in Wyoming.

SCORPION (*Scorpionida*). A terrestrial arthropod with a conspicuous pair of pincers and a slender terminal region of the body bearing a clawlike sting, and the order made up of these animals. They are grouped with the spiders, ticks and other forms in the class *Arachnida*.


Scorpion.

The order is characterized by the conspicuous pincers and by the sting. The body is divided into cephalothorax and abdomen, and the latter consists of a broad anterior portion and a slender post-abdomen. The sting, a modified telson, bears the opening of the duct of a poison gland. The genital ducts open on the ventral surface of the first abdominal segment, just in front of a pair of comb-like pectines of the second segment which are regarded as accessory reproductive organs.

Scorpions are common in warm dry regions. In the United States they occur as far north as Kentucky. Their poison is not virulent as a rule.

SCORPION FLY (*Insecta, Mecoptera*). Moderately large insects with four membranous wings. They are named from the peculiar modification of the terminal segments of the abdomen, which fairly resembles that of the scorpions. The apparent sting is, however, made up of the external genital organs. See also **Mecoptera.**

SCORPIUS (the scorpion). Scorpius is the eighth sign of the zodiac. The constellation is rather far south for observation in Europe and North America, but is a beautifully grouped constellation presenting more resemblance to the figure for which it is named than is the case with most of the others.

The brightest star in the group Antares (α Scorpii) is one of the most beautiful stars in the sky. It is distinctly reddish in color and gets its name from the fact that it opposes or rivals Mars, the red planet, in color. It is the largest star whose diameter has thus far been measured, having a diameter approximately 450 times that of the sun. The star is so large that if Antares were in the position of the sun it would extend out beyond the planet Mars. (See map accompanying entry on **Constellations.**)

SCREAMER (*Aves, Ciconiiformes*). Peculiar South American birds with moderately long legs and large feet. The beak is like that of the domestic fowl and the wings are provided with two stout spurs on the front margin. The birds are as large as geese and swim readily, although the toes are not webbed. See also **Anseriformes.**

SCREW-WORM (*Insecta, Diptera*). The term screw-worm is usually reserved for those maggots of flies that attack a wounded or diseased animal, as contrasted with the egg-laying and larval activities of those flies (such as bot fly) which attack a healthy animal for the purpose of laying eggs within its body just below the skin. Screw-worms are of the family *Calliphoridae*, order *Diptera*. One species is *Callitroga* or *Cochliomyia hominivorax* (Coquerel), or *C. americana*. The adult screw-worm fly appears much like a house fly, but is about double its size. There are 3 prominent black stripes on the back behind the head. This fly lays its eggs on the edges of wounds of injured animals. The resulting maggots first feed on wound tissue and then proceed to sound tissue, a process aided by hooks in their mouth parts. Thus, old wounds do not heal, but spread and the affected animal becomes sullen and withdrawn. A chain reaction is activated because the odor from the spreading wound attracts more and more of the screw-worm flies to further the process. An untreated animal can die in relatively short order because, in this condition, the animal usually will not feed.

The species, *Callitroga macellaria* (Fabricius), known as the *secondary screw-worm*, lays its eggs on the bodies of dead animals, but also attacks wounded animals. The species is responsible for a very high percentage of the animal screw-worm attacks in the southern states. Reactions from an infestation can include meningitis and peritonitis. Serious epidemics, involving tens of thousands of animals, have occurred in the Gulf States. Losses in some years run into the several millions of dollars.

Control is aided by continuing inspections of animals for any breaks in their skin that will attract the flies. Operations, such as dehorning, castrating, earmarking, docking of lamb tails, branding, etc., should be performed only during late fall and winter when flies are not active. Dogs should not be trained to bite or nip at livestock. Unavoidable injuries should be treated immediately. Infested wounds can be treated by applying smears of various formulations directly to the wound, using a small paint brush. The material should be thoroughly worked into any pockets where maggots may be hiding.

SCROPHULARIACEAE. This family contains some 2,500 species, most of which are herbs or shrubs. Its members are numerous in temperate regions, where many of them are common plants, as for example mullein, "butter-and-eggs," speedwell, and lousewort. Annuals, biennials and perennials are found in the family.

The flowers are zygomorphic, or bilaterally symmetrical, with the calyx and corolla both tubular and each composed of four or five lobes. In many plants of this family the corolla is distinctly 2-lipped, as in the Snapdragon. Usually there are four stamens, which are inserted on the corolla tube. The ovary, composed of two united carpels, becomes a dry capsule containing many small seeds. The flowers of this family are mostly pollinated by insects, such as bees, wasps, and flies, which seek the nectar secreted in a disk at the base of the ovary.

In this family are found many plants grown by man as ornamentals. Some, such as Foxglove (*Digitalis*) and *Veronica*, are hardy biennials, or perennials; others, like Snapdragon (*Antirrhinum*), are not hardy; while *Calceolaria*, a native of South America and Mexico, is a hothouse plant grown for its bizarre sac-like flowers of brilliant color. Drugs of medicinal value are also found in several plants of this family, the most important being digitalis, from species of foxglove. In early days many species were used as a source for homemade brews. Others are poisonous herbs.

SCUD. See **Clouds and Cloud Formation.**

SCULPINS (*Osteichthyes*). Of the order *Scleroparei*, family *Cottidae*, sculpins are peculiar fishes, usually small, with a broad depressed head and large pectoral fins. They are of no importance as food fishes. Many of the included species bear other names, including the little miller's thumbs of freshwaters and the sea raven which ranges from Cape Cod to the Arctic. One species of the northern Atlantic coast is called the big sculpin, or daddy sculpin. The *Hemilepidotus hemilepidotus* (spotted irishlord) occurs in the American Pacific and reaches a length of about 20 inches (51 centimeters). *Enophrys bison* (buffalo sculpin) also occurs in the American Pacific. One of the larger sculpins is *Scorpaenichthys marmoratus* (crab-eating cabezon) which attains a length of about 30 inches (76 centimeters) and weighs up to 25 pounds (11 kilograms). It is found in Pacific waters from Lower California northward to British Columbia. Although the roe is poisonous, the flesh is considered good (even if a green color). The *Leptocottus armatus* (Pacific staghorn sculpin) is common and found in shallow waters of the Pacific coast, usually quite abundant in bays ranging from Lower California northward to Alaska. Their size usually does not exceed 6 inches (15 centimeters). *Hemitripterus americanus* (Atlantic sea raven) has qualities like a puffer in that it can inflate itself by swallowing air when taken from the water. Although rumored to be edible, its major use is for baiting lobster traps.

The grunt sculpin (*Rhamphocottus*) is of the family *Rhamphocottidae* and is named for the grunting noise it makes when taken from the water. Measuring only about 3 inches (7.5 centimeters), the grunt sculpin ranges in Pacific waters from northern California to Alaska. It is well known for its vertical temperature distribution. This sculpin does well in aquariums.

SCUP. See **Porgies.**

SCURVY. See **Ascorbic Acid (Vitamin C).**

SCYPHOZOA. The jellyfishes. A class of the phylum *Coelenterata* made up entirely of marine animals which are, with very few exceptions, floating forms. The jellyfishes represent the highest development of the medusa form of coelenterates, and have lost the polyp stage with the exception of the reduced hydratuba larva.

Jellyfishes owe their name to the great development of the middle layer of the body (mesogloea), which is a bulky and jelly-like mass. They contain a high percentage of water, sometimes as great as 96%, and are consequently soft-bodied and without rigid support. In the water, however, they are delicate and beautiful. Many are filmy transparent creatures while others are beautifully colored. They are found at various depths and in various seas, and in size they range from species less than 1 inch in diameter to the large *Cyanea* with a body 6–7 feet (1.8– 2.1 meters) in diameter and tentacles 120 feet (36 meters) long. They are of no economic importance.

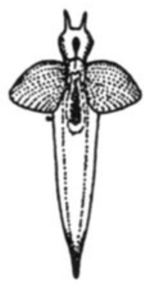
SEA ANEMONE. A complex polyp of the class Anthozoa (*Actinozoa*). Although closely related to the alcyonarians and corals the sea anemones are usually solitary and in some species are large and beautifully colored. They are without hard supporting structures such as the related forms possess.

SEA ARROW (*Mollusca, Cephalopoda*). Small slender squids, *Omnastrephes*, which swim very rapidly. Also called flying squids.

SEA BASS. See **Bass.**

SEA BEAR. See **Sea Lions and Seals.**

SEA BUTTERFLY (*Mollusca, Gasteropoda*). Mollusks with the foot formed into two wing-like lobes which propel the animal through the sea by slow flapping movements. They make up the order *Pteropoda*.


Sea butterfly.

SEA COWS (*Mammalia, Sirenia*). The manatee and dugong, collectively sea cows, constitute an order of mammals, *Sirenia*. The manatee is a fully aquatic animal with a horizontally-flattened oval tail. There are no hind limbs, and the fore limbs are developed as flippers. See accompanying figure. They have a superficial similarity to whales but differ in many details of structure. They live only in shallow coastal waters and estuaries and eat aquatic plants. The several species of the genus *Manatus* are distributed on both shores of the Atlantic and in the Oriental and Australian regions. *M. latirostris* is found off Florida. The dugong is related to the American manatee and is found along the shores of the Oriental region. This animal (*Halicore*) has a blunt muzzle, a broad horizontal tail, and pectoral flippers. Principal diet is seaweed. The dugong is hunted for its flesh and oil.

Manatee (sea cow). (*A. M. Winchester.*)

SEA ELEPHANT. See **Sea Lions and Seals.**

SEA FAN (*Coelenterata, Anthozoa*). Marine polyps of the order *Alcyonaria* whose colonies are in the form of thin lacy fans.

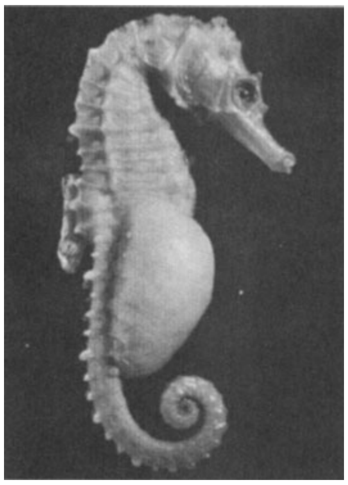
SEA FEATHER (*Coelenterata, Anthozoa*). Marine polyps of the order *Alcyonaria*. The colony has a central stalk bearing lateral branches, the whole resembling a feather.

SEAFOOD POISONING. See **Foodborne Diseases.**

SEA GULL. See **Petrels and Albatrosses; Shorebirds and Gulls.**

SEA HARE (*Mollusca, Gasteropoda*). A marine mollusk of oval form with two ear-like tentacles near the anterior end which give it this name. The mantle almost conceals the shell and the foot forms two lobes by which the animal swims. Sea hares live on seaweed.

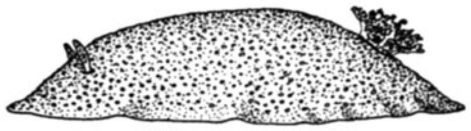
SEAHORSES (*Osteichthyes*). Of the order *Solenichthys* (tube-mouthed fishes), family *Syngnathidae*, seahorses are of the same family as the pipefishes. See also **Pipefishes (Osteichthyes).** Because of what might be termed their "cute" figures and resemblance to equine mermaids, seahorses have been a favorite of commercial and scientific aquarium operators and would also be of the fish hobbyists were it not for the difficulties encountered in maintaining them in tanks. See accompanying photo. There are about 25 species of seahorses, distributed well, but the majority occurring in the Indo-Australian area. Seahorses are exclusively marine in habitat and are quite small, ranging from a length of about $1\frac{1}{2}$ inches (3.8 centimeters) to a maximum of 8 inches (20 centimeters). There are numerous similarities between seahorses and pipefishes. For example, as with pipefishes, there is independent movement of the eyes. Seahorses are incapable of rapid movement and hence utilize an accurate suction mechanism which is operable within a range of about $1\frac{1}{2}$ inches (3.8 centimeters). It has been observed that juvenile seahorses may feed for a period of up to 10 hours and, during that time, consume from 3,000 to 4,000 brine shrimp. The large species (*Hippocampus hudsonius* or Atlantic seahorse) survives well in tanks but does not reproduce in captivity. In contrast, the *Hippocampus zosterae* (Gulf of Mexico) pigmy seahorse may survive through a number of generations. Seahorses are well known for the display of unusual courting habits. These include odd shaking movements, and in some species, the holding of tails, and, in the female, simply turning of the head stimulates the male. In the forward direction, the tail of the seahorse is used in a prehensile manner.


Seahorse.

SEA ICE. See **Polar Research.**

SEAL. See **Sea Lions and Seals.**

SEA LEMON (*Mollusca, Gasteropoda; Nudibranchiata*). A flattened oval marine mollusk without shell, which feeds on sponges and other sessile animals. The roughened skin and the oval form suggest the fruit for which the animal is named.

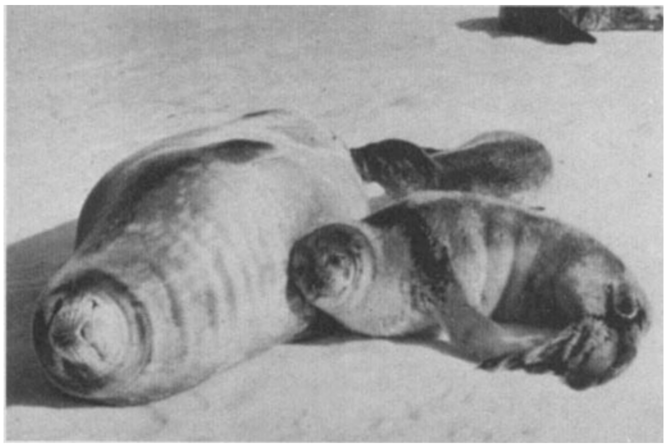

Sea lemon.

SEA LEOPARD. See **Sea Lions and Seals.**

SEA LEVEL. See **Hydrology; Ocean.**

SEA LIONS AND SEALS (*Mammalia, Pinnipedia*). This order of mammals contains large swimming animals, including sea lions, seals, walruses, and sea elephants, among others. At one time, some authorities placed these animals under the *Carnivora*.

Sea lions are animals closely related to the seals. They belong to the same family as the fur seals and, with them, differ from the other seals in having external ears. The common sea lion, *Zalophus caliifornianus*, is found on the Pacific coast of Mexico and California, and Steller's sea lion, *Eumetopias stellari*, ranges from the Bering Strait to southern California. See accompanying figure.


Weddel seal. (*A. M. Winchester.*)

Seals are animals whose bodies are highly specialized for life in the ocean, although they are able to move about on land rather clumsily. The fore limbs are paddlelike flippers, and the hind limbs are shifted so that they lie close together on opposite sides of the rudimentary tail and serve as a powerful propeller, increasing the effectiveness of the vertical undulating movements by which the animal swims. The body proper is formed to offer little resistance to the water in swimming. The fur seal represented on the Pacific coast by *Callorhinus alascanus* (*ursinus*) was once abundant in both hemispheres but was threatened with extermination by relentless hunting practices many years ago. The Pribilof Islands are an important center for these well known seals. Fur seals are sometimes called sea bears. Seals are relatively long-lived, the average bull seal living up to 20 years of age. The bulls do not mate until they are 10 years old. An adult bull may mate many females within a few days. The cows live even longer, up to 30 years. Some species of seals are found throughout the North Atlantic and in waters around the British Isles. Reports indicate that seals were hunted for their meat as early as 500 B.C.

Although mature seals live to a good age, studies in which pups have been tagged and traced indicate that about 60% of the pups die before they reach one year of age. The pups are quite small and have only a relatively thin skin for protection. The adult seals develop thick layers (up to 2 inches); 5 centimeters of blubber immediately under the outer skin. The females identify their young by sense of smell. The pups are nursed, seal milk being among the richest of all mammal milks. Large bulls may attain a length of some 18 feet (5 meters) and a weight of about 3 tons (2.7 metric tons). The elephant seal may have a trunk some 2 feet (0.6 meter) in length. Also known as the sea elephant, the name refers to the proboscis of the male. They are found principally along the California coast. In arctic regions, polar bears are among the seals' worst enemies. The bears may pounce upon a seal sleeping on the ice, attack it in the water while swimming, or await a seal to come out of its breathing hole. Alaskan fur seals have been studied extensively. At the start of a mating season, a bull will stake out a specific area, to which females gradually arrive. Competing bulls herd as many females as possible into an area which can be protected effectively. The bulls are ferocious fighters and jealously guard their territories. See illustration that accompanies entry on **Sexual Selection.**

The sea leopard is a large seal, *Stenorhynchus leptonyx*, of the Southern Hemisphere. It is yellowish or tawny with gray spots.

The walrus is a giant marine animal related to the seals, but constituting a distinct family. Adults reach a length of more than 12 feet and a weight in excess of a ton. The feet of the walrus are adapted for swimming, but they are used mainly for clumsy locomotion on land, as in the case of seals. In early life, the body is covered with thick lightbrown fur, that tends to disappear after middle age. The muzzle bears a number of very thick bristles. Walruses have the canine teeth of the upper jaw prolonged as tusks. The ivory of these tusks is valued by Eskimos. These animals are confined to the arctic seas and are commonly regarded as constituting an Atlantic, *Odobanus rosmarus*, and a Pacific, *O. obesus*, species, the latter with longer tusks.

SEA MOTHS (*Osteichthyes*). Of order the *Hypostomides,*, family *Pegasidae*, sea moths are well named because, except upon close examination, they appear possibly more like an insect than a fish. In fact, biologists and zoologists have not fully agreed upon the exact fit for this creature. Doubtless, they belong in the great domain of fishes, but the exact place remains to be scientifically determined. Sea moths have been a curiosity for centuries as they were brought by travelers into Europe from the Orient and South Pacific. Because of similarities of the body armor, there is some resemblance to pipefishes, seahorses, and sea poachers. Some relationship with the flying gurnards is indicated by the expanded pectoral fins. Some five species of sea moths are known, all occurring in tropical waters from Hawaii to Africa. No specimens have been encountered in the Atlantic. The largest specimens measure about 5 inches (12.5 centimeters) in length.

SEAMOUNTS. Generally, isolated peaks rising 3,000 or more feet (915 meters) above the floor of the ocean basin. Also known as *guyots*, they are believed to be volcanic cones that once rose above sea level. Because they usually have flat tops, it is believed that sea erosion flattened them, after which the earth's crust in that location subsided, lowering the seamouts far below the surface to their present depths. See also **Ocean.**

SEA MOUSE (*Annelida, Polychaeta*). A marine worm, *Aphrodite* (*A. hastata* in American waters and *A aculeata* in British waters), of compact oval form, covered above and on the sides with a felt-like material. It is recorded from Vineyard Sound on the Atlantic coast.

SEA OTTER. See **Mustelines.**

SEA RAVEN. See **Sculpins.**

SEA ROBINS (*Osteichthyes*). Of the suborder *Polynemoidea*, family *Triglidae*, sea robins occur in American Atlantic waters from Nova Scotia south to Venezuela. They attain a length of about 16 inches (40 centimeters), are highly colorful, with bony heads well protected with spines. Separate fanlike pectoral fins help the sea robin "walk" on the bottom. They are carnivorous, mollusks and small crustaceans comprising their main diet. Some of the larger members can attain a length of some 2 to 3 feet (0.6 to 0.9 meter). It is believed that all triglids are capable of noise-making and, as with other "acoustic" fish, such as croakers and grunts, their sounds are created by vibrating muscles with the assistance of a large air bladder which acts as a sounding box. The tubfish (*Trigla hirundo*), also known as the yellow gurnard, is found in the waters off Europe and the African coasts. It is a very striking fish, featuring bright blue-edged pectorals. There is also a group of armored sea robins (*Peristediidae*) whose entire bodies are protected by bony projections. They prefer the deep tropical waters. Some of the larger triglids are considered good as food.

SEA SLUG (*Mollusca, Gasteropoda*). Marine mollusks with compact bodies and without shells. Some species have branching processes on the surface of the body by which they breathe. The creeping habits and general form are similar to those of the terrestrial slugs, although there is no closer relationship between the two. The sea slugs make up the order *Nudibranchiata*.

SEA SNAKES. See **Snakes.**

SEASON CRACKING. Spontaneous cracking of brass and other metals on standing. Intergranular cracks result from the action of residual internal stresses from cold-working operations aided by surface corrosion. Cold-worked high-zinc brasses sometimes fail during storage in ordinary atmosphere. Ammonia salts and other specific reagents greatly accelerate cracking in brasses subject to this defect. Many other metals, including stainless steels, are subject to cracking under certain conditions of stress and corrosion.

SEAWATER (Desalination). See **Desalination.**

SEAWEEDS. The production of edible seaweeds is a significant industry in the Orient. It is particularly well established in Japan where seaweed "farming" is practiced on a large scale. China and Korea are also notable producers of edible seaweeds, and improved hybrids have been reported from China. Increased demand for seaweeds has focused attention on methods of cultivating selected species. In the late 1950s, the Japanese *Gelidum* beds, which had proved adequate for centuries, declined rather mysteriously and caused a world shortage. This resulted in the large-scale exploitation of supplies of suitable seaweeds for the manufacture of agar in Chile and Portugal. By the early 1960s, the demand for carrageenin, and to a lesser extent for furcellaran, by the food processing industry caused a shortage of other red seaweeds. The main producers of these products constructed drying units in Nova Scotia and Prince Edward Island to encourage the collection of seaweed.

The economically important seaweeds fall into two main groups: (1) that which grows in the inter-tidal fringe; and (2) that which is permanently submerged. This apparently simple distinction is, however, complicated by many factors. The extent of the inter-tidal zone is partly controlled by the gradient of the beach, and it also increases with latitude and tidal range. There are considerable local variations in tidal range, varying from less than 3 meters (10 feet) in Shetland (British Isles) to nearly 10 meters (30 feet) on parts of the west coast of the isles. One prominent seaweed of the inter-tidal zone (*Chondrus crispus*) extends into the sublittoral; and a substantial part of the Canadian harvest grows sublittorally. (Littoral refers to of, on, or along the shore.) The important free-floating seaweeds are *Furcellaria fastigiata*, harvested in the central Kattegat; and *Gracilaria* spp., found off the coasts of Chile and Portugal. These forms are usually harvested by netting.

The cast seaweeds are important in some areas, such as the cast red seaweeds of Prince Edward Island and the cast stripes (*Laminaria hyperborea*) of the Scottish coasts.

Apart from edible uses, seaweed is essentially the raw material for the extraction of a range of carbohydrates, such as agar, carrageenin, and furcelleran from the red seaweeds; and sodium alginate from brown seaweeds—all products used in food processing. Edible seaweeds produced in Japan include: *Konbu* (*Laminaria* spp.); *wakeme* (*Undaria pinnatifida*); *amanori* (*Porphyra* spp.); and *aonori* (*Monostroma* spp. and *Enteromorpha* spp.). The production of amanori approximates upwards of 7000 sheets in Japan; and about one-third of that quantity in Korea, much of which is exported to Japan. Converted into weight, this production represents about 300,000 tons (270,000 metric tons) annually.

SEBACEOUS CYST. A cystic structure developing in the skin, due to plugging of a duct leading from a sebaceous gland. The cyst may increase in size, and some become infected. Sebaceous cysts are commonly called wens and are often seen on the scalp and face as well as other parts of the body. They always contain a white cheesy material.

Treatment is by surgical excision. If the entire cyst wall is not removed recurrence is likely.

SEBACEOUS GLAND. A gland of the skin of mammals which secretes an oily substance (sebum). These glands are usually associated with hair follicles and in man are especially abundant in the scalp, although they occur all over the body with the exception of the palms of the hands and soles of the feet. Their secretion keeps the skin pliable and anoints the hair.

In structure these glands are of the compound alveolar type. Their secretory parts are saccular, discharging to a common duct which often opens in the hair follicle.

SEBORRHEA. A variety of disorders of sebaceous glands are included in this clinical term. The dry, scaling form, "dandruff," is characterized by the presence of fine, branny, slightly greasy scales, which are readily shed. It is common on the scalp but may spread down over the face, neck and ears, and may even be a generalized dermatitis. The oily variety is associated with an excess secretion of oil by the sebaceous glands of the scalp and skin. The scalp is covered with a greasy layer which may mat the hair together; this type is associated with permanent baldness.

In some cases, psoriasis of the scalp may be involved. This disorder usually forms discrete plaques covered by silvery scales, whereas seborheic dermatitis characteristically is diffuse and patchy and is associated with a greasy scale.

See also **Dermatitis and Dermatosis.**

SECONDARY EMISSION. This term refers to the result of any of several different processes, in each of which some kind of "primary" emission, when it encounters some form of matter, gives rise to another emission of the same or of different character.

The most familar example of a secondary emission is the x-rays, which have their origin in the impacts of high-speed electrons (cathode rays) upon atoms of matter. The resulting x-rays may themselves act, in turn, as the primary emission and, falling upon solid bodies, cause a secondary x-ray emission. Or they may fall upon a fluorescent substance (see **Luminescence**) and give rise to a secondary radiation of visible light. X-rays, or any other photons, falling upon a photosensitive metal, may cause a secondary emission of photoelectrons. (See **Photoelectric Effect.**) The "recoil" electrons from the Compton scattering of x-rays constitute one form of secondary emission. (See **Compton Effect.**)

The most common use of the term denotes the emission of electrons from a solid as the result of the collision of higher energy electrons with the solid. Since the energy levels of the emitted electrons depend upon the atoms emitting them, as well as upon the initial energy, a method of surface layer analysis (i.e., analysis of atoms at the surface or several atomic layers below) has been based upon secondary electron emission. The method consists of placing the solid sample in a vacuum chamber and bombarding it with a beam of relatively low-energy electrons. The secondary electrons emitted (see **Auger Effect**) are recorded according to their characteristic energies, which correspond to the atoms emitting them. The method is especially useful in detecting impurities of the lighter elements.

SECONDARY SEXUAL CHARACTERISTICS. The characteristics of living things whose appearance is definitely associated with the sex of the individual, although the characteristics have no direct connection with the process of reproduction.

The different colors and patterns of the two sexes in many species of birds and insects are familiar examples of secondary sexual characteristics. Horns of some males, the manes of many male mammals, and the spurs of cocks are also in this category. In some species the differences resulting from such characteristics are so great that the two sexes can scarcely be associated by appearance. Cases are on record among the insects of the classification of males and females of the same species in different genera prior to the discovery of their relationship through other evidence. See also **Gonads;** and **Hormones.**

SECOND ORDER REACTION. See **Chemical Reaction Rate.**

SECOND ORDER SYSTEM. A system whose performance characteristics are presented in the form of a second order differential equation

$$r(t) = k_1 \frac{d^2C(t)}{dt^2} + k_2 \frac{dC(t)}{dt} + k_3C(t) \cdots$$

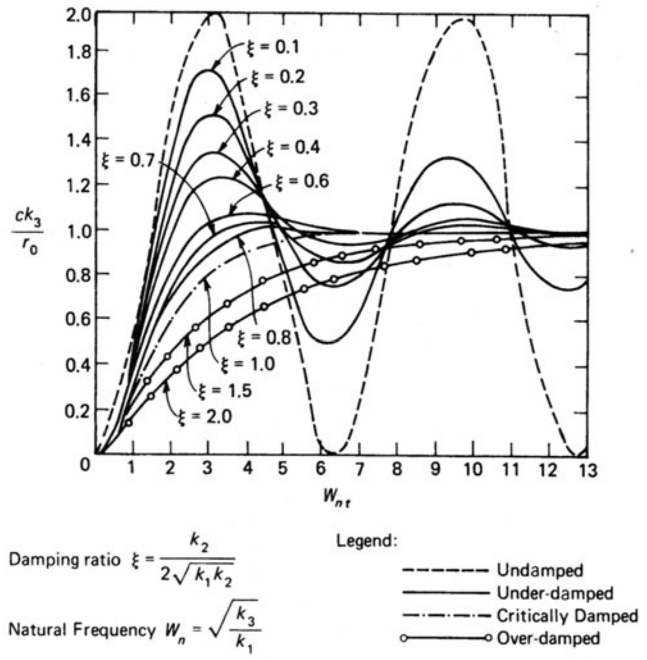
where r = system input
C = system output
t = time

k_1, k_2 and k_3 are coefficients.

If all coefficients are constants, then the Laplace transform of foregoing equation also should be second order:

$$\frac{C(S)}{R(S)} = \frac{1}{k_1S^2 + k_2S + k_3} \cdots$$

For simplicity, all initial conditions were assumed zero. The time response of a second order system subjected to a step input, r_0, is shown graphically in the accompanying figure.



Time response curve of a second order system subjected to a step input.

SECRETARY BIRD. See **Eagle.**

SECRETION. See **Kidney and Urinary Tract; Urine.**

SECTILE. Capable of being cut. In mineralogy, sectile refers to substances, such as talc, mica, and steatite, which can be cut smoothly by a knife.

SECTION MODULUS. An inspection of flexure will reveal that the maximum stress in a member subjected to a transverse bending is directly proportional to the external bending moment, and inversely proportional to the ratio of

$$\frac{\text{moment of inertia}}{\text{distance of the farthest stressed element from the nuetral axis}}$$

It is apparent that this ratio is entirely a property of the shape and size of the cross section of the structural member. This ratio is known as the section modulus, and is an important property of rolled steel sections and other shapes which are used as structural members. When the bending moment to be withstood by a beam or column is divided by this section modulus, the quotient is the maximum bending stress which will exist in that member.

SECULAR DETERMINANT. An equation of the form

$$K(\lambda) = |a_{ij} - b_{ij}\lambda| = 0$$

which becomes a polynomial in λ when the determinant is expanded. In matrix algebra it is the determinant of the characteristic matrix of A, $[\lambda E - A]$ where E is a unit matrix. In this case,

$$K(\lambda) = \lambda^n + a_1\lambda^{n-1} + \cdots + a_n = 0$$

is the characteristic equation or function, the coefficients a_i involve the elements of A, and the n roots of $K(\lambda)$ are the eigenvalues, latent or characteristic roots. All of them need not be different; if two or more of them are equal they are said to be degenerate. An important property of the latent roots is the following. Suppose two matrices A and B are related by a similarity transformation, $B = Q^{-1}AQ$, then the latent roots of A are identical with those of B.

See also **Determinant.**

SECULAR PARALLAX. The parallactic motion of a star due to the sun's motion through space (see **Stellar Parallax**).

SECULAR TERMS. In the mathematical expression of an orbit, terms for very long period perturbations, in contrast to periodic terms, terms of short period.

SEDGE (*Cyperaceae*). This family of monocotyledons is composed of grass-like plants which are found chiefly in marshy places. Most of its members are perennials with creeping rhizomes and grass-like leaves. The basal portion of the leaf is a sheath which completely surrounds the stem. The stem is generally solid and triangular in cross section. The inflorescence is a spike or a panicle, composed of one- to many-flowered spikelets. Each flower, borne in the axil of a bract, has three stamens and a single pistil; in some sedges there is also a perianth of six or many bristles. The fruit is an achene. All sedges are wind-pollinated.

Few of the sedges are of any importance. Some of them yield a coarse hay which may be fed to livestock, and is sometimes used for packing material. Papyrus, used as paper by the ancient peoples, was made from the stems of *Cyperus papyrus*. The erect stems and leaves of species of *Scirpus* are sometimes dried and woven to form chair seats; chairs finished with this material are called rush-bottomed chairs. The plants are sometimes called bulrushes. Several sedges form tubers which are sometimes used for food.

SEEBECK EFFECT. Production of electromotive force as a result of a circuit containing two conducting materials having two junctions between the materials at different temperatures. This effect is the basis of all thermocouples. The Seebeck effect is one of a number of related thermoelectric phenomena, is the inverse of the Peltier effect, and is closely related to the Thomson effect. See also **Thermocouple.**

SEED. In the vegetable kingdom, seeds are a vital link to the past and to the future. For the majority of food plants, of all factors, including climate, water management, fertilizers, insecticides, etc., that are required to ensure a successful crop, healthy, vigorous seeds are, without qualification, the singular indispensable requirement. Seed requirements for planting are a significant cost factor and support a major seed growing, processing, and supply industry, as is evidenced by the accompanying table.

Seeds per se are consumed as food items and the oils and proteins which they contain are expressed, extracted, and separated from them to yield edible fats and oils and vegetable protein. The grain industry is based upon seed processing. See also **Protein;** and **Vegetable Oils (Edible).**

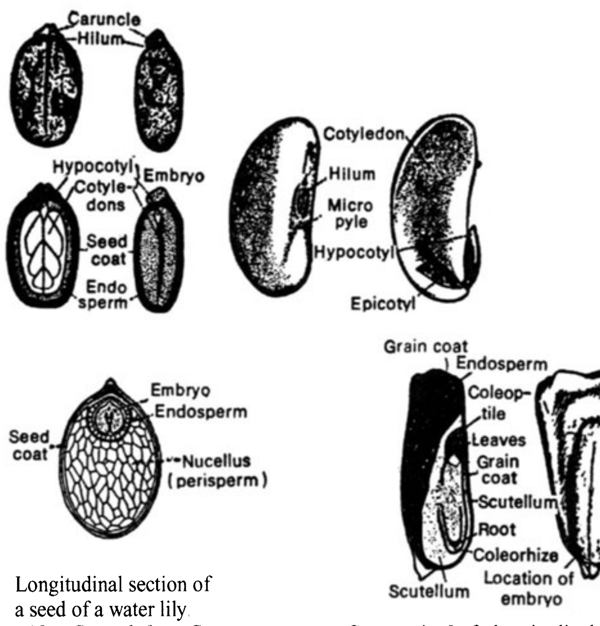
QUANTITY OF SEED REQUIRED FOR PLANTING SELECTED MAJOR CROPS[1]

Crop	Pounds/Acre	Kilograms/Hectare
Barley	80	90
Beans (dry edible)	54	60
Groundnuts (peanuts)	136	152
Maize (corn)	13.4	15
Oats	83	93
Peas (dry field)	167	187
Potatoes (Irish)	1830	1050
Rice	142	159
Rye	90	101
Sorghum	6.76	7.57
Soybeans	63	71
Sweet potatoes	600	672
Wheat (Durum)	89	100
Wheat (winter)	69	77

[1]Average of entire United States plantings.
SOURCE: U.S. Department of Agriculture.

The collection and preservation of seeds as a source of germ plasm for improving crops in numerous ways (yields, resistance to disease, insects, and severe climates) are extremely important scientific undertakings.

Botanical Factors A seed consists of a dormant embryo, together with a quantity of stored food which may be absorbed in the embryo, or may surround it, and one or two seed coats or integuments. The seed develops from an ovule. In angiosperms, the seed is completely enclosed by the ovary wall. In gymnosperms, the seed lies exposed on the surface of a scale of the cone. Representative seeds are shown in the accompanying diagram.



Longitudinal section of a seed of a water lily (*After Conrad, from Curtis, Nature and Development of Plants, Henry Holt & Co.*)

Corn grain. Left, longitudinal section perpendicular to the broad face of the grain; right, surface view

Representative types of seeds.

The fertilized egg develops into the embryo which is a young plant contained in a seed. This embryo may be an undifferentiated mass of cells, as it is in the orchid family, but usually is more highly organized. It then consists of a short axis which is called the hypocotyl. At one end of the hypocotyl there is a primitive root called the radicle. At the other end is a terminal bud, called the plumule. This plumule may be nothing more than a small mass of undifferentiated cells, recognizable only as a small bulge at the apex of the hypocotyl, or it may be a well-developed shoot having a short internode and two distinct leaves. Borne laterally at the apex of the hypocotyl there are one or more seed leaves or cotyledons. In many seeds these cotyledons are thin and more or less leaf-like, while in others they are very fleshy, filled with stored food material, and form the greater part of the seed. The number of cotyledons varies. In monocotyledons there is usually only one cotyledon, in dicotyledons there are two, and in gymnosperms there are often many.

In angiosperms the endosperm is the tissue which results from the endosperm nucleus. It is a tissue which is rich in stored food. The food reserves stored in the seed are carbohydrates, especially starches and sugars, fats and proteins. The developing embryo gets its food from the endosperm. In many seeds the embryo uses only a part of the endosperm during its development, so that the mature seed contains much endosperm surrounding the embryo. These are called albuminous seeds. In other plants, the food reserves of the endosperm have been entirely absorbed and restored in the embryo, especially in the cotyledons, which then becomes very fleshy. Such seeds are exalbuminous.

The embryo sac of the ovule is surrounded by a mass of tissue called the nucellus. In most plants this is completely absorbed before the seed reaches maturity. In some seeds it persists and becomes much enlarged. It is then known as the perisperm, and serves as an additional source of stored food.

Surrounding all these are the seed coats, which develop from the integuments of the ovule. The outer coat may be variously modified to aid in the dissemination of the seeds. On the seeds of many plants there is a fleshy structure called the aril which grows up around and more or less covers the outer integument. In some seeds, the outer coat produces an outgrowth called the caruncle, which seems to aid in absorbing water from the soil and passing it on to the seed. Passing through the integuments is a minute hole called the micropyle. It is through this that the pollen tube entered the young ovule; the radicle generally points directly towards it. The seed is attached to the ovary wall by a small stalk or funiculus which, when the seed falls off, leaves a scar called the hilum on the seed coats.

Natural Broadcasting of Seeds. Once started, without human intervention the plant must remain in a fixed position throughout its life. In higher plants, the vegetative parts are very rarely able to colonize new territories. The fruit, or less frequently the seed, is the part which is carried to new regions. There are several agents effecting this transfer. The most important is wind. Sometimes the seeds of a plant are very small and light, and so easily carried by the wind. The minute seeds of orchids are carried by this means; in these seeds, additional buoyancy is gained by a loose, thin case which surrounds the embryo tissue within and acts as a float. In other plants the seeds are carried away by hairs which grow from the seeds. Milkweed seeds are provided with a tuft of long silky hairs attached at one end of the thin light seed. Cotton fibers serve the same purpose; they completely cover the cotton seed. In other plants the seed is provided with a wing, a flat thin outgrowth from the seed coats. Catalpa seeds are thus equipped, as also are pine seeds. The distances to which the wind carries seeds is considerable. By this means plants are disseminated over many square miles or kilometers.

Another way in which new land is reached is by ejecting seeds violently from the fruit. When the fruit of the witch hazel is ripe, the dry, thick wall of the ovary suddenly snaps and hurls the seeds violently to a distance of many feet (several meters). The common Jewelweed or Touch-me-not scatters its seeds in similar fashion. At maturity the fruit abruptly splits open, and the valves roll back, throwing the seeds many feet away. In similar fashion the pods of many legumes split apart forcefully and scatter the seeds within. Seeds scattered by this means cannot attain the wide dispersal which wind-borne seeds do.

A few plants form seeds which float readily on water for some time without harm. Currents of water may carry the seeds to new shores. Other seeds are provided with hooks or barbs or have a sticky surface which causes them to adhere to the bodies of passing animals which scatter them. In most plants, however, it is the fruit which is so carried. Such fruits, as beggar's lice, burdock, goldenrod and many others, are commonly mistaken for seeds.

Germination and Dormancy. Having reached an environment where suitable conditions exist, the seed germinates. The seeds of many plants must reach such a place in a very short time or perish, since they remain viable but a very short time. Those of the willow, for example, live only a few days after falling from the parent plant. The seeds of most garden vegetables grow best if planted within a year from the time they ripen, though they may retain their vitality for three or four years with decreasing vigor. On the other hand, some seeds lie dormant for a long time before germinating. Seeds of many weeds, including the common ragweed, the pollen of which causes hay-fever, may live for years before germinating, making it very difficult to eradicate the species by pulling up the plants for a single season. Tests show that the seeds of some plants may remain viable for 20–50 years. It is recorded that the seeds of the Asiatic lotus have germinated after lying dormant for 200 years. But records of viable seeds found in ancient vaults, such as contained the mummies of Egypt, are entirely unfounded.

Certain conditions favor the continued vitality of dormant seeds. Sometimes the seed has a very thick wall which is impervious to water or to oxygen gas, and so excludes two things necessary to start germination. Until the wall has softened or rotted the seed does not germinate. In other seeds the thick wall resists the pressure of the developing embryo within. Many important crop plants have seeds which germinate slowly because of their thick coats. To hasten germination and to insure a uniform stand of plants the seeds are scarified before planting; that is, the seed coat is rubbed with abrasive substances which break down the impervious wall layers. Often the wall of the seed is sufficiently damaged during mechanical threshing to insure prompt germination when planted. Prolonged soaking sometimes hastens germination.

In other seeds dormancy is inherent in the embryo itself. The embryo may be entirely undeveloped, requiring a long period of slow development before it can break the seed coats. Other seeds germinate only after a period of "after-ripening" which varies from a single winter to many years. The changes occurring during this period are as yet not well understood. The time of "after-ripening" may be considerably shortened by burying the seeds in sand or other suitable material and keeping them cold and moist.

The external conditions necessary to cause germination are adequate water, suitable temperature and oxygen. With some seeds light is an important factor. Most seeds contain very little water, which is one of the reasons why they can survive under adverse conditions such as cold and drought. To germinate they must receive additional water. This added water favors digestion, a process which makes available to the plant the stored food. Both water and oxygen are needed by the germinating embryo, because of the great increase in respiration, the process which frees to the plant the energy stored in the carbohydrates and other compounds.

The temperature at which a seed will germinate varies with different plants. For each there is a considerable range of temperature. The lowest temperature at which germination occurs is called the minimum temperature, and varies from 0–10°C or even higher. The maximum or highest temperature at which germination takes place is usually between 45 and 50°C. The most favorable temperature, or optimum, is about 30°C. Light favors the germination of many common plants, such as many grasses and troublesome weeds. Other plants, including many common crop plants, are unaffected by light.

Germination is the development of the embryo into a young plant. It becomes completed when the young plant is independent of the food stored in the seed. In most seeds the first visible change is swelling of the seed, which is a result of the increased water content. Often the seed coats are ruptured by the swelling of the contents of the seed. Respiration increases greatly, and much energy is made available in those regions where active growth occurs, that is, the hypocotyl and plumule. The radicle pushes out of the seed and attaches itself, by means of root hairs, to the soil particles. These then begin absorbing water from the soil. In some seeds the hypocotyl elongates considerably, often forming an arch which subsequently straightens, lifting the cotyledons and plumule out of the soil and into the air. In other seeds the cotyledons remain permanently underground, the plumule elongating and pushing out into the air. There the first leaves of the plant appear. With their formation the plant becomes independent.

Many foodstuffs are seeds, especially those of the cereal grains, rice, wheat, corn, barley, and oats. But seeds are used in many other ways. Many medicinal products are obtained from seeds. Linseed oil, soybean oil, and coconut oil are but a few of the many oils which come from seeds. Poppy seeds, caraway seeds, and mustard add flavor to other foods. Clothing is made from the hairy covering of the cotton seed. Beads, buttons, and ornaments of various kinds are also often made from seeds.

See also **Germ Plasm;** and **Plant Breeding.**

SEEDING (Gas Conductivity). Introduction of atoms, such as sodium, with a low ionization potential into a hot gas for the purpose of increasing the electrical conductivity.

SEEING (Astronomy). A term used by ground-based optical astronomers to describe image quality.

When describing stellar (point source) images, good seeing means the images are small and stationary; bad seeing means the images are large (sometimes changing in size in a rapid and irregular fashion) and unsteady. When describing images of extended objects (such as planets or nebulae), good seeing means the images are sharp and steady; bad seeing means the images are blurred and unsteady.

Seeing quality is determined by motions of air masses of differing index of refraction (due to differences in density and/or humidity) so that light rays are refracted or bent in a time-variable fashion. These motions can be both near the telescope or at some distance from it. A similar phenomenon is observed in viewing a distant object over a nearby fire or other source of heat. The heat causes rising currents of warmer air which produce time variable refraction so that the object's image is seen to "shimmer" or be blurred.

For most astronomical observations, the better the seeing, the fainter one can observe. Hence, seeing quality is an important criterion in selecting the site of a ground-based observatory. (Other important criteria are the number of clear hours, absence of anthropogenic night-time illumination, and, for infrared observations, dryness of the atmosphere.) The sites with the best seeing are generally islands and mountains near oceans or in isolated flat areas. This is generally believed to be because the large-scale motions of the atmosphere in such places are largely laminar rather than turbulent. Local conditions, such as local topography, height of dome, temperature of dome floor, among other factors, can be very important and much effort is made to control these conditions.

See also **Light Pollution.**

Peter Pesch, Chairman, Astronomy Department,
Case Western Reserve University, Cleveland, Ohio.

SEGER CONE. A series of substances having different fusion temperatures might serve roughly to measure the temperature of high-temperature regions such as furnaces, since, with a series of substances having progressively increasing fusing temperatures, the temperature naturally lies between the fusion temperature of the last substance fused, and that of the next not yet fused. A series of artificially prepared mixtures, mostly of the oxides such as clays, lime, feldspar, have been designed to form a series of "Seger cones." There are 60 mixtures covering a temperature range from 590 to 2,000°C.

SEICHE. 1. A standing wave oscillation of an enclosed water body that continues, pendulum fashion, after the cessation of the orginating force, which may have been either seismic or atmospheric.

2. An oscillation of a fluid body in response to a disturbing force having the same frequency as the natural frequency of the fluid system. Tides are now considered to be seiches induced primarily by the periodic forces caused by the sun and moon.

3. In the Great Lakes area, any sudden rise in the water of a harbor or a lake, whether or not it is oscillatory. Although inaccurate in a strict sense, this usage is well established in the Great Lakes area. See also **Estuary.**

SEISMIC SEA WAVE. See **Tsunami.**

SEISMOLOGY. See **Earth Tectonics and Earthquakes.**

SEIZURE (Neurological). Epilepsy is characterized by sudden, brief disturbances in brain function. A clinical seizure may be defined as one resulting from the excessive discharge of aggregates of neurons, which in some manner become depolarized in a synchronous fashion. Epilepsy is not a disease in the usual sense, but rather it is a disorder, the root causes of which remain poorly understood. Some researchers have suggested that the paroxysmal depolarization of neurons may derive from abnormalities of neuronal membranes, disturbances in synaptic transmission, or defective functioning of glial cells. The causative factors will probably be revealed when there is a better understanding of the fundamental biochemistry and bioelectronics of the brain and the nervous system. Epilepsy may be divided into two broad categories: (1) *idiopathic epilepsy*, where there is no known organic injury to the brain prior to the first seizure; and (2) *symptomatic epilepsy*, where such damage has been confirmed. The genetics of seizure disorders (epilepsy) are poorly understood and difficult to research because of the probable large number of causative factors and, from a statistical standpoint, because of the prevalence of the disorder. It is estimated that a minimum of 0.5% of the population suffer from recurrent seizures. This, in a country of the size of the United States, translates into 1 million persons. It is generally concluded that the close relatives of persons with idiopathic epilepsy may experience a seizure incidence rate of three times that of the general population. Of course, where seizures are the result of injury (accidents to the brain and spine during and after birth—at any time of life), genetic factors do not enter.

Even though there is a poor understanding of the biochemistry and mechanics which precipitate seizures, over the years very effective treatment regimens have been developed empirically.

Types of Seizures

Some authorities have classified seizures on the basis of symptoms as follows:

- Generalized Seizures
  - Grand Mal (*tonic-clonic*)
  - Petit Mal
  - Drop Attack (*akinetic*)
  - Myoclonic
- Partial Seizures
  - Psychomotor
  - Focal Cortical
- Continual (continuing *status epilepticus*)

The treatment, including the drugs of choice, varies from one type to the next.

Grand Mal Epilepsy. In this type of seizure, the patient may feel unusually good or poor for a day or two prior to the attack. This vague state warns the patient of an impending attack. It is peculiar to the individual and identical in different attacks. There may be queer sensations in some part of the body. Flashes of light or color may be seen, strange sounds may be heard, and pleasant or disturbing emotions may be experienced. Consciousness is swiftly lost, often with a wild, harsh cry. Breathing stops; the legs and body stiffen; and the patient falls to the ground, the elbows bent at rigid right angles. During this part of the spasm, which lasts from 10 to 30 seconds, the bladder or bowel, or both, may empty. The face becomes blue as the features contort, and the patient seems about to die when the spasm breaks. Rhythmic muscular contractions, at first small and rapid, begin and then become slower and more powerful. Gasps for breath come through heavy froth, often blood-stained from a bitten tongue or cheek. The contractions become less and less frequent. This phase usually lasts 2 to 3 minutes, but may persist longer. At the end, the patient may sleep heavily for several hours or rouse with aimless, thrashing movements, dazed and forgetful and unable to understand what is said to him. There is often a severe headache for several hours. Occasionally, the patient may have one convulsion after another without regaining consciousness. The dangerous state is called *status epilepticus* and demands immediate attention by the physician, since it can result in death from exhaustion.

In summary, the clinical features of grand mal are generalized major motor convulsions with a loss of consciousness and a depression of cerebral function.

Petit Mal. In petit mal attacks, a milder form of epilepsy, the loss of consciousness is fleeting and variable—from 1 to 40 to 50 times per day. It may be so short that it goes unnoticed. The head may nod momentarily, the flow of speech may halt a second or two and then may be normally resumed; or perhaps only a vacant stare marks the attack. Sometimes there are one or two contractions of the arms flickering of the eyelids. In petit mal, an electroencephalogram a spike-and-wave pattern, which occurs at the rate of 3 per second.

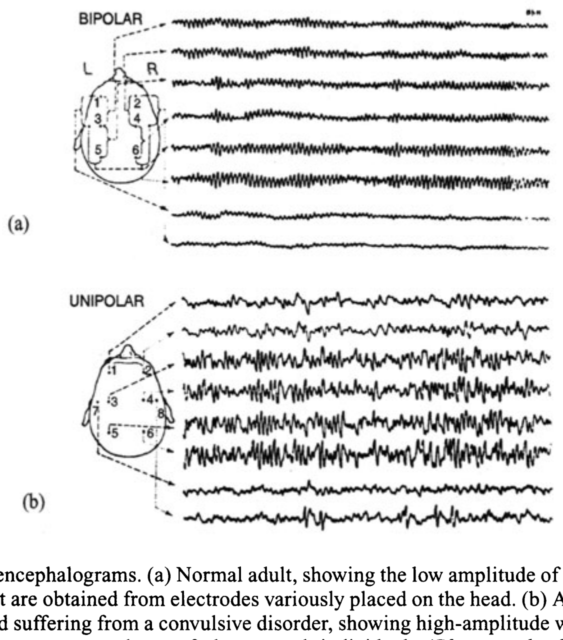
Drop Attack. In this type of attack, there is a brief loss of consciousness and loss of muscle tone.

Myoclonic Type. The clinical features include isolated muscle jerks, which may be precipitated by various sensory stimuli. This condition is frequently associated with degenerative and metabolic brain disease.

Psychomotor Epilepsy. This manifestation may follow grand or petit mal seizures or occur independently. These attacks last from a few minutes to a day or two. The patient may remember nothing of the episode on recovery. The behavior is confused and unusual. There may be uncontrollable emotional outbursts which may be violently destructive or the patient may be dazed and apathetic. There is also evidence that some psychomotor cases may be the result of subtle injury to the brain at the time of birth. In this state, there is some impairment of consciousness.

Focal Seizures. Also known as Jacksonian epilepsy, these seizures assume two forms. There may be unusual sensations or uncontrolled movements remaining localized to one part of the body, while consciousness is undisturbed. Thus, the head and eyes may turn irresistibly to one side despite the patient's awareness. In the other form, the movements or sensations which begin in one part of the body may spread upward in a slow, orderly fashion, or they may cross to the other side. This form may develop into a typical grand mal attack with loss of consciousness. Many convulsive seizures resulting from brain injury are of the Jacksonian type.

Continual Seizures. There is incomplete recovery between attacks, which may be generalized or partial. Drugs of choice in treatment include phenytoin and phenobarbital. Patients are treated for *status epilepticus*. This condition is considered a medical emergency and invariably results in death when not treated. Factors which the physician and other emergency personnel must consider include: (1) measures for ensuring an unblocked airway and measures for ensuring adequate fluids, such as intravenous infusion with 5% dextrose; and (2) administration of drug of choice.



Electroencephalograms. (a) Normal adult, showing the low amplitude of the tracings that are obtained from electrodes variously placed on the head. (b) A 4-year-old child suffering from a convulsive disorder, showing high-amplitude waves, in marked contrast to those of the normal individual. (*Photography by F. W. Schmidth.*)

Electroencephalographic tracings can be of considerable value in the assessment of various seizure disorders. See accompanying diagram. See also **Nervous System and the Brain.**

SELACHII. See **Sharks.**

SELECTION RULES (Energy Levels). It was found early in the study of atomic spectra that radiative transitions between certain pairs of energy levels seldom or never occur. A set of rules which are expressed in terms of the differences of the quantum numbers of the two states involved allow a prediction of allowed transitions and forbidden transitions. The conditions for allowed transitions are:

$$\begin{aligned}\Delta L \text{ (orbital angular momentum)} &= \pm 1\\ \Delta J \text{ (total angular momentum)} &= 0 \text{ or } \pm 1\\ \Delta M \text{ (magnetic orientation)} &= 0 \text{ or } \pm 1\end{aligned}$$

The selection rules are not rigorously obeyed. In atoms which do not exhibit Russell-Saunders coupling, the quantum numbers L and S are not defined. Even in atoms which do have this type of coupling, forbidden transitions are merely of lower probability than allowed ones and may occur from a state from which no transitions are allowed by the rules if conditions are such that collisions of the second kind do not remove the atom from the initial state before it radiates (e.g., at extremely low pressures).

Similar selection rules hold for molecular spectra. In fact, let ψ_i and ψ_j be wave functions for two levels in any quantum mechanical system. Then if P is the appropriate operator, a transition between levels i and j is permitted if the matrix element

$$\int \psi_i^* P \psi_j \, d\tau$$

does not vanish. Here ψ_i^* is the complex conjugate of ψ_i, $d\tau$ is a volume element including all of the variables involved in the two wave functions, and the operator may refer to electric or magnetic dipole radiation, quadrupole radiation, polarizability, etc. If the integral vanishes, the transition is forbidden. Frequently, symmetry properties and group theory may be used to determine whether the matrix element does or does not vanish. This is very helpful since evaluation of the integral itself may be difficult or impossible.

SELECTION RULES (Nuclear). A set of statements that serve to classify transitions of a given type (emission or absorption of radiation, beta decay, and so forth) in terms of the spin and parity (I and π) quantum numbers of the initial and final states of the systems involved in the transitions, in such a way that transitions of a given order of inherent probability (after making allowance for the influence of varying energy, charge and size of system, and so forth) are grouped together. The group having highest probability of taking place per unit time is said to consist of allowed transitions; all others are called forbidden transitions. Table 1 lists the selection rules for radiative transitions: each entry gives the character (E = electric, M = magnetic) and the multipole order (1 for dipole, 2 for quadrupole, 3 for octopole, . . .) of the predominant radiation mechanism for the indicated spin change ΔI and parity change $\Delta\pi$; the entry "none" means that radiative transitions are strictly forbidden.

Fermi selection rules and Gamow-Teller (GT) selection rules are alternative sets of rules for allowed beta transitions; both are currently believed to be valid, so that a transition allowed according to either set is actually allowed.

TABLE 1. SELECTIVE RULES FOR RADIATIVE TRANSITIONS

$\Delta\pi$	ΔI						
	0 $I = 0$	0 $I \neq 0$	1	2	3	4	5
No	None	$M1$	$M1$	$E2$	$M3$	$E4$	$M5$
Yes	None	$E1$	$E1$	$M2$	$E3$	$M4$	$E5$

TABLE 2. SELECTION RULES FOR BETA DECAY

$\Delta\pi$	ΔI						
	0	1	2	3	4	5	6
No	*A*	*A*	II	II	IV	IV	VI
Yes	I	I	I	III	III	V	V

Table 2 lists the selection rules for beta decay: the entry *A* means that for the indicated spin and parity change the transition is allowed; I, means that it is first forbidden; II, second forbidden . . .

SELENIUM. Chemical element, symbol Se, at no. 34, at. wt. 78.96, periodic table group 16, mp 217°C, bp 685°C, density 4.82 g/cm^3 (solid), 4.86 (single crystal). Selenium has a large number of allotropes, some of which have not been fully investigated. On heating selenium above its melting point and cooling it, a red vitreous mass is formed, probably a mixture of allotropes. A red, amorphous allotrope is precipitated by SO_2 from selenious acid solutions. On heating at above 150°C, the red vitreous form changes to a gray hexagonal form, the stable form at ordinary temperatures, with metallic properties, one of which is photo-conductivity. By evaporation of a CS_2 solution of the red vitreous form below 72°C, a red α-monoclinic form is obtained; evaporation above 72°C gives β-monoclinic selenium. Black hexagonal selenium, believed to have a ring structure, is produced by heating amorphous selenium to near its melting point. Unlike sulfur, liquid selenium apparently has only one form.

There are six natural occurring isotopes: ^{74}Se, ^{76}Se through ^{78}Se, ^{80}Se, and ^{82}Se, and seven known radioactive isotopes ^{72}Se, ^{73}Se, ^{75}Se, ^{79}Se, ^{81}Se, ^{83}Se, and ^{84}Se. With exception of ^{79}Se which has a half-life of something less than 6×10^4 years, the half-lives of the other isotopes are comparatively short, measured in minutes, hours, or days. In terms of abundance, selenium ranks 34th among the elements occurring in the earth's crust. It is estimated that a cubic mile of seawater contains about 14 tons (3 metric tons per cubic kilometer) of selenium. First ionization potential 9.75 eV; second 21.3 eV; third, 33.9 eV; fourth, 42.72 eV; fifth, 72.8 eV. Oxidation potentials $H_2Se(aq) \rightarrow Se + 2H^+ + 2e^-$, 0.36 V; $Se + 3H_2O \rightarrow H_2SeO_3 + 4H^+ + 4e^-$, −0.740 V; $H_2SeO_3 + H_2O \rightarrow SeO_4^{2-} + 4H^+ + 2e^-$, −1.15 V; $Se^{2-} \rightarrow Se + 2e^-$, 0.78 V; $Se + 6OH^- \rightarrow SeO_3^{2-} + 3H_2O + 4e^-$, 0.36 V; $SeO_3^{2-} + 2OH^- \rightarrow SeO_4^{2-} + H_2O + 2e^-$, −0.03 V. Other important physical properties of selenium are given under **Chemical Elements.**

Selenium was first identified by Berzelius in 1817. The element is found associated with volcanic activity, as for example in cavities of Vesuvian lavas and in the volcanic tuff of Wyoming (about 150 parts per million).

Selenium occurs as selenide in many sulfide ores, especially those of copper, silver, lead, and iron, and is obtained as a by-product from the anode mud of copper refineries. The mud is (1) fused with sodium nitrate and silica, or (2) oxidized with HNO_3, and the H_2O extract is then treated with HCl and SO_2, whereupon free selenium is separated.

Uses: Selenium is widely used in photoelectric cells. The element alters its electrical resistance upon exposure to light. The response is proportional to the square root of incident energy. Selenium cells are most sensitive in the red portion of the spectrum. Although an external emf must be applied, the resistance is low and amplification is easy. In the selenium photovoltaic cell configuration, a thin film of vitreous or metallic selenium is coated onto a metal surface. Then, a transparent film of another metal, often platinum, is placed over the selenium. A cell of this type generates its own emf, with a decrease in internal resistance with increasing irradiation. The response essentially is proportional to incident energy. The cells are not importantly sensitive to small temperature changes.

Advantage of the unipolar conduction characteristic of selenium is taken in arc rectifiers. In a typical unit, a nickel or nickel-plated steel or an aluminum disk with a thin layer of selenium applied to one side is used. Selenium also is added to copper alloys and to stainless steel to increase machinability. Advantages claimed for selenium copper are high machinability, combined with hot-working properties and high electrical conductivity. As a decolorizer in glass, selenium counteracts green shades arising from ferrous ingredients. Sodium selenite is used in the production of red enamels and in the manufacture of clear red glass. Addition of from 1 to 3% selenium to vulcanized rubber increases abrasion resistance. The element also is used in photographic and printing reproduction chemicals.

Selenium is also used as an additive to lead-antimony battery grid metal and as a vulcanizing agent to improve temperature and abrasion resistance of rubber.

Chemistry and Compounds: Due to its $4s^24p^4$ electron configuration, selenium, like sulfur, forms many divalent compounds with two covalent bonds and two lone pairs, and *d* hybridization is quite common, to form compounds with Se oxidation states of 4+ and 6+.

While selenium dioxide, SeO_2, can be produced by direct reaction of the element with oxygen activated by passage through HNO_3, the compound is easily made by heating selenious acid, H_2SeO_3. Selenium dioxide sublimes at 315–317°C, and is readily reduced by SO_2 to elemental selenium. Selenium troxide, SeO_3, is not prepared from the dioxide by oxidation, although selenium does react with oxygen to form SeO_3 and SeO_2 in an electric discharge. Preferred method of preparing SeO_3 is by refluxing potassium selenate with sulfur trioxide. The reverse reaction, hydration of SeO_3 to selenic acid, H_2SeO_4, occurs easily. Selenious acid, H_2SeO_3, produced by hydration of SeO_2, is a stronger oxidizing agent than sulfurous acid as judged by its quantitative oxidation of iodide ion in acid solution, but is a weaker acid (ionization constants 2.4×10^{-3} and 4.8×10^{-9} at 25°C). It forms salts, the selenites, many of which, especially those of the heavy metals, are reduced to selenides by hydrazine. Many of the selenites, e.g., those of nickel, mercury, and ferric ion, are very slightly soluble in H_2O. Selenious acid is readily oxidized by halogens in the presence of silver ion or 30% H_2O_2 to selenic acid, H_2SeO_4. Selenic acid is as strong an acid as H_2SO_4, and it is more readily reduced, reacting with hydrobromic acid and hydriodic acid to form selenious acid or (at high concentration) elemental selenium. Like sulfate ion, SO_4^{2-}, SeO_4^{2-} is tetrahedral in crystals.

Hydrogen selenide, H_2Se, is a stronger acid than H_2S (ionization constants of H_2Se, 1.88×10^{-4} and about 10^{-10}) and is less readily obtained from selenides than H_2S from sulfides (the selenides of aluminum, iron and magnesium, Al_2Se_3, FeSe, and MgSe, require heating with H_2O or dilute acids). In general, the metal selenides are prepared by direct combination of the elements. Those of transition groups 3–8, 1 and 2 and main groups 3 and 4 exhibit many instances of well-defined compounds, berthollide compounds, and substitutional solid solutions. Thus four intermediate phases are found in the palladium-selenium system, Pd_4Se, $Pd_{2.8}Se$, $Pd_{1.1}Se$, and $PdSe_2$.

Selenium hexafluoride, SeF_6, the only clearly defined hexahalide, is formed by reaction of fluorine with molten selenium. It is more reactive than the corresponding sulfur compound, SF_6, undergoing slow hydrolysis. Selenium forms tetrahalides with fluorine, chlorine, and bromine, and dihalides with chlorine and bromine. However, other halides can be found in complexes, e.g., treatment of the pyridine complex of SeF_4 in ether solution with HBr yields $(py)_2SeBr_6$. Selenium tetrafluoride also forms complexes with metal fluorides, giving $MSeF_5$ complexes with the alkali metals.

Selenium forms several oxyhalides, e.g., $SeOF_2$, $SeOCl_2$, and $SeOBr_2$, the first two being liquids and the last a crystalline solid, mp 41.6°C. Selenium also forms tetraselenium tetranitride, Se_4N_4.

Selenocyanates, M^ISeCN, corresponding to the thiocyanates, are prepared by addition of selenium to soluble cyanides. They are similar to the thiocyanates except that HSeCN immediately decomposes in acid to selenium and hydrogen cyanide. The heavy metal selenocyanates are less soluble than the corresponding thiocyanates.

Selenium forms "thio"-type compounds, such as $SeSO_3$ by reaction of selenium and sulfur trioxide, $SeSO_3^{2-}$ (selenosulfates) by reaction of selenium and sulfites, SeS^{2-} (selenosulfides) by reaction of selenium with sulfides, as well as diselenides, Se_2^{2-}, and polyselenides, Se_x^{2-}.

Carbon diselenide is an evil-smelling liquid, and COSe and CSSe are also known.

Biological Role of Selenium: Some very interesting examples of the effect of soils on the nutritional quality of plants are associated with selenium. The element has not been found to be required by plants, but it is required in very small amounts by warm-blooded animals and probably by humans. However, selenium in larger quantities can be very toxic to animals and humans.

In large areas of the world, the soils contain very little selenium in forms that can be taken up by plants. Crops produced in these areas are, therefore, very low in selenium. A selenium deficiency in livestock is a serious problem. A deficiency causes a form of muscular dystrophy in younger animals and poor reproductive qualities in the adult animals. For prevention, sodium selenate or sodium selenite, sometimes augmented with vitamin E, is added in proper proportions to feedstuffs. Some areas, including the Plains and Rocky Mountain states in the United States have soils that are rich in available selenium. In regions like these, selenium toxicity is a problem. The situation is particularly serious in Arizona, California, Montana, Nevada, New Mexico, and South Dakota.

An interesting feature of selenium is that it occurs naturally in several compounds and these vary greatly in their toxicity and in their value in preventing selenium-deficiency diseases. In its elemental form, selenium is essentially insoluble and biologically inactive. Inorganic selenates or selenites and some of the selenoamino acids in plants are very active biologically, whereas some of their metabolites that are excreted by animals are not biologically active. In well-drained alkaline soils, selenium tends to be oxidized to selenates and these are readily taken up by plants, even to levels that may be toxic to the animals that eat them. In acid and neutral soils, selenium tends to form selenites and these are insoluble and unavailable to plants. Selenium deficiency in livestock is most often found in areas with acid soils and especially soils formed from rocks low in selenium.

In 1934, the mysterious livestock maladies on certain farms and ranches of the Plains and Rocky Mountain states were discovered to be due to plants with so much selenium that they were poisoning grazing animals. Affected animals had sore feet and lost some of their hair; many died. Over the next 20 years, researchers found that the high levels of selenium occurred only in soils derived from certain geological formations of high selenium content. They also found that a group of plants, called *selenium accumulators*, had an extraordinary ability to extract selenium from the soil. These accumulators were mainly shrubs or weeds native to semiarid and desert rangelands. They usually contained about 50 parts per million (ppm) or more of selenium, whereas range grasses and field crops growing nearby contained less than 5 ppm selenium. These findings helped ranchers to avoid the most dangerous areas when grazing livestock.

In 1957, selenium was found to be essential in preventing liver degeneration of laboratory rats. Since then, research workers have found that certain selenium compounds, either added to the diet or injected into the animal, would prevent some serious disease of lambs, calves, and chicks. That selenium is an essential nutrient element for birds and animals has been established.

In most diets used in livestock production, from 0.04 to 0.10 ppm of selenium protects the animal from deficiency diseases. If the diet is very high in vitamin E, the required level of selenium may be lower.

In terms of human dietary requirements, much of the wheat for breadmaking in the United States is produced in selenium-adequate sections of the country. Bread is generally a good source of dietary selenium.

Selenomethionine decomposes lipid peroxides and inhibits in vivo lipid peroxidation in tissues of vitamin-E-deficient chicks. Selenocystine catalyzes the decomposition of organic hydroperoxides. Selenoproteins show a high degree of inhibition of lipid peroxidation in livers of sheep, chickens, and rats. Thus, some forms of selenium exhibit in vivo antioxidant behavior.

Additional Reading

Carter, G. F., and D. E. Paul: "Materials Science and Engineering," ASM International, Materials Park, Ohio, 1991.

Liotta, D., and R. Monahan III: "Selenium in Organic Synthesis," *Science*, **221** 356–361 (1986).

Marshall, E.: "High Selenium Levels Confirmed in Six States," *Science*, **231**, 111 (1986).

Meyers, R. A.: "Handbook of Chemicals Production Processes," McGraw-Hill, New York, 1986.

Reamer, D. C., and W. H. Zoller: "Selenium Biomethylation Products from Soil and Sewage Sludge," *Science*, **208**, 500–502 (1980).

Sax, N. R., and R. J. Lewis, Sr.: "Dangerous Properties of Industrial Materials," 8th Edition, Van Nostrand Reinhold, New York, 1992.

Staff: "Plants Can Eat Toxic Soil," *National Food Review*, 42 (October–December, 1989).

Staff: "ASM Handbook—Properties and Selection: Nonferrous Alloys and Pure Metals," ASM International, Materials Park, Ohio, 1990.

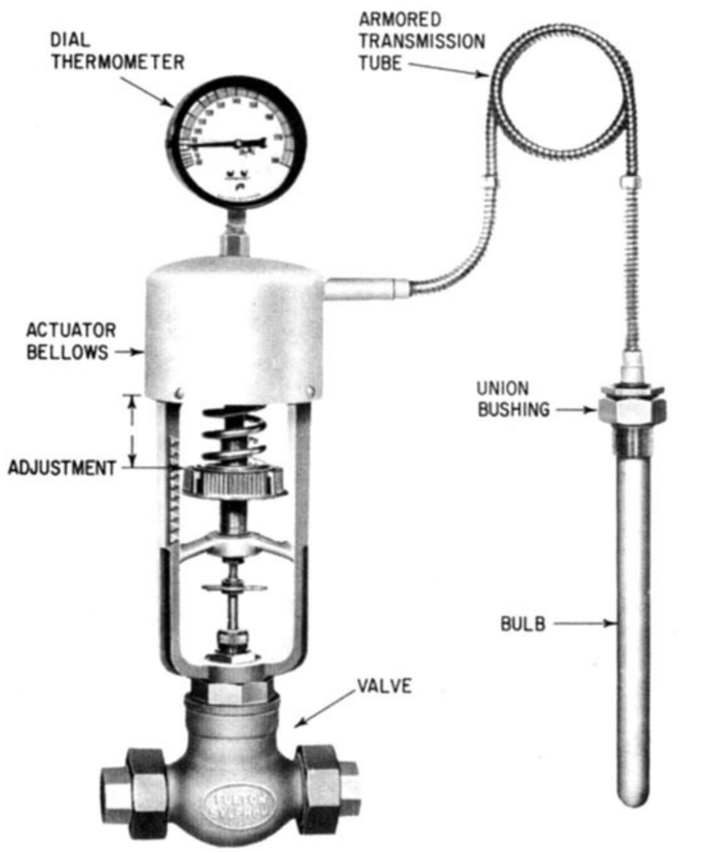
Staff: "Handbook of Chemistry and Physics," 73rd Edition, CRC Press, Boca Raton, Florida, 1992–1993.

SELF-DIAGNOSTICS. A term commonly used in the sensors, instrumentation, process control, and computer field to designate separate circuitry and instructions (self-contained) to test and sometimes recalibrate an electronic measurement and supervisory piece of equipment.

SELF-ENERGY (Particle). Classically, the energy of interaction between different parts of the particle (considered, for example, as a ball of charge). In quantized field theory (see **Field Theory**) the contribution to the Hamiltonian arising from the virtual emission and absorption of other particles, especially photons or mesons. In terms of mass-energy equivalence, the energy equivalent of the rest mass of the particle.

SELF-INDUCTANCE. The ratio of the magnetic flux linking a circuit to the flux-producing current in that circuit.

SELF-OPERATED CONTROLLER. Often termed a *regulator*, a self-operated controller requires no external power for automatically controlling a variable, notably pressure, temperature, liquid level, or flow. Some of the styles of self-operated controllers are: (1) *directly-actuated* (DA), where the valve is positioned directly by the measuring element; (2) *pilot-actuated* (PA), where the measuring element controls a pilot valve. The latter admits supply-line pressure to a piston or diaphragm to move the main valve; (3) *self-contained* (SC), where the measuring element is an intimate part of the valve structure. This type of device responds only to variations, such as pressure, temperature, or



Direct-actuated remote-sensing self-operated temperature controller or regulator. (*Robertshaw.*)

flow, that occur in the flowing medium; and (4) *remote-sensing* (RS), where the measuring element is separated from the valve.

Usually, regulators are furnished with only a proportional control action, with the proportional band established by the manufacturer. Control accuracy obtainable with self-operated controllers is on the order of: temperature regulators, $\pm 1°F$; pressure regulators, ± 0.5 psi; level regulators, $\pm 1\%$. Self-operated controllers are advantageous where control specifications can be met because of their simplicity, ease of service, general ruggedness, comparatively low cost, no need for external power (hence extra wiring or piping), and normally long useful life. Limitations include the rather coarse accuracy performance previously mentioned, allowable pressure drop through the valve, limited availability of ranges, and the usually relatively large measuring element.

With reference to the accompanying figure, the direct-actuated, remote-sensing temperature regulator employs a vapor-filled thermal system. When the temperature at the bulb rises, vapor pressure in the system is increased and this, in turn, is transmitted to the bellows. When the force produced equals the adjustable spring force, further pressure increase will cause a downward movement of the stem, thus causing the valve to commence closing. The reverse process occurs, of course, upon a drop in temperature.

SELLMEIER EQUATION. An equation often used for media which show anomalous dispersion

$$n^2 = 1 + \sum \frac{A\lambda^2}{\lambda^2 - \lambda_1^2} + \frac{B\lambda^2}{\lambda^2 - \lambda_2^2} + \cdots$$

where n is the refractive index, λ the wavelength of the light, λ_1 and λ_2 are the wavelengths of absorption lines, and $A, B, \ldots$ are constants to be determined from experimental data. If the equation is expanded in a series it becomes

$$n^2 = a + \frac{b}{\lambda^2} + \frac{c}{\lambda^4} \cdots$$

which has the same form as the Cauchy dispersion formula (but note that n occurs to the first power there).

SEMICARBAZONES. The products of the reaction between an aldehyde or a ketone with semicarbazide are termed *semicarbazones.*

$CH_3{\cdot}CHO + NH_2{\cdot}CO{\cdot}NH{\cdot}NH_2 \rightarrow$
(acetaldehyde) (semicarbazide)

$CH_3{\cdot}CH{:}N{\cdot}NH{\cdot}CO{\cdot}NH_2 + H_2O$
(acetaldehyde semicarbazone)

$(CH_3)_2CO + NH_2{\cdot}CO{\cdot}NH{\cdot}NH_2 \rightarrow$
(acetone)

$(CH_3)_2C{:}N{\cdot}NH{\cdot}CO{\cdot}NH_2 + H_2O$
(acetone semicarbazone)

SEMICONDUCTORS. Materials and devices known as *semiconductors* have been the backbone of the electronics industry for many years. Semiconductors did not enter the industry in a major way, however, until several years after the vacuum tube (valve) had been well established. In terms of perspective, it is interesting to note that at least one semiconductor device predated the vacuum tube in the early days of radio communication. This was the then familiar galena crystal and accompanying whisker used in early crystal set radio receivers.

With the continuing attention to developing reliable and efficient vacuum tubes, which occurred over a long time span, interest in semiconductors stagnated—with the exception of the emerging radar technology of the World War II era. Massachusetts Institute of Technology's Radiation Laboratory became active in the investigation of crystal rectifiers and engaged in exploratory studies and in the development of very pure semiconducting materials, notably silicon and germanium. Several other research institutions during this same period became interested in the theoretical aspects of solid state and semiconductors, including the energy levels of solids and the charged carrier transport in silicon and a few other materials. Out of this early phase of semiconductor technology came, in 1947, the invention of the transistor (contraction for transfer resistor). Inventors Shockley, Bardeen, and Brattain (Bell Laboratories) received the Nobel Prize in physics in 1956 for their accomplishment.

The transistor had a tremendous impact, constituting the birth of modern electronics. The transistor led to the development of almost innumerable semiconductor device configurations and ultimately to the phasing out of the vacuum tube except for rather special and limited applications.

Innovations to fabricate semiconductors with improved performance and smaller size (microelectronics) for many years has been a continuing process without apparent end. As of the very late 1980s, the enhancement of semiconductors continues apace. Circuit integration has gone through several major phases—from IC (integrated circuit) to LCI (large-scale integration) to VLSI (very large-scale integration) to VHSIC (very high-speed integrated circuits) to ULSI (ultralarge-scale integration)—to the point where it is difficult to find appropriate adjectives to describe developments in ICs.

With few exceptions, it is only within recent years that serious concern has been expressed by experts in what the physical limits on size and performance may be. Fortunately, thus far the advancements in fabrication (chip-making processes), such as electron beam and molecular beam lithography, have allayed these concerns in the short term. Although some interest in nonsilicon materials has always been present, there has been a recent reawakening of interest in the use of gallium, indium, arsenic, phosphorus, and antimony—this interest intensifying because of the opportunities such elements offer over silicon. Although silicon which for years has been the basis for volume-produced devices, is not seriously threatened at this juncture, some authorities feel that, in the long term, the emphasis on silicon will be heavily shared with these other materials.

Even with the aforementioned materials, the microelectronics and optoelectronics industries have relied almost exclusively on *inorganic* materials. This is expected to change. Some authorities now forecast that by the early 2000s, there will be a major shift, namely, to *molecular electronics.* Even today, considerable attention for the long term is being paid to the *organic* solid state. The richness of the variety of organic molecular materials available offers enormous potential compared with the relative paucity of structures achievable with inorganic compounds, even when due allowance is made for the exciting developments in inorganic quantum well semiconductors. A hint of what may be achieved along these lines is given by the progress thus far made in liquid crystals, piezoelectric and photoconducting polymers. This implies, of course, that organics no longer will be confined to their traditional applications in electronics, such as for insulation, adhesion, or encapsulation.

Molecular electronics currently is defined as the use of organic molecular materials to perform an active function in the processing of information and its transmission and storage. An alternative definition has been suggested, namely, the achievement of *switching* on a molecular scale. As observed by G. G. Roberts (University of Oxford), "It is interesting to note that only a modest diminution in the size of electronic circuit components is required before the scale of individual molecules is reached; in fact many existing circuit elements could already be accommodated within the area occupied by a leukemia virus."

Some investigators forecast that during the first quarter of the 21st Century, molecular electronics will lead to *supermolecular electronics* where signal transport and control will be effected by nanometer-scale assemblies.

Numerous applications of contemporary semiconductors are described throughout this encyclopedia. See **Molecular and Supermolecular Electronics.**

Nature of Semiconductors. From the standpoint of their use in electronics, semiconductors are distinguished from other classes of materials by their characteristic electrical conductivity σ. The electrical conductivities of materials vary by many orders of magnitude, and consequently can be classified as: (1) the perfectly conducting superconductors; (2) the highly conducting metals ($\sigma \approx 10^6$ mho/centimeter); (3) the somewhat less conducting semimetals ($\sigma \approx 10^4$ mho/centimeter); (4) the semiconductors covering a wide range of conductivities ($10^3 \gtrsim \sigma \gtrsim 10^{-7}$ mho/ centimeter); and (5) the insulators, also covering a wide range ($10^{-10} \gtrsim \sigma \gtrsim 10^{-20}$ mho/centimeter).

These low-conductivity materials are characterized by the great sensitivity of their electrical conductivities to sample purity, crystal perfection, and external parameters, such as temperature, pressure, and frequency of the applied electric field. For example, the addition of less than 0.01% of a particular type of impurity can increase the electrical conductivity of a typical semiconductor like silicon or germanium by six or seven orders of magnitude. In contrast, the addition of impurities to typical metals and semimetals tends to decrease the electrical conductivity, but this decrease is usually small. Furthermore, the conductivity of semiconductors and insulators characteristically decreases by many orders of magnitude as the temperature is lowered from room temperature to 1 K. On the other hand, the conductivity of metals and semimetals characteristically increases in going to low temperatures, and the relative magnitude of this increase is much smaller than are the characteristic changes for semiconductors. The principal conduction mechanism in metals, semimetals, and semiconductors is electronic, whereas both electrons and the heavier charged ions may participate in the conduction processes of insulators.

Classification. It is customary to classify a semiconductor according to the sign of the majority of its charged carriers, so that a semiconductor with an excess of negatively charged carriers is termed *n*-type. A semiconductor with an excess of positively charged carriers is called *p*-type, while a material with no excess of charged carriers is considered to be perfectly compensated. Many of the important semiconductor devices depend upon fabricating a sharp discontinuity between the *n*- and *p*- type materials, the discontinuity being called a *p-n* junction.

Other Characteristics. Even though most semiconductors exhibit a metallic luster when inspected visually, this does not provide a reliable criterion for their classification, since the electrical conductivity of all materials is frequency dependent. Visual inspection tends to be sensitive to the conductivity properties at visible frequencies ($\sim 10^{15}$ Hz). Although materials with a high optical reflectivity tend also to exhibit high dc conductivity, these two properties are not necessarily correlated in semiconductors and metals. An example of a metal without metallic luster is ReO_3 (rhenium trioxide), a semitransparent reddish solid. On the other hand, most of the common semiconductors do exhibit metallic luster primarily because electronic excitation across their fundamental energy gaps can be achieved at infrared frequencies. At low frequencies, the principal conduction mechanism is free carrier conduction, which is important in metals and is present to some extent in semiconductors which contain impurities or are found at elevated temperatures. In contrast, interband transitions dominate the conduction process at very high frequencies. Interband transitions contribute to the conductivity by about the same order of magnitude in semiconductors, metals, and insulators.

Since the dc conductivity due to free carriers is characteristically low in semiconductors and insulators, the generation of free carriers by exposure to light at infrared, visible, and ultraviolet frequencies can lead to a large increase in the dc conductivity. This photoconductive effect, which is not observed in metals or semimetals, can be enormous in low-conductivity semiconductors (an increase in the dc conductivity of CdS (cadmium sulfide) by 8 orders of magnitude is observed). The effects of ultraviolet light, for example have been used successfully in reprogramming EPROM (electrically programmable read-only memory) devices, exposure to UV light causing trapped charges to leak off.

Because of the extreme sensitivity of semiconductors to impurities, temperature, pressure, light exposure, and certain other factors, these materials can be exploited in the fabrication of useful devices, such as the crystal diode, the transistor, integrated circuits, photodetectors, and light switches.

Flow of Current in Semiconductors. The flow of electric current depends upon the acceleration of charges by an externally applied electric field. Only those charges that resist collisions or scattering events are effective in the conduction process. Because of collisions, charged particles in a solid are not accelerated indefinitely by the applied field, but rather, after every scattering event, the velocity of a charged particle tends to be randomized. Thus the acceleration process must start anew after each scattering event and charged particles achieve only a finite velocity along the electric field **E**, the average value of the velocity being denoted by v_D, the drift velocity. The effectiveness of the charge transport by a particular charged particle is expressed by the mobility μ, which is defined as $\mu = v_D/E$. The mobility of a particle with charge *e* and mass *m* can be related directly to the mean time between scattering events (also called the relaxation time) by the expressed $\mu = e\tau/m$. The electrical conductivity σ depends upon the mobility of the charged carriers as well as on their concentration *n*, and is simply written as $\sigma = ne\mu$, where *e* is the charge of the carriers. The advantage of expressing the conductivity in this form is the explicit separation into a factor *n* which is highly sensitive to external parameters, such as temperature, pressure, optical excitation, irradiation, and into another factor μ, which depends characteristically on scattering mechanisms and on the electronic structure of the semiconductor.

The classical theory for electronic conduction in solids was developed by Drude in 1900. This theory has since been reinterpreted to explain why all contributions to the conductivity are made by electrons which can be excited into unoccupied states (Pauli principle) and why electrons moving through a perfectly periodic lattice are not scattered (wave-particle duality in quantum mechanics). Because of the wavelike character of an electron in quantum mechanics, the electron is subject to diffraction by the periodic array, yielding diffraction maxima in certain crystalline directions and diffraction minima in other directions. Although the periodic lattice does not scatter the electrons, it nevertheless modifies the mobility of the electrons. The cyclotron resonance technique is used in making detailed investigations in this field.

The origin of the energy barrier[1] for carrier generation is directly connected with the energy levels for electrons in a solid. Considering electrons in a solid from a tight-binding point of view, the discrete energy levels of the free atom broaden in the solid to form energy bands. For materials which are well described by the tight-binding approximation, the width of the energy bands is sufficiently small so that an energy gap is formed between the energy bands; in the forbidden energy gap there are no bound states. Of particular importance to the conduction properties of a solid is the fact that *all* of the available states in each band would be filled if each atom were to contribute exactly two electrons, thereby causing every solid with an odd number of electrons per atom to be metallic; while solids with an even number of electrons per atom would be insulating or semiconducting. The occurrence of energy bandgaps is also a consequence of the weak binding approximation, whereby the periodic potential itself is responsible for creating bandgaps through the mixing of states separated by a reciprocal lattice vector. See also **Solid-State Physics.**

For semiconductors, the excitation energy lies in the range 0.1 to about 2 eV. Thermal fluctuations are sufficient to excite a small, but significant, fraction of electrons from the occupied levels (the valence band) into the unoccupied levels (the conduction band). Both the excited electrons and the empty states in the valence band (aptly called *holes*) may move under the influence of an electric field, providing a means for conduction of current. (A hole acts like an electron with a positive charge.) Such electron-hole pairs may be produced not only by thermal energy, but also by incident light, providing photo-effects.

Crystallographic defects, in general, are also electronic defects. In metals, they provide scattering centers for electrons, increasing the resistance to charge flow. The resistance wire in many electric heaters, in fact, consists of an ordinary metal, such as iron, with additional alloying elements, such as nickel or chromium, providing scattering centers for electrons. In semiconductors and insulators, alloying elements and defects provide an even greater variety of effects, since they can change the electron-hole concentrations drastically in addition to providing scattering centers. The semiconductor industry has been built on the alloying of silicon, and a few other elements, including germanium, with *selected impurities* in carefully controlled concentration and geometry. Some of the other emerging semiconductor materials are described later.

[1]The sets of discrete but closely adjacent energy levels, equal in number to the number of atoms, that arise from each of the quantum states of the atoms of a substance when the atoms condense to a solid from a nondegenerate gaseous condition make up the energy band (also called the Bloch band). For a semiconductor, the highest energy level is the conduction band, containing only the excess electrons resulting from crystal impurities. The next highest level is the valence band, usually completely filled with electrons. In between these bands is the forbidden band, which is wider for an insulating material than for a semiconductor and vanishes in a conducting material.

Doping. This is a process for purposely adding impurities to a semiconductor (or production of a deviation from stoichiometric composition in order to achieve a desired characteristic). Doped material thus is no longer intrinsic but is impure and called extrinsic. If a trivalent impurity is introduced into silicon or germanium, *holes* are created and the material is said to be *p*-type. Introduction of a pentavalent element into silicon or germanium, on the other hand, creates *free electrons* and the material is said to be *n*-type. Because of thermal effects, free electrons and holes are always being produced in silicon and germanium (intrinsic generation of electron-hole pairs). Consequently, there will be some electrons in *p*-type material and some holes in the *n*-type material. These carriers are referred to as *minority carriers*. Electrons in *n*-type material and holes in *p*-type material are termed *majority carriers*.

The process of placing impurities in the near-surface region of solids is accomplished by a procedure known as *implanting*. A commonly used implanting procedure is to accelerate impurity ions in an electrostatic field with the energy sufficient to impinge with the desired force on the solid target. Known as *ion implantation*, this carefully controlled and reproducible procedure has been widely used to dope semiconductors to create *p-n* junction formations. A certain amount of damage, however, occurs in the semiconductor material in some cases. The surface may become amorphous, or because the implanted dopants may not reach substitutional spots in the crystal lattice the ions may not become electrically active. Thus, it is necessary to anneal the solid for electrical activation of the implanted ions as well as to remove any damage.

Semiconductor Device Configurations

Representative of semiconductor devices is the diode, a two-terminal device which has the property of permitting current to flow with practically no resistance in one direction and offering nearly infinite resistance to current flow in the opposite direction. Applications of the diode are numerous, as in gating circuits used in digital computers.

A widely used semiconductor diode is the *p-n junction diode*. Imagine a crystal (single) of silicon doped so half the material is *p*-type and the other half is *n*-type. The internal boundary between the two extrinsic regions is a *p-n junction*, and the resulting device is a *diode*. See Fig. 1.

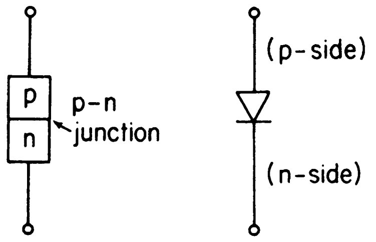


Fig. 1. (*Left*) Configuration of *p-n* junction diode. (*Right*) electrical symbol.

Three possible configurations of the *p-n* junction are shown in Fig. 2. The energy diagrams for these three configurations are also shown in Fig. 2. Similar diagrams can be generated for holes. When the diode is unbiased, no net flow of electrons takes place across the junction. Assuming that some electrons on the *n*-side have sufficient energy to overcome the potential hill, electrons on the *p*-side (minority carriers) "slide down" the hill, making the net current flow zero. For the reverse biased

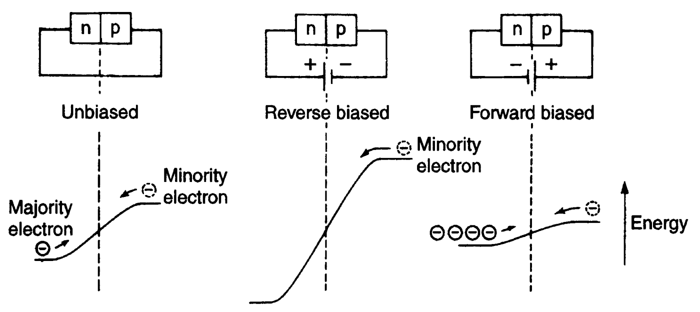


Fig. 2. Three possible configurations of a *p-n* junction diode.

example, the potential hill is raised and only the few minority carriers from the *p*-side slide down. This results in a minute reverse saturation current. When the diode is forward biased, the potential hill is lowered. This enables electrons to climb over the hill and current flow occurs. The same considerations apply to holes. In fact, the total diode current is equal to the sum of the electrons and holes flowing across the junction.

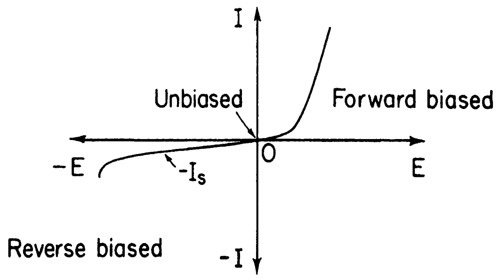


Fig. 3. Characteristic curve of a semiconductor diode.

The characteristic curve of a semiconductor diode is shown in Fig. 3. An equation for this curve, called the *rectifier equation*, is expressed as:

$$I = I_s\,(e^{-11600\,E/T} - 1)$$

where I = diode current, amperes
I_s = reverse saturated current (temperature dependent), amperes
E = diode biasing voltage ($+E$ for forward bias; $-E$ for reverse bias), volts
T = absolute temperature (0°C + 273°), degrees Kelvin

At room temperature (300 K) and $E > 0.1$ volt,

$$I \cong I_s e^{39E}$$

Where E is more negative than 0.1 volt,

$$I \cong -I_s$$

An example of a simple rectifier employing a *p-n* junction diode is given in Fig. 4. During the positive half-cycle (0° to 180°) of the ac sinusoidal waveform v_s, the diode is forward-biased and conducts. The voltage v_L across load resistance R_L is, therefore, nearly identical to that of v_s for the positive half-cycle. For the negative half-cycle (180° to 360°), the diode is reverse biased and does not conduct. No current flows in R_L, and $v_L = 0$ during the negative half-cycle. Because the diode conducts for only one-half cycle, the circuit of Fig. 4 is called a *half-wave rectifier*. The waveform of v_L is only unidirectional. To obtain steady dc, like that from a battery, a filter is required. An example of an elementary filter is a large-valued capacitor placed across the load resistor.

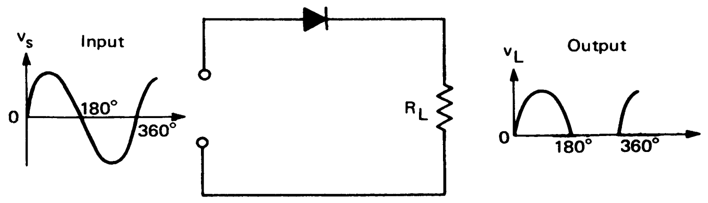


Fig. 4. Simple rectifier employing a *p-n* junction diode.

The circuit of Fig. 4 can also be used as a detector of amplitude-modulated (AM) radio waves. Fig. 5(a) illustrates the components of an AM wave. If this is applied to the input of Fig. 4, the wave is rectified and the output appears as shown in Fig. 5(b). Placing a small-valued

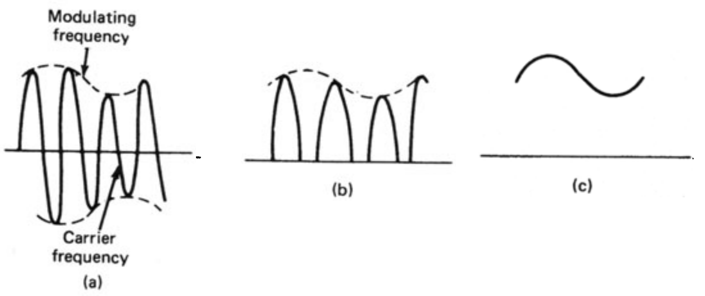


Fig. 5. Use of *p-n* junction diode as AM radio detector.

capacitor across R_L filters out the carrier frequence and the desired modulating signal is obtained, as shown in Fig. 5(c).

A bipolar transistor on a single silicon crystal is diagrammed in Fig. 6.

Metal Oxide Semiconductors. The metal oxide semiconductor field effect transistor (MOSFET) is representative of another class of semiconductors. In *n*-MOS device, two islands of *n*-type silicon are created in a *p*-type silicon substrate. A thin layer of nonconducting SiO_2 lies on top of the silicon substrate. Direct connections on a *source* and *drain* are made to the two islands, while a metal *gate* is coupled to the silicon substrate by capacitance. Usually the source and substrate are electrically connected and held at a potential of zero volts. The drain is held at a positive voltage. In this condition, no current flows into the MOS device. When a positive potential is applied to the gate, the electric field attracts a majority of electrons to the thin layer at the surface of the crystal under the gate. Since this region is normally *p*-type, the surface becomes "inverted" creating a continuous *n*-type channel between source and drain, thus allowing large currents to flow. This creates a current amplification as in a bipolar transistor. An advantage of MOSFET over bipolar transistors is that they require no isolation islands and thus can be packed more closely on a silicon chip.

Complementary MOS devices (CMOS) have been widely used in recent years. See Fig. 7. The CMOS is made up of a *n*-MOS and a *p*-MOS. A main advantage of the CMOS is its low power consumption.

Other Materials for Semiconductors. Although silicon (and germanium at one time) is the unquestioned principal semiconductor material as of the late 1980s, silicon does have limitations. For example, it is not easy to integrate electronic and photonic devices in the same microchip. Silicon has a relatively narrow range of temperature tolerance, is susceptible to radiation damage and has "slow" electrons compared with some other materials. Elements in Periodic Table Groups III and IV (now officially called Groups 13 and 15) have fast electrons. For example, the differences between electrons in gallium arsenide and silicon stem basically from the differing chemical characteristics. Electrons in gallium arsenide at low electric fields behave like very light particles which can move easily through the vibrating (and obstructing) crystal lattice of atoms. In contrast, the electrons in silicon behave like heavy particles which move sluggishly under the influence of an applied voltage. The result is significantly faster operating times in microchip operation. See Fig. 8. Fast electrons translate into fast switches. Such switches, when multiplied by thousands or even hundreds of thousands, comprise the basic building blocks of a digital integrated circuit (commonly called a microchip). As pointed out by Allyn, Flahive, and Wemple (1986), there are two classes of speed: (1) Maximum speed achievable, no matter how much "push" is provided by the applied voltage. This is known as *saturated voltage*. Gallium arsenide materials have an advantage in saturated voltage over silicon of 1.5; and with indium—gallium arsenide compounds, the advantage reaches 2.5. (2) The second speed relates to the ease with which electrons can be brought up to full speed (*low-field electron mobility*). Higher mobility in the Groups 13–15 (III–V) semiconductors means that the electrons reach full speed at lower operating voltages. These speed advantages are particularly important in terms of interdevice wiring, which tends to dominate the speed of high-density microchips. Major emphasis on these newer semiconductor materials is directed on the Schottky gate field effect transistor. See Fig. 9.

In the 1950s and 1960s, considerable investigation was made of amorphous *chalcogenide glasses* for possible use in semiconductor devices. The glasses are named for the chalcogens (Group 16, formerly Group VI in the Periodic Table). Early in their consideration, these materials created a considerable controversy among solid-state physicists. Claims were made and challenged as regards their possible impact on further revolutionizing the semiconductor industry. However, it has been shown that chalcogenide glasses can "switch," but some scientists observe that almost any material will switch under the right conditions. Compositions proposed for memory switches are exemplified by $Te_{81}Ge_{15}Sb_2S_2$, and for non-memory switch materials, $Te_{40}As_{35}Si_{18}Ge_7$. It has also been shown that transitions occur in these glasses when they are exposed to intense light and thus possible photographic uses have been proposed.

Semiconductors used in solar cells are described under **Solar Energy.**

Gallium Arsenide Power Sources. GaAs was first synthesized in 1929 by V. M. Goldschmidt. Its semiconducting properties were not studied until 1952 by H. Welker. The first GaAs p-n junction used for power generation at microwave frequencies was the tunnel diode. Later, GaAs varactor diodes were used in harmonic frequency multipliers and parametric amplifiers at microwave and mm-wave frequencies because of the inherent higher cut-off frequencies possible with gallium arsenide. In 1963, J. B. Gunn discovered the negative resistance property of GaAs, after which GaAs diodes and field effect transistors (FETs) were developed.

The idea for using diodes for generation and amplification of power at microwave frequencies was suggested by A. Uhlir, Jr. Frequency multipliers have been used for power generation since 1958. These devices depend on the nonlinear reactance or resistance characteristics of semiconductor diodes. Generally, there are three types of multiplier diodes—step recovery diodes, variable resistance multiplier diodes, and variable capacitance multiplier diodes.

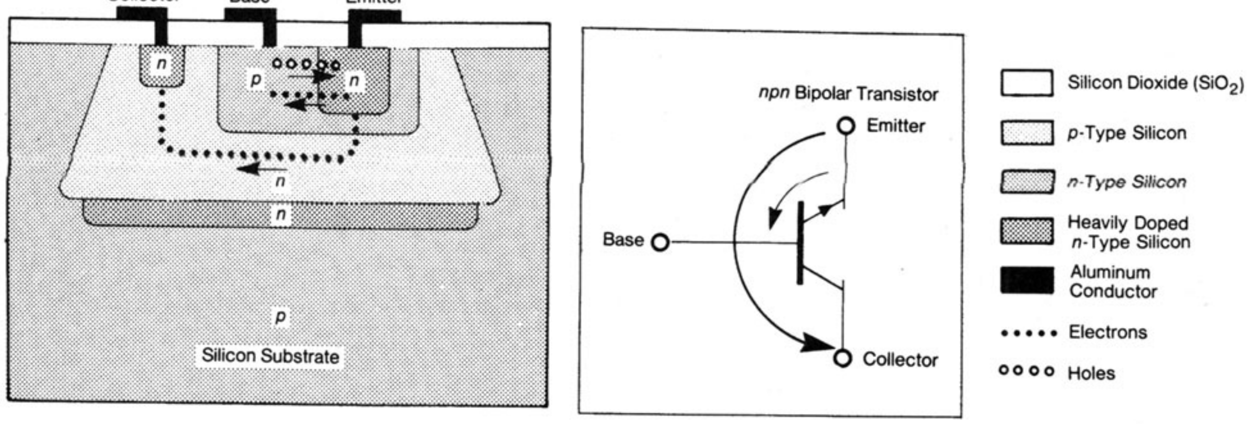


Fig. 6. A bipolar transistor on a single silicon crystal. (*Kurnik, Chemical Engineering Progress.*)

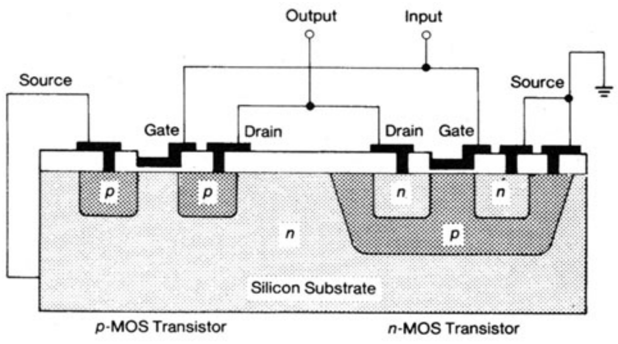


Fig. 7. Complementary MOS device (CMOS) on a single silicon crystal. (*Kurnik, Chemical Engineering Progress.*)

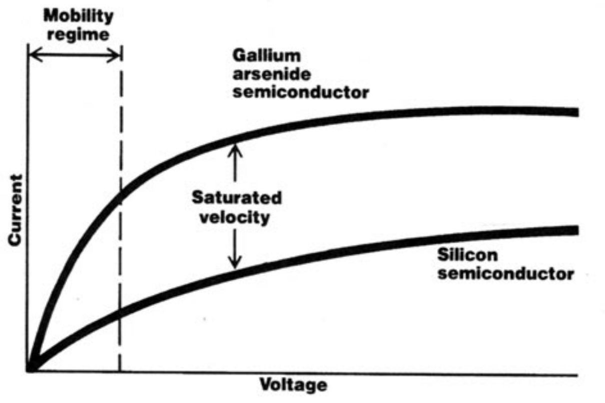


Fig. 8. Current-voltage characteristics of two hypothetical devices of identical physical size. The gallium arsenide curve rises faster and reaches peak velocity faster than the silicon. This means that the group III–V (13–15) electrons produce significantly faster operating times in microchips. (*AT&T Technology.*)

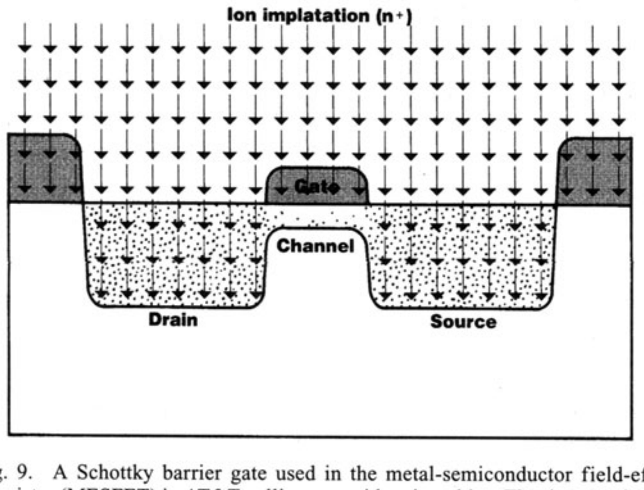


Fig. 9. A Schottky barrier gate used in the metal-semiconductor field-effect transistor (MESFET) in AT&T gallium arsenide microchips. The tiny gate is only one micrometer wide ($\frac{1}{25,400}$ inch). The gate electrode is deposited before the ion-implantation process so that the gate material will "shade" the channel under it from the "ion rain" that doses the exposed material. (*AT&T Technology.*)

T. B. Ramachandran (Microwave Device Technology Corporation) notes that there are two inherent major disadvantages for current GaAs FET power devices:

1. The devices are surface oriented. Since the active region is close to the surface, the surface effects tend to affect the device performance. This may be seen in the noise performance of GaAs FETS close to the carrier.
2. To increase the power output, the breakdown voltage must be increased. Active channel doping has to be decreased in order to increase the breakdown voltage. This reduction in doping density decreases the maximum current density, and this tends to reduce the total power output.

J. B. Gunn (International Business Machines) noticed in 1963 the current instabilities in GaAs at high electric fields. Known as the Gunn effect, or the transferred electron effect, Gunn diodes have been used as a low-cost source for microwaves since 1968. These components are comparatively easy to manufacture and hence the cost is low. GaAs Gunn diodes are used from C-band (4 GHz) through W band (100 GHz).

W. T. Read (AT&T Bell Laboratories) first reported microwave oscillations in silicon p-n junctions in 1965. During the interim, much research has gone into developing impact ionization avalanche transit time (IMPATT) diodes for a variety of applications.

Research and Development Trends

Quantum-Effect Devices. There is a limit on the components of ordinary integrated circuits because "smallness" of size can interfere with their functionality. Such problems may be overcome through the use of quantum-effect semiconductor devices.

It has been predicted by a number of authorities that, before the year 2000, the physical laws that govern the behavior of circuit components may impede the ultimate shrinkage of the chip. As early as 1982, P. K. Chatterjee stressed how close the end point on downscaling components may be. Estimates of minimum feature size as of the early 1990s range between 100 and 500 billionths of a meter. As observed by R. T. Bate (Texas Instruments Incorporated), "The same solution that some of the very phenomena that impose size limits on ordinary circuits could be exploited in a new generation of vastly more efficient devices. The functional bases for these devices are quantum-mechanical effects that carry semiconductor technology into a realm of physics where subatomic particles behave like waves and pass through formerly impenetrable barriers. With the so-called quantum semiconductor device, I believe it will be possible to put the circuitry of a supercomputer on a single chip."

Doped silicon, doped and undoped gallium arsenide, and aluminum gallium arsenide have been used as the basis for quantum devices. Of course, size reduction of these proportions pose difficult production tasks. In addition to shrinking size, quantum devices can be expected to be faster and more efficient. A prototype quantum chip, with features one-hundredth of the size of an ordinary chip may appear as shown in Fig. 10. An operational semiconductor device based upon the quantum effect should appear prior to the year 2000. Aggressive research currently is being carried out by AT&T Bell Laboratories, IBM Corpora-

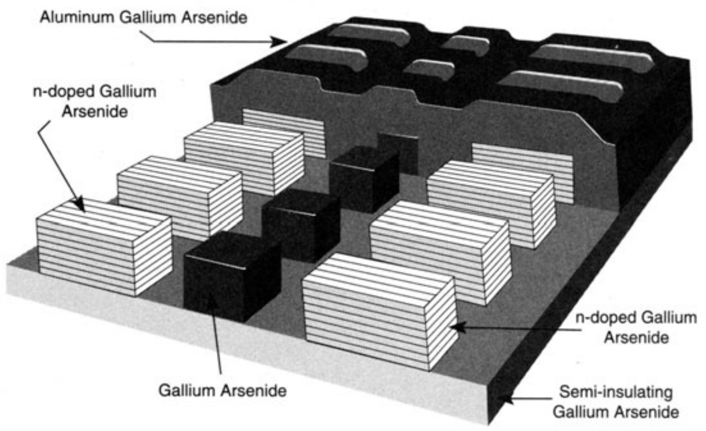


Fig. 10. Quantum chip consisting of four materials. Final product is about $\frac{1}{100}$th size of conventional chip. Current flows from one negatively doped (n-doped) gallium arsenide block to another by way of a layer of aluminum gallium arsenide, a gallium arsenide cube, and thence to an other aluminum gallium arsenide layer. Current conductivity of a quantum device is extremely sensitivity and thus capable of exacting control. (*This idealized model is suggested by R. T. Bate in the scholarly reference cited.*)

tion, the Massachusetts Institute of Technology, Hughes Research Laboratories, Texas Instruments Corporation, the University of Cambridge, Philips Research Laboratory, and the University of Glasgow, among others. As stressed by R. T. Bate, "The commitment of so many research teams to a problematic technology attests to the tremendous potential of these devices and to the faith that they will take the lead in the next semiconductor revolution. The costs and risks involved must be borne in order to revitalize a rapidly maturing electronics industry; the results can only benefit a society that has learned to depend on integrated circuits in many ways."

Atom Switch. By employing the technique of the Scanning Tunneling Microscope, D. M. Eigler and a research team at the IBM Almaden Research Division, San Jose, California, have improvised an "atom switch." Through careful movement of a single xenon atom between the microscope's tip or a nickel surface, the researchers have altered the amount of tunneling current between tip and sample. When the xenon rests on the surface, this is tantamount to the switch's off position. The switch is turned on by applying a 64-millisecond (0.8 V pulse) to the tip. This causes the xenon to jump to the tip, thus increasing the tunneling current by a factor of about seven. As of the present, no practical applications of this switching action are planned because the apparatus involved is bulky and costly. Some scientists believe that the principle ultimately may be useful for information storage systems. Other scientists have observed that, if storing a bit in a cluster of 1000 atoms ever becomes practical, a machine could be developed that would store the contents of the U.S. Library of Congress on a silicon disk only 12 inches (30 cm) wide. For more detail, see Yam reference listed.

Dynamical Phenomena at Metal and Semiconductor Surfaces. This topic has been investigated in recent years through the use of ultrafast measuring techniques involving lasers and nonlinear optics. As reported by J. Bokor (AT&T Bell Laboratories), "Understanding of the rates and mechanisms for relaxation of optical excitation of the surface itself as well as those of adsorbates on the surface is providing new insight into surface chemistry, surface phase transitions, and surface recombination of charge carriers in semiconductors." The combination of lasers and nonlinear optical techniques is now being brought to bear on the next frontier in surface physics, namely surface dynamics. Ultrafast lasers allow for the study of picosecond and femtosecond processes directly in the time domain, circumventing the ambiguities attendant on linewidth measurements for the determination of lifetimes. One may anticipate continued growth in the diversity of applications of these techniques to the understanding of the complexities of surface dynamics. See Fig. 11.

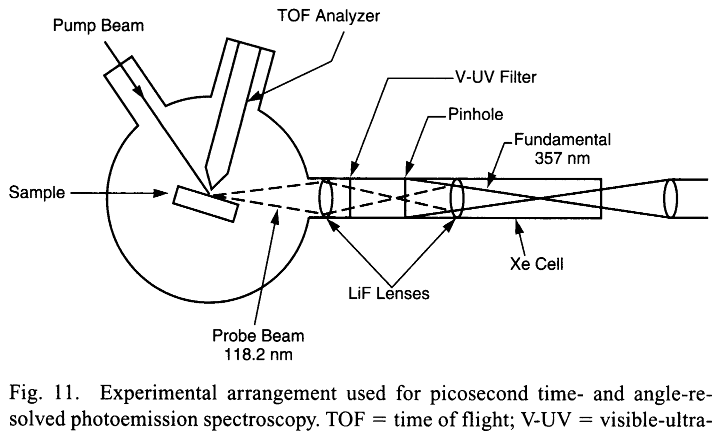


Fig. 11. Experimental arrangement used for picosecond time- and angle-resolved photoemission spectroscopy. TOF = time of flight; V-UV = visible-ultraviolet; LiF = lithium fluoride Xe = xenon. (*Source: AT&T Bell Laboratories.*)

Amorphous Silicon. According to P. G. LeComber (University of Dundee), the most important difference between crystalline silicon and an amorphous semiconductor is that in the latter there is a continuous distribution of localized states within the forbidden energy gap. Another important difference concerns the mobility of the electrons or holes. LeComber observes, "In an amorphous material the periodicity of the lattice only extends over a few atomic spacings. Under these conditions, the electron transport may no longer be considered as band motion with occasional scattering, as in crystalline theory. In this case, the electron motion is essentially a diffusive process which can be considered to be similar to the Brownian motion of small particles in liquids. Properties of particular importance in the application of amorphous silicon films include:

- Thin films (about 1 micrometer thick).
- Low deposition temperature.
- Large area growth on many substrates, such as glass, metals, and flexible plastics.
- Mechanically very hard.
- Chemically very stable.
- Inert material.
- Extremely photoconductive.
- Room temperature electrical conductivity can be controlled over ten orders of magnitude by doping for both n-type and p-type material.
- Ease of sequentially producing p-type and n-type material by switching from one gas mixture to another.
- Easy to pattern arrays of devices using conventional photolithographic techniques developed for crystalline silicon.

Hybrid Ferromagnetic-Semiconductor Structures. G. A. Prinz and researchers at the Materials Science and Technology Division of the Naval Research Laboratory, Washington, D.C., have been studying hybrid ferromagnetic-semiconductor materials through the use of modern thin-film techniques. Thus far, the team has researched and demonstrated combinations of Fe/Ge, Fe/GaAs, Fe/ZnSe, and Co/GaAs. The researchers observe, "Ultrahigh-vacuum growth techniques are being used to grow single-crystal films of magnetic materials. These growth procedures, carried out in the same molecular beam epitaxy systems commonly used for the growth of semiconductor films, have yielded a variety of new materials and structures that may prove useful for integrated electronics and integrated optical device applications."

Useful characteristics of hybrid ferromagnetic-semiconductor structures include:

1. Produce significant changes in the electrical and optical properties.
2. Coupling of devices to a radiation field, particularly in the microwave range.
3. Such devices provide a source of spin-polarized carriers.

Details of this research can be found in the Prinz reference listed.

Microclusters. These may be defined as small aggregates of atoms that make up a distinct phase of matter. The chemistry of clusters is highly reactive and selective. Their principle area of future application is catalysis. However, clusters also hold some promise for electronic applications. As observed by M. A. Duncan and D. H. Rouvray (University of Georgia), "Thin films of clusters possessing desirable electronic qualities could be of great interest in microelectronics. It is possible to envision applications in optical memories, image processing and superconductivity. Given the potential for construction of parts from networks of clusters, it may eventually be possible to make electronic devices on a molecular scale. Ultimately a machine might be designed that could serve as a link between solid-state electronics and biological systems, such as systems of neurons. Such a link might convey data from a television camera to the brain of a blind person."

See also separate article on **Molecular and Supermolecular Electronics**.

Additional Reading

Allison, J.: "Electronic Engineering Semiconductors and Devices," 2nd Edition, McGraw-Hill, New York, 1990.

Bate, R. T.: "The Quantum-Effect Device: Tomorrow's Transistor?" *Sci. Amer.*, 96 (March 1988).

Bierman, H.: "Material Advances Pave the Way for Device and System Improvements," *Microwave J.* 26 (October 1990).

Bokor, J.: "Ultrafast Dynamics at Semiconductor and Metal Surfaces," *Science*, 1130 (December 1, 1989).

Brodsky, M. H.: "Progress in Gallium Arsenide Semiconductors," *Sci. Amer.*, 68 (February 1990).

Brophy, J. J.: "Basic Electronics for Scientists," 5th Edition, McGraw-Hill, New York, 1990.

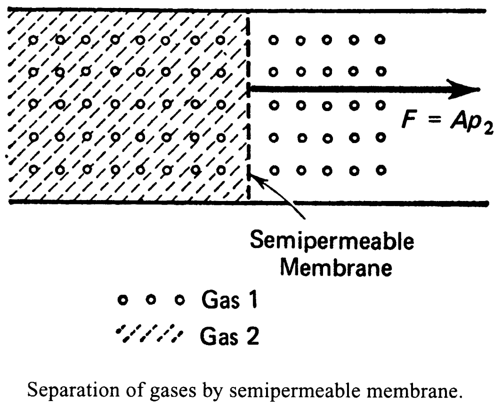
DiSalvo, F. J.: "Solid-State Chemistry: A Rediscovered Chemical Frontier," *Science*, 649 (February 9, 1990).

Duncan, M. A., and D. H. Rouvray: "Microclusters," *Sci. Amer.*, 110 (December 1989).
Ellowitz, H. I.: "1991 U.S. GaAs Foundry Update," *Microwave J.*, 42 (August 1991).
Fink, D. G., and D. Christiansen: "Electronics Engineers' Handbook," 3rd Edition, McGraw-Hill, New York, 1989.
Fisk, Z., et al.: "Heavy-Electron Metals: New Highly Correlated States of Matter," *Science*, 33 (January 1, 1988).
Geinovatch, V. G.: "Prognostications from the Edge," *Microwave J.*, 26 (April 1991).
Geis, M. W., and J. C. Angus: "Diamond Film Semiconductors," *Sci. Amer.*, 84 (October 1992).
Goldstein, A. N., Echer, C. M., and A. P. Alivisatos: "Melting in Semiconductor Nanocrystals," *Science*, 1425 (June 5, 1992).
Kemerley, R. T., and D. F. Fayette: "Affordable MMICs for Air Force Systems," *Microwave J.*, 172 (May 1991).
LeComber, P. G.: "Amorphous Silicon—Electronics Into the 21st Century," *University of Wales Review*, 31 (Spring 1988).
Pool, R.: "Clusters: Strange Morsels of Matter," *Science*, 1186 (June 8, 1990).
Prinz, G. A.: "Hybrid Ferromagnetic Semiconductor Structures," *Science*, 1092 (November 23, 1990).
Ramachandran, T. B.: "Gallium Arsenide Power Sources," *Microwave J.*, 91 (January 1990).
Soref, R.: "Silicon-Based Optical-Microwave Integrated Circuits," *Microwave J.*, 230 (May 1992).
Van Zant, P.: "Microchip Fabrication," 2nd Edition, McGraw-Hill, New York, 1990.
Wiley, J. B., and R. B. Kaner: "Rapid Solid-State Precursor Synthesis of Materials," *Science*, 1093 (February 28, 1992).
Yablonovitch, E.: "The Chemistry of Solid-State Electronics," *Science*, 347 (October 20, 1989).
Yam, P.: "Atomic Turn-On: First Atom Switch," *Sci. Amer.*, 20 (November 1991).

SEMIDIAMETER CORRECTION (Sextant). When the altitude of a celestial object that is close enough to the earth to present a finite disk (e.g., moon, sun, planet) is measured with reference to the visible sea horizon or to an artificial horizon, other than that contained in the bubble sextant, it is more convenient to use the upper or lower limb (edge) of the disk than to estimate the center of the object. When solving the astronomical triangle using this measured altitude, as in determining position at sea, the position of the object given in the almanac is that of the center. Hence, to obtain the observed altitude of the center, a correction must be applied known as the correction for semidiameter. The correction is to be added or subtracted from that obtained with the sextant, depending upon whether the lower or upper limb of the object is observed. The value of the semidiameter is given in the almanac for the given date of observation. In using the moon for accurate determination of position, an additional factor, known as augmentation, must be considered due to the fact that the distance of the object from the observer varies with the altitude of the object. See also **Sextant.**

SEMIPERMEABLE MEMBRANE (or Semipermeable Diaphragm). A membrane or septum through which one (or more) of the substances composing a mixture or solution may pass, but not all.

In osmotic pressure determinations, semipermeable membranes permit the passage of a solvent but not of certain colloidal or dissolved substances. Many natural membranes are semipermeable, e.g., cell walls; other membranes may be made artificially, e.g., by precipitating copper cyanoferrate(II) in the interstices of a porous cup, the cup serving as a frame to give the membrane stability.



Separation of gases by semipermeable membrane.

Semipermeable membranes are also used in the separation of gases. (See accompanying figure.) When a semipermeable membrane is placed in a gas mixture, being impermeable to gas 2 and allowing gas 1 to pass, the force exerted on it will equal the area times the partial pressure of gas 2 only. While there are no ideal semipermeable membranes for gases, there exist in practice reasonable approximations to them, such as incandescent platinum or palladium sheets, which can be penetrated by hydrogen but not by other gases. A film of water also acts as a semipermeable membrane for gases, since it is pervious to NH_3 or SO_2 because of their solubility in water, but gases which are not easily soluble are held back.

See also **Desalination.**

SEMITONE (Half-Step). The interval between two sounds whose basic frequency ratio is approximately equal to the twelfth root of two. The interval, in equally tempered semitones, between any two frequencies, is 12 times the logarithm to the base 2 (or 39.86 times the logarithm to the base 10) of the frequency ratio.

SENSILLAE. The sense organs of insects. The term is usually applied to the integumentary sense organs of the group but it is also extended to include the scolophores on which organs of hearing and chordotonal organs are based, and the ommatidia and retinulae of the eyes.

The sensillae of other kinds include some form of cuticular structure, often a projection, associated with a nerve ending and in some cases with gland cells. These organs include some of tactile function and chemoreceptors, both of taste and of smell. In form, their external parts are classed as six types: (1) Placoid sensillae end in a thin porous plate or membrane covering a canal. (2) Trichoid sensillae end with a slender seta. (3) Basiconic sensillae have a conical protuberance. (4) Styloconic sensillae have a fixed conical base bearing subordinate projections. (5) Coeloconic sensillae end with a depression containing a conical projection. (6) Ampulliform sensillae end with a slender projection in an expanded chamber at the inner end of a long tubule.

Tactile sensillae are distributed over the entire body but are often much more abundant on the legs and sensory appendages such as antennae and palpi. Organs of smell are often abundant on the antennae and in some species appear to be limited to these appendages. There is some possibility that they may occur on other parts of the body. Organs of taste are undoubtedly associated with the mouth parts, but they are supposed to be present in aquatic insects on the outer surface of the body as well.

SENSITIVITY (Instrument). With reference to industrial and scientific instruments, the Instrument Society of America defines *sensitivity* as the ratio of a change in output magnitude to the change of input which causes it after the steady-state has been reached. Sensitivity is expressed as a ratio with the units of measurement of the two quantities stated. The ratio is constant over the range of a linear device. For a nonlinear device, the applicable input level must be stated.

Sensitivity has been used frequently to denote the *dead band*. However, its usage in this sense is deprecated since it is not in accord with accepted standard definitions of the term. See also **Dead Band.**

SENSOR (Measurement). A device that detects or senses the value or change of value of a variable being measured. The first link in the measurement-system chain. In some cases, the sensor essentially comprises the total measurement system, as in the case of a liquid-in-glass thermometer where complete measurement is accomplished by calibrating a capillary column connected directly to the temperature-sensitive thermometer bulb. In other instances, the output of the sensor may be converted from one form of energy to another (for example, mechanical motion transduced to an electrical voltage or current) and amplified and conditioned one or more times before a useful output signal for display, recording, or control is obtained. There are hundreds of different kinds of sensors to detect changes, ranging over dozens of major variables. The term *sensor* is essentially synonymous with detector or primary element.

SENSORY ORGANS. Structures in the animal body which are influenced by certain factors in the environment. Also known as receptors. The action of the environmental factor on the living substance is known as a stimulus. It results in the transmission of a nerve impulse to some nerve center and from this point may influence appropriate reactions of the animal or may be stored in memory. Special sense organs are found only in animals with nervous systems. Their development involves high specialization in some phase of the general property of living matter called irritability, and to some extent this property persists in all living tissue, whether nervous or sensory or not.

Stimuli arise from contacts with solid objects, from chemical compounds, either dissolved or in the gaseous state, from the incidence of light rays, and from factors which damage the body. It is known from various evidences that some animals perceive factors to which human organs are not sensitive, but as far as is known, the only stimulating factors are in the following groups.

Contact results in variable pressures to which organs of several kinds are sensitive. In vertebrates tactile corpuscles and other similar structures located in surface tissues may be classed as organs of touch. They are sensitive to simple pressure and give rise to images of form through the varying pressures due to uneven surfaces and gross contours. Since sound waves are due to rhythmic compression of the air, the ears and other auditory organs such as those of insects are also sensitive to pressures, but only to fluctuations of relatively high frequency (in man 30–30,000 per second). Between auditory organs and simple organs of touch are the lateral line organs of fishes and the chordotonal organs of insects, both related in some anatomical details to the auditory organs of the groups to which they belong. These organs are supposed to be sensitive to fluctuating pressures of lower than auditory frequency.

Tactile organs are also closely allied to sensory organs of insects which apparently enable them to avoid obstacles when flying. Supposedly these organs are sensitive to the changes of air pressure, often extremely delicate, resulting from approach to objects. They are located in the wings.

Dissolved substances stimulate organs of taste and gases or vapors act on organs of smell (olfactory organs). In the vertebrates the olfactory organs are associated with the nasal passages or occupy a similar position; nasal structures of fishes are limited to the olfactory function. Vertebrate organs of taste are known as taste buds and are located in the oral cavity principally, although aquatic forms may have them also in the skin. Sensory organs of this class in the invertebrates are extremely varied. Some are known as sensillae. In addition to organs of taste and smell, a general chemical sense is recognized. It is resident in various surface layers and is the least sensitive of the group. These sense organs are known collectively as chemoreceptors.

The varied integumentary sense organs of the human body are known to include some sensitive only to heat, cold, or pain. These organs may be the free nerve endings found in the skin. It has been suggested that pain may also result from overstimulation of other types of sense organs.

Sense organs that are stimulated by light are familiar to us in our own eyes. From this stage of complexity they range downward to simple light-sensitive cells. The transition includes organs capable of perceiving fluctuations of light, the direction from which it comes, and movements, as well as organs which form images in varying degrees of precision.

In contrast with sense organs of the kinds mentioned, which are classed as exteroreceptors, the body contains others called interoceptors. They are the source of sensations of hunger, thirst, nausea, and pain. Other interoceptors in the muscles, joints, and tendons are associated with the maintenance of equilibrium and are classed as proprioceptors. They are probably subject to varying pressure due to tension of muscles and shifting of the weight of the body. The semicircular canals of the inner ear of vertebrates are also organs of equilibration.

Organs of many invertebrates, such as the tentaculocysts of jellyfishes, can be interpreted only by experimental evidence. By testing the animal with different stimuli definite conclusions can often be drawn from its reactions. In most cases there is evidence of functions like those of our own sense organs.

All sense organs consist of nerve endings associated with various specialized cells or tissues. The nerves are not limited to one type of stimulation but their response may be identical under various stimuli. Thus a mechanical shock to the eye produces a sensation of light. The nerve fibers leading from the sense organ toward the central system are sensory or afferent.

SENSORY RECEPTOR. Specialized dendrites of certain neurons (sensory neurons; see **Nervous System and The Brain.**) which are sensitive to some physical state such as stretch, temperature, or chemical environment, e.g., stretch receptor, cold receptor, etc. In the case of the visual, auditory, and olfactory systems the dendrites of the sensory nerve fiber form a synapse with a specialized sensory cell which is highly sensitive to the given physical agent, e.g., rods and cones of the retina are specifically sensitive to light.

Some receptors lie within the body (*interoceptors*) and some are on the surface (*exteroceptors*). It is the function of receptors to initiate sensory impulses.

SEPIOLITE. The mineral sepiolite or meerschaum is soft, white, light in weight, and occurs in claylike nodular masses. It is a complex, hydrous magnesium silicate corresponding to the formula $Mg_4Si_6O_{15}(OH)_2 \cdot 6H_2O$. It crystallizes in the orthorhombic system; hardness, 2–2.5; specific gravity, 2; color, white, grayish white, sometimes a yellowish- or bluish-green; opaque. It is capable of floating on water, hence the name meerschaum or sea foam. It occurs in Asia Minor associated with serpentine and magnesite, and may be derived from the latter. Other deposits are in the Czech Republic and Slovakia, Morocco, and Spain; and in the United States in Pennsylvania and New Mexico. The name meerschaum is from the German. Sepiolite is from the Greek, meaning cuttlefish, referring to the similarity of the bone of that animal to the light, porous sepiolite. The material is used in the manufacture of smoking pipes.

SEPSIS. A toxic condition of the human body as caused by the spread of bacteria from a focal infection to other portions of the body. See also **Septicemia.** *Sepsis neonatorum* is an invasive bacterial infection that occurs in the first week of life, accounting for 10–20% of neonatal deaths, particularly in premature, low-birth-weight infants.

SEPTARIAN STRUCTURE. Mineralized irregular polyground joints or cracks in certain concretions. The structure resembles the pattern of cracks developed by desiccation of mud, and probably resulted from a similar cause—contraction due to desiccation of colloidal material.

SEPTICEMIA. A condition wherein pathogenic organisms circulate in the blood, causing fever and other symptoms of their presence. A number of years ago, this condition was rather aptly called "blood poisoning," a term which purveyed the grave significance of septicemia prior to the appearance of antibiotics and other effectual therapies. Septicemia persists as a serious complication of numerous acute infections, as caused by hemolytic streptococci, staphylococci, pneumococci, meningococci, color bacilli, and other pathogens. Portals of entry of these pathogens into the blood include the lungs, middle ear, the mastoid process, the skin, and the genitourinary tract. Although there are no consistent signs of septicemia, the condition is suspected when, during the course of an infection, a patient develops unusual fever, hemorrhages into the skin (purpura) or joints, symptoms of endocarditis, jaundice, and widespread abscesses. In pneumonia, meningococcus meningitis, osteomyelitis, and puerperal or post-abortion infections, the physician is on the alert for the development of the signs of septicemia. Whenever the complication is suspected, a blood culture is made immediately.

Acute septicemia is a major manifestation of meningococcal disease caused by *Neisseria meningitidis*. In nongonococcal infection of the female genital tract, anaerobic bacteria of a type found in the intestinal tract are frequently the causative agents of septicemia. These infections may involve vulvovaginal abscesses, tubo-ovarian abscesses, and pelvic peritonitis. Similar microorganisms are present following gynecologic surgery, parturition, and abortion. A dreaded complication of septic abortion is clostridial myometritis with septicemia. Septicemia is sometimes a complication which may occur in patients on hemodialysis. See **Kidney and Urinary Tract.**

Since the availability of antibiotics and other drugs, the former high mortality from septicemia has been markedly reduced.

SEPTUM. A thin wall or partition. The term is applied to the radiating plates on the foot of the coral polyp, to the transverse partitions which subdivide the body cavity of the annelid worms into chambers, and to the partitions between chambers of the shell of Nautilus, among the invertebrates. Its most familiar use among the vertebrates is to designate the nasal septum which separates the right and left nasal passages, although it applies also to the partition between the right and left chambers of the heart and to numerous other structures.

SEQUENCE. A set of quantities $s_1, s_2, \ldots, s_n, \ldots$, called elements, which can be arranged in an order so that when n is given, the nth member of the sequence s_n is completely specified. A relatively simple type of sequence is a progression. The elements are usually arranged by matching them up, one by one, with the positive integers $1, 2, 3, \ldots, n, \ldots$. A common symbol for a sequence is $\{S_n\}$. Let N be an arbitrary positive number. If it can be chosen so that $N \geq |s_n|$ for all absolute values of the members of a sequence, then the sequence is bounded. If there is at least one $|s_n| \geq N$ it is unbounded. This definition applies to an upper bound but a lower bound can be described in a similar way. A sequence is convergent if it has a limit; divergent otherwise. For example, the sequence of integers $1, 2, 3, \ldots, n, \ldots$ is divergent.

Examples of other sequences are series, finite or infinite; infinite products; continued fractions.

See also **Progression.**

SEQUENTIAL ANALYSIS. The analysis of material derived by a sequential method of sampling, that is to say, it is the data, not the analysis, which are sequential.

In sequential sampling the members are drawn one by one (or in groups) in order, and the results of the drawing at any stage decide whether sampling is to continue. The sample size is thus not fixed in advance but depends on the actual results and varies from one sample to another. The sampling terminates according to predetermined rules which are decided by the degree of precision required. See **Sampling (Statistics).**

SEQUESTERING AGENTS. See **Chelates and Chelation.**

SEQUOIA. See **Giant Sequoia; Redwood (Coast).**

SERANDITE. The mineral serandite is a hydrated manganese-sodium silicate corresponding to the formula $Mn_2NaSi_3O_8(OH)$, crystallizing in the triclinic system, of pseudo-monoclinic character. Color, rose-red, pink; transparent; brittle, and uneven fracture. Prominent basal and prismatic cleavage; vitreous to pearly luster. Crystals thick tabular or prismatic, and as intergrown aggregates. Occurs as superb crystals in a carbonatite zone in a host body of nepheline-syenite in association with analcime, aegerine, and other rare minerals at Mt. St. Hilaire, Quebec, Canada. Its only other known world occurrence is on the Island of Rouma, Los Islands, Guinea.

SERIATE FABRIC. A geological term proposed in 1906 by Cross, Iddings, Pirsson, and Washington, for the texture of an igneous rock whose granular crystals form a complete gradation in size.

SERIES. An expression of the form $a_1 + a_2 + a_3 + \cdots + a_n + \cdots$ which may have a finite or an infinite number of terms. Its partial sums constitute a sequence $\{S_n\}$, where $s_1 = a_1$, $s_2 = a_1 + a_2, \ldots, s_n = \Sigma_{k=1}^{n} a_k$. If the number of terms in its partial sums is allowed to increase without limit, either: (a) S_n approaches a limit; (b) it does not approach a limit. In the first case, if $\lim_{n\to\infty} s_n = s$, then s is the sum of the convergent infinite series and one writes $\Sigma_{k=1}^{\infty} a_k = s$. If $\lim_{n\to\infty} s_n = \pm\infty$, the series is said to be definitely divergent to $\pm\infty$. It can also be indefinitely divergent, where M and m are upper and lower limits of the sequence and the series oscillates. A simple example is the alternating series, $S_n = 1 - 1 + 1 - 1 \pm \cdots + (-1)^{n-1}$, for its sum is either unity or zero, depending on whether n is even or odd.

Convergent series are generally the most useful type for practical applications (but see **Asymptotic Series**), hence it is of great importance to test them for this property (see **Convergence**).

Algebraic combination of series should not be made carelessly for it does not follow that the usual associative, distributive, and commutative laws of algebra will always hold. One important case is the Cauchy product. If

$$S_1 = \sum_{n=0}^{\infty} a_n \quad \text{and} \quad S_2 = \sum_{n=0}^{\infty} b_n$$

are absolutely convergent to the sums A and B, respectively. then the Cauchy product,

$$S_1S_2 = \sum_{n=0}^{\infty}\sum_{i=0}^{\infty} a_i b_{n-1}$$

is absolutely convergent to the sum AB. This procedure is especially useful for power series, for if

$$y_1 = \sum_{n=0}^{\infty} a_n(x - x_0)^n \quad \text{and} \quad y_2 = \sum_{n=0}^{\infty} b_n(x - x_0)^n$$

then

$$z = y_1y_2 = \sum_{n=0}^{\infty}\sum_{i=0}^{\infty} a_i b_{n-1}(x - x_0)^n$$

Similarly, a uniformly convergent series may be either differentiated or integrated, term by term. See also **Reversion of Series.**

SERIES MOTOR. See **Motor (Electric).**

SEROUS GLAND. A gland that produces a watery secretion, in contrast with mucous glands, whose secretions are composed of or contain mucus. The term is used in connection with the salivary glands of vertebrates, which are partly serous and partly mixed. The serous gland cells are distinguished in part by their more granular cytoplasm and rounded nucleus, located near the middle of the cell.

SEROUS OTITIS MEDIA. See **Hearing and the Ear.**

SEROWS. See **Goats and Sheep.**

SERPENTINE. This is a group name for minerals encompassing two principal polymorphic forms: *chrysotile* and *antigorite*. This monoclinic mineral of hydrous magnesium silicate composition $Mg_3Si_2O_5(OH)_4$ is essentially a product of metamorphic alteration of ultrabasic rocks rich in olivine, pyroxene, and amphibole. Serpentine crystals are unknown except as pseudomorphic replacements of other minerals, e.g., after clinochlore crystals at the Tilly Foster Mine, Brewster, New York, Antigorite occurs as platy masses; *chrysotile* as silky fibers. Most massive serpentine rocks are composed essentially of antigorite. The hardness is 2–5, specific gravity ranges from 2.2 (fibrous varieties) to 2.65 (massive varieties). Color usually mottled green. The name *serpentine* stems from the mottled character, somewhat resembling the skin of a serpent. There is a greasy to wax-like luster in massive material; silky in fibrous material. The minerals are translucent.

Chrysotile fibers are the source of commercial asbestos, although fibrous amphiboles also contribute to similar usage. Asbestos is economically valuable for its incombustibility and low conductivity of heat, thus as fireproofing and insulating material. See also **Asbestos.**

Chrysotile deposits of economic value are found in Quebec, Canada, in the former U.S.S.R., and in South Africa. Minor occurrences are found in the United States in Vermont, New York, New Jersey, and Arizona. *Verd antique* marble (serpentine marble) is quarried extensively near West Rutland, Vermont.

Elmer B. Rowley, F.M.S.A., formerly Mineral Curator,
Department of Civil Engineering, Union College,
Schenectady, New York.

SERVAL. See **Cats.**

SERVICEBERRY. See **Rose Family.**

SERVICE TREE. See **Ash Trees.**

SERVOMECHANISM. A closed-loop system which depends upon the feedback concept for operation. The terms *control system, regulator,* and *servomechanism* are often used interchangeably. There are some historical differences, but the distinctions frequently are fine. In terms of current usage, the terms servomechanism and regulator are gradually phasing out of the literature in deference to control system or controller.

Conventionally, for the servomechanism, it is assumed that the output or controlled variable is forced to be a preassigned function of the reference input, where the reference input is, in general, an arbitrary function of time. A typical example is a gun fire control system. Since the target is moving, the reference input (which could be a radar output signal of target position) must be variable. If the target is to be hit, gun position (controlled variable) must be forced to a preassigned function of the reference input. This characteristic of a servomechanism is contrasted with regulator operation in which the controlled variable is maintained substantially constant. Control of gas line pressure is an example of a regulator system. The reference input (desired value of line pressure) is a constant quantity. The system works to maintain the controlled variable (line pressure) at the desired value regardless of user load requirements. From an analytical standpoint, there is very little difference between the two systems.

A servomechanism also has been defined as a feedback control system in which one or more of the system signals represent mechanical motion. Thus, a servomechanism usually is used to control an output position mechanically in response to input signal changes.

Similarly a regulator has been defined as a device that maintains a desired quantity at a predetermined value or varies it in accordance with a predetermined plan. A controller is aptly termed a regulator when it is relatively simple in design, easy to apply, relatively inexpensive, and has comparatively coarse performance characteristics. There are exceptions, of course, where manufacturers of sophisticated equipment elect to term their products regulators rather than controllers. Normally, a regulator does not incorporate means for transmitting or receiving signals from remote locations, indication, or recording. Usually a regulator is close-coupled to the equipment that it is controlling. Often, regulators are self-contained, self-actuating without requiring external energy sources. They depend upon energy derived from the equipment or system that they are regulating. A regulator in this category is logically termed a self-actuated regulator, or self-operated controller. See also **Self-Operated Controller.**

SERVOMOTORS. Electric motors are widely used in industry for effecting desired speeds and motions of machines, the positioning of parts or machines, and indirectly to control draw, thickness, stretch, and a number of other variables that must be controlled. Thus, servomotors are integral parts of automatic control and automation systems.

Servomotor and Servosystem Design Trends[1]

Whether it be *X-Y* or point-to-point positioning, or a constant or variable speed requirement, an electric motor provides precise motion control in a diverse group of products, ranging from simple conveyors to more complex machine tools and computer peripherals. The more complex systems utilize a 4-quadrant servo drive system in conjunction with the servomotor. With emphasis on increased industrial productivity and reliability, numerous advances in servotechnology have been made in the comparatively recent past. Advancements are leading to more effective use of the microprocessor in servo loop control. This article explores trends in both servomotors and servosystems and how these trends relate to applications when a manufacturing firm upgrades its facilities to reflect the philosophy of modern control and automation technology.

[1]By John Mazurkiewicz, Manager, Application Engineering, Pacific Scientific, Rockford, Illinois.

Industrial Motors—A Perspective

The direct current (DC) motor was one of the first machines devised to convert electrical energy to mechanical power. Origin of the DC motor can be traced to machines conceived and tested by Michael Faraday.

Since DC motor speeds can easily be varied, they are utilized for applications where speed control, servo control, and/or positioning tasks exist. Most small motors used in industry are alternating current (AC) motors. These motors are relatively constant-speed devices and are applied where speed *control* is not required. The speed of an AC motor is determined by the frequency of the voltage applied (and the number of magnetic poles). See also **Motor (Electric).**

Alternating Current (AC) Motors

There are basically two types of AC motors: (1) induction, and (2) synchronous. If the *induction motor* is viewed as a type of transformer, with the stator as the primary and the rotor as the secondary, it becomes easier to understand its operation. The currents which flow in the stator induce currents in the rotor, and two magnetic fields are set up. These two magnetic fields *interact* to produce *motion*. The speed of the magnetic field going around the stator will determine the speed of the rotor. The rotor will attempt to follow the stator's magnetic field, but will "slip" when a load is attached. Therefore, induction motors always rotate slower than the stator's rotating field. The *synchronous motor* is basically the same as the induction motor, but with slightly different rotor construction. The rotor is either (1) self-excited (same as induction), or (2) directly excited to set up the rotor field. The salient poles (or teeth) construction prevents slippage of the rotor field with respect to the stator field. Thus, this type of motor always rotates at the same speed (in synchronization) as the stator field. A single-phase AC motor is not self-starting. They employ a starting mechanism in order to start rotation—in the form of a start winding or a capacitor in a winding. Thus both motor types, induction and synchronous, utilize stators with rotating magnetic fields. They suffer from low starting torque, slow acceleration, and torque breakdown at overload.

Direct Current (DC) Motors

In a DC motor, the stator field can be set up by either permanent magnetics or a field winding. Thus, in contrast with the AC stator field which is rotating, the stator field is stationary. The second field, the rotor field, is set up by passing current through a commutator and into the rotor assembly. The rotor field rotates in an effort to align itself with the stator field, but at the appropriate time (due to the commutator) the rotor field is switched. Thus, by this method the rotor field never catches up to the stator field. Rotational speed depends on the strength of the rotor field, i.e., the more voltage on the rotor, the faster the rotor will turn. Thus, the DC motor is straightforward—it has predictable speed-torque characteristics, and suffers none of the speed control problems associated with AC motors.

Brushless Motors

In recent years, there has been a trend to favor *brushless* motors. The brushless motor technology emerged in the 1930s, along with vacuum tube power technology, more sophisticated control systems, and the industrial needs for velocity control and position control of basic motors. The transistor became an efficient power-handling device in the 1950s when PWM (pulse width modulation) and PFM (pulse frequency modulation) techniques became practical. With subsequent developments of transistor circuits, analog operational amplifiers, low-cost logic components, memory arrays, and microprocessor chips, control systems became oriented toward the retention and processing of information and thus able to handle more complex calculations. In the 1970s, the development of new magnetic materials provided the opportunity to explore and to design innovation in terms of pulse-modulated DC motors. It was not until this rapid expansion of modern semiconductors and new magnetic materials, along with the requirements for upgrading existing products for the automation of industry, that brushless motor development and use began in earnest.

Motivations for considering the brushless motor include the desire for improved productivity, improved product requirements in terms of higher speed, greater acceleration/deceleration, reduced maintenance, reduced size, improved power-to-weight ratio, and increased reliability.

Operating Principle of Brushless Motors

A cross-sectional view of a brushless motor is shown in Fig. 1. Brushless motors are similar to AC motors in that a moving magnet field causes rotor movement or rotation. Both motor types use stator windings and have no brushes. Brushless motors are also similar to permanent magnet (PM) DC motors since they have linear characteristics. Also, both motor types use permanent magnets to generate one field. The brushless motor is, in essence, a *hybrid*, which combines the best attributes of both the AC and DC motors.

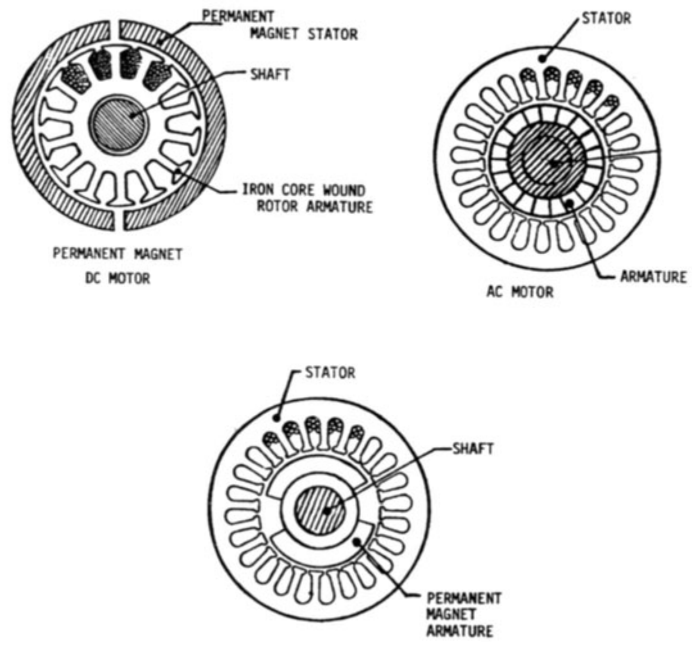
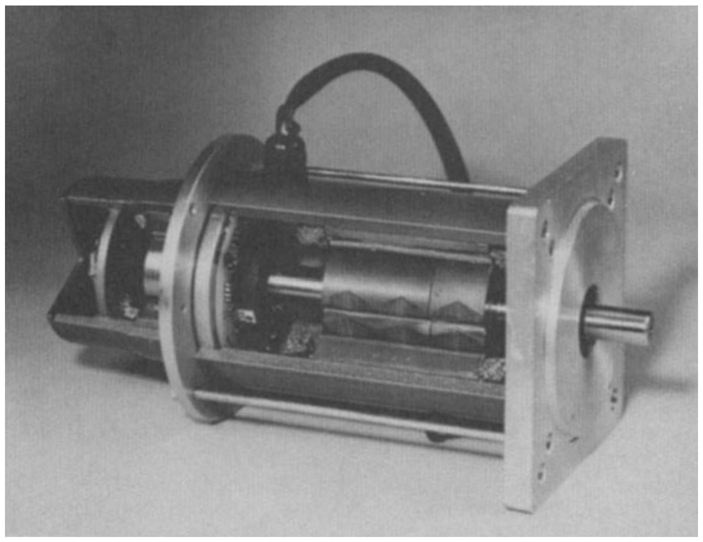
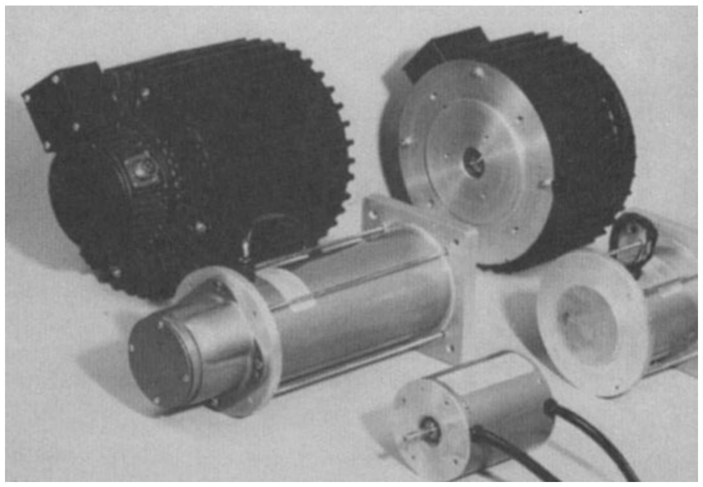

PERMANENT MAGNET STATOR
SHAFT
IRON CORE WOUND ROTOR ARMATURE
PERMANENT MAGNET DC MOTOR
STATOR
ARMATURE
AC MOTOR
STATOR
SHAFT
PERMANENT MAGNET ARMATURE

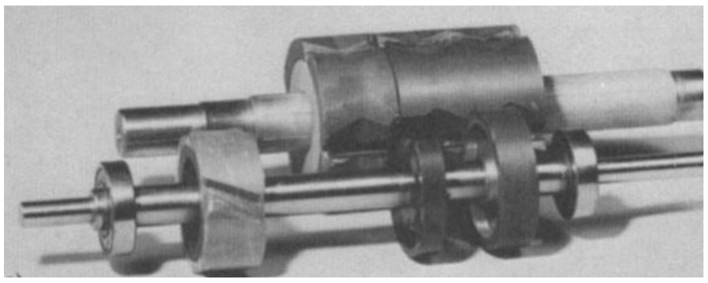
Fig. 1. Cross section of various types of electric motors used in servosystems.



(a)



(b)



(c)

Fig. 2. Brushless motors: (a) cutaway view; (b) representative family of contemporary designs; (c) brushless rotor assemblies. (*Pacific Scientific, Rockford, Illinois.*)

The configuration of the brushless motor most commonly used in contemporary systems is shown in Fig. 2(a). In this motor the rotor consists of permanent magnets and the stator consists of windings. These windings are termed "commutation" windings. By passing a current through a winding, a magnetic field is set up with which permanent magnets on the rotor interact. This results in rotation of the rotor. A representative family of modern brushless motors is shown in Fig. 2(b). Examples of brushless rotor assemblies are illustrated in Fig. 2(c).

Figure 3 illustrates, in simplified form, how rotation occurs. With a current passing through *winding 1* (Fig. 3(a)), a south pole is set up with which the permanent magnet will react and movement will begin. If, at the appropriate time, current is shut off in *winding 1* and turned on in *winding 2* (Fig. 3(b)), then the rotor will continue to move. By continuation of this timing sequence, complete rotation will occur as the rotor repeatedly tries to catch up to the stator magnetic field. In this example, the operation is simplified for explanation by exciting only one winding at a time. In practical situations, two and sometimes three windings are energized at a time. This procedure permits the development of higher torques.

As indicated, if current is properly switched from winding to winding, the rotor will continue to rotate. Switching is accomplished in conjunction with a position sensor. Frequently, solid-state Hall-effect sensors are located on the shaft assembly. These extremely rapid sensors note the position of the shaft and provide an output signal. This output signal informs the motor's electronics when to switch current from winding to winding.

In comparing two motors which develop the same torque, the brushless type has advantages. Figure 4 illustrates this point by showing a locus of safe operating areas. The maximum speed is the maximum recommended top speed of the motor, determined by (1) commutator bar-to-bar breakdown voltage in a brush type motor, and (2) by mechanical centrifuge conditions in a brushless motor. The maximum temperature limit is determined by the motor's hot armature temperature. These are quite close inasmuch as the motors develop the same approximate stall torque. Operation above this line will result in the motor's armature temperature exceeding the recommended manufacturer's maximum limit.

Control of Brushless Motors

Control of the brushless motor is accomplished by incorporating two additional output stages over the conventional servo control design.

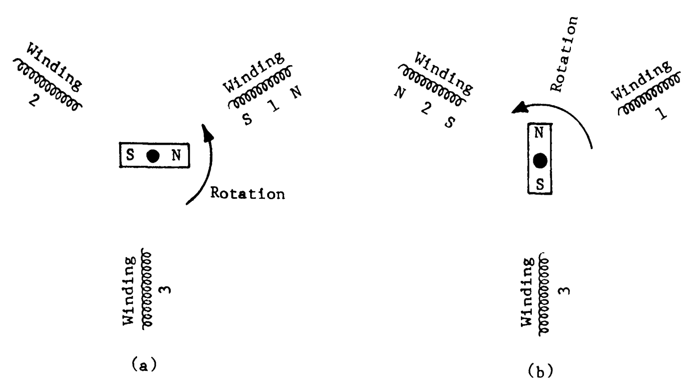


Fig. 3. Basic rotation of a brushless motor.

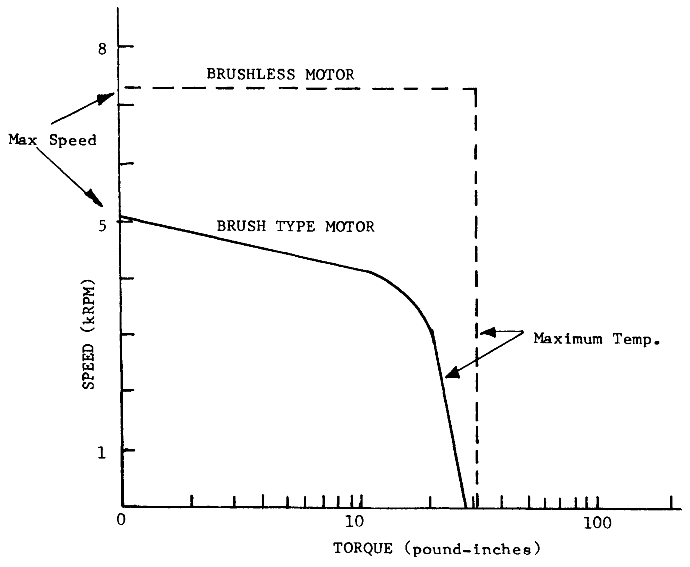


Fig. 4. Safe operating areas of two equivalent motors.

Figure 5 illustrates a simplified block diagram. *Basic operation*: Once the "run" command is given, the binary decoder compares outputs from the Hall sensors (which Halls are *on*). For example, if the shaft is sitting at "zero" (see Fig. 6), input to the logic circuit (Hall sensor output) informs the logic that Hall sensors #1 and #3 are *on*. The binary decoder interprets these data and outputs a binary code. In this example, the output code from the switch logic is a "5." This code will turn on appropriate OR gates, which turn on switches 3 and 2. The switches apply power from the voltage supply to motor windings (active legs T and R). As the motor rotates through 60°, the logic input will change state, as Hall #3 shuts off. The binary decoder output changes to a "1." This code will turn on appropriate OR gates, which turn on switches 2 and 5. (Note that during this transition, switch 2 has remained on, whereas switch 3 shuts off and switch 5 is turned on.) Again, the switches apply power to legs R and S. As the motor continues to rotate, the sequence continually progresses, changing current flow through the motor windings until complete rotation through 360° is attained, after which the sequence repeats. As current flow changes from winding to winding, the magnetic field also changes. In effect, the magnetic field is sequencing around the stator. The permanent magnet rotor will try to catch up to the field created, but never will—due to switching of the magnetic fields as a result of the Hall-effect sensor signals.

Thus, brushless rotation depends upon the stator field rotating (similar to AC motors). The significant difference is the presence of internal shaft position feedback in the brushless design. This element gives brushless motors their linear and predictable speed-torque characteristics (similar to DC motors).

Microprocessors in Servo Control Systems

Several types of controls have been developed over the years for DC motors, including SCR (silicon-controller rectifier), linear, pulse-width modulation, among others. These controls have served the needs of a diverse group of applications and generally they are basic and simple. The DC approach was chosen because no economically equivalent AC package existed. The AC motor manufacturers traditionally ignored this market. However, after the rapid escalation of energy costs in the 1970s, the motor industry perspective changed. Controls for AC motors quickly assumed a new level of sophistication. Two basic types of control emerged and are currently available: (1) the six step, and (2) pulse-width modulation. In development are more sophisticated controllers that employ SCR microprocessors and large scale integrated (LSI) dedicated electronics. In these cases, the control converts 60 Hz to direct current and then synthesizes a sine wave at a frequency to produce the desired speed.

The traditional cumbersome approach with discrete components would be unacceptable when using higher technology involving brushless designs. A search for a new approach began and, in light of the microprocessor, with its power and flexibility and its increased use in numerous applications, the concept of servosystem design was revolutionized. This approach has succeeded in providing greater flexibility in design of new systems and has potential to enhance the capabilities of existing systems. This could be termed the "intelligent" approach that utilizes microprocessor technology—an approach that impacts very favorably on design time, setup time, and implementation time.

Adjustable-Speed Brushless Control. A control of this type is self-contained, including power supply and heat sink, with the objective of driving a permanent-magnet brushless motor with Hall sensor feedback. A typical block diagram is shown in Fig. 7. The microprocessor-based control operates by energizing two of the motor's three windings, and switches power from winding to winding according to the feedback from the Hall sensors as previously described. This is coupled with pulse-width modulation drive techniques to make effective use of the output power transistors.

A user-friendly operating panel activates the control, while status lights provide easy readout. Also on the front panel is a speed adjustment pot, and a digital speed readout indicator. The tight speed regulation offered by the microprocessor approach improves system accuracy. Accuracy of ± 5 RPM over speed ranges of 100 to 10,000 RPM is possible even with dynamic load variations of 50%. The digital readout also serves as a diagnostic indicator should any of the system's protection require activation to shut down the system. This provides simple, easy-to-use user interfacing.

The Servosystem (Servocontroller) Approach. This is used with brushless technology in the same basic manner as it is employed with other prime mover (motor) technologies. Traditionally, these closed-loop servosystem designs will involve determination of load conditions and velocity profile, then prime mover (motor) selection, and determination of amplifier requirements. Following is the tedious task of compensation for gain and bandwidth adjustments for accuracy and response. The latter may involve a paperwork analysis prior to breadboarding a prototype. Or, if the individual components are purchased from independent suppliers, the "tweeking" of potentiometers would begin—in an effort to set up and adjust the compensation values. This approach can be difficult. As an example, in some servo amplifiers, there are ten potentiometers for a variety of adjustments. Throughout the compensation process, the main consideration is to set up the compensation for a given load condition. If the load changes, then the entire process must be repeated. To alleviate this problem, the system is designed to accommodate worst case conditions. This results in a system design that will be overdamped for all conditions (except the worst case), thus severely compromising system performance (speed and accuracy).

A much less cumbersome and effective approach is to utilize microprocessor techniques—an approach that uses digital control and digital filtering. The "intelligent" system compensates by simply inputting or programming the controller with parameters which describe the servo loop parameters. The algorithms precisely control servo loop velocity. Execution is under microprocessor control and all servo loop compensation (motor, load, or environmental conditions) is monitored and con-

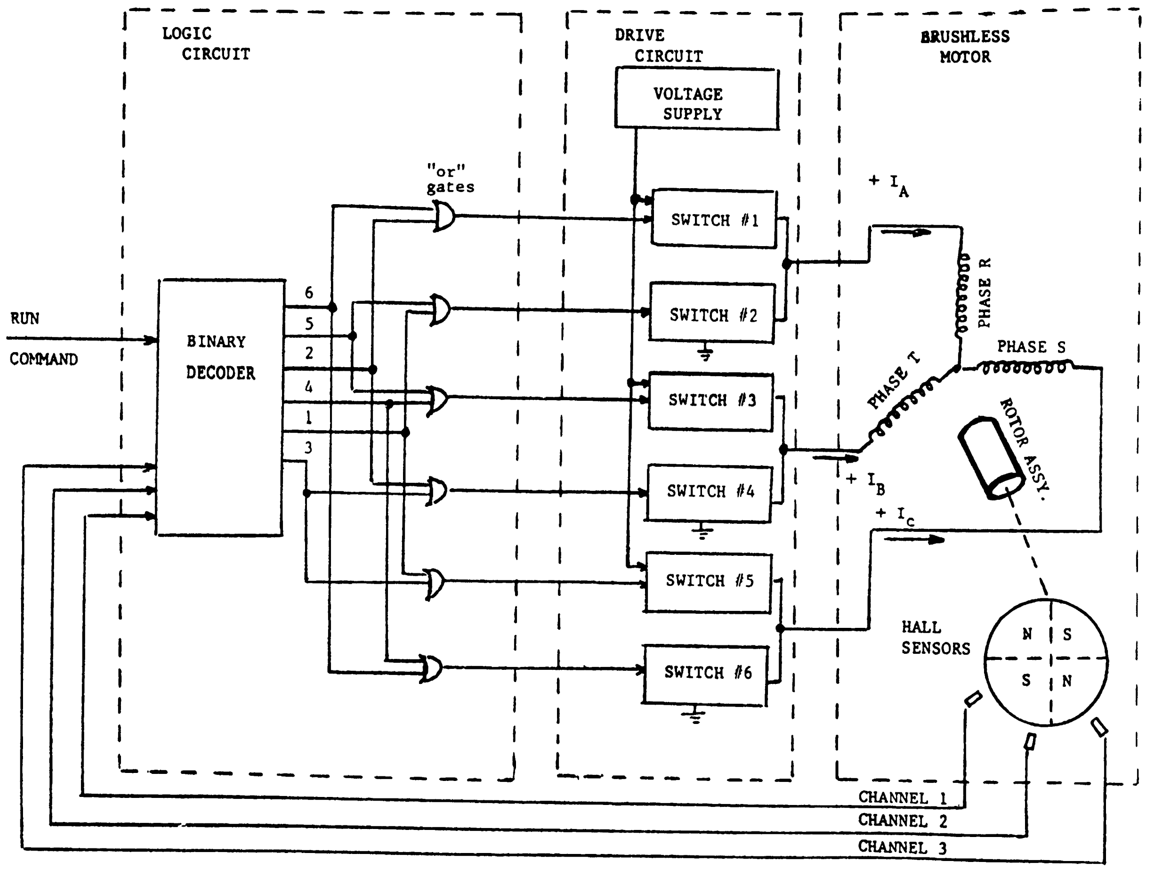
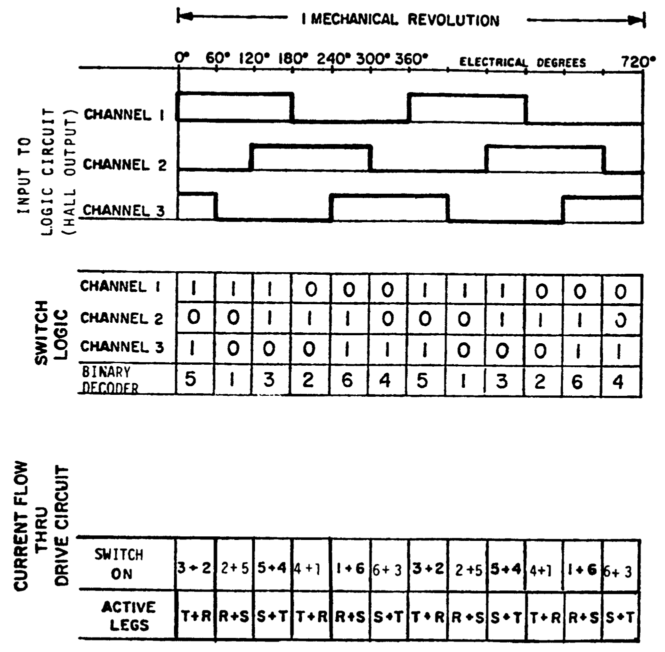


Fig. 5. Simplified block diagram of control system for brushless motor.



CHANNEL 1	1	1	1	0	0	0	1	1	1	0	0	0
CHANNEL 2	0	0	1	1	1	0	0	0	1	1	1	0
CHANNEL 3	1	0	0	0	1	1	1	0	0	0	1	1
BINARY DECODER	5	1	3	2	6	4	5	1	3	2	6	4

SWITCH ON	3+2	2+5	5+4	4+1	1+6	6+3	3+2	2+5	5+4	4+1	1+6	6+3
ACTIVE LEGS	T+R	R+S	S+T	T+R	R+S	S+T	T+R	R+S	S+T	T+R	R+S	S+T

Fig. 6. Timing diagram for clockwise rotation of brushless motor.

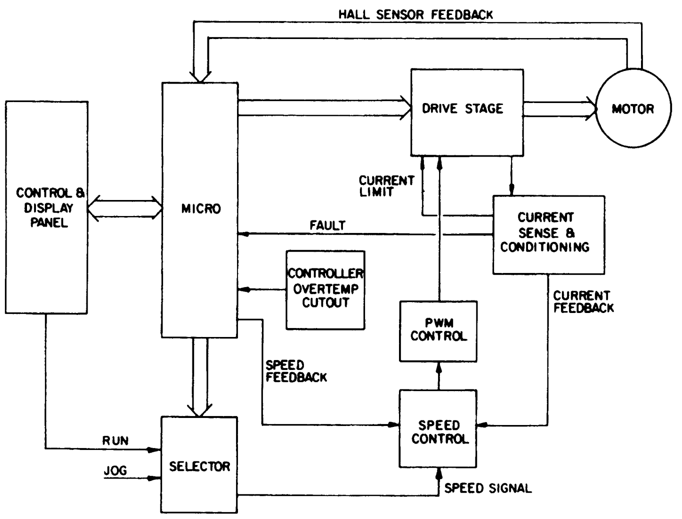


Fig. 7. An adjustable-speed brushless motor control system.

trolled should changes occur. This simplifies design and gains flexibility for the system.

One approach is the use of a general-purpose microprocessor-based control as shown in Fig. 8. The advantages of this approach include fewer components needed, which in turn lowers cost, reduces circuitry, and improves reliability. Components must be added to complete the system design, such as a drive amplifier. However, there are microprocessor-based controls, which include an internal drive amplifier, available today in production quantities to drive a brushless motor. These devices are termed "intelligent servo drives" as shown in the block diagram of Fig. 9.

These brushless servocontrols are software compensated, thus eliminating the need for pot adjustments. The microprocessor-based control can accept a variety of inputs—either analog or digital (serial or parallel) velocity commands. Parameters are factory-loaded in nonvolatile EEPROM (electrically erasable/programmable read-only memory) so

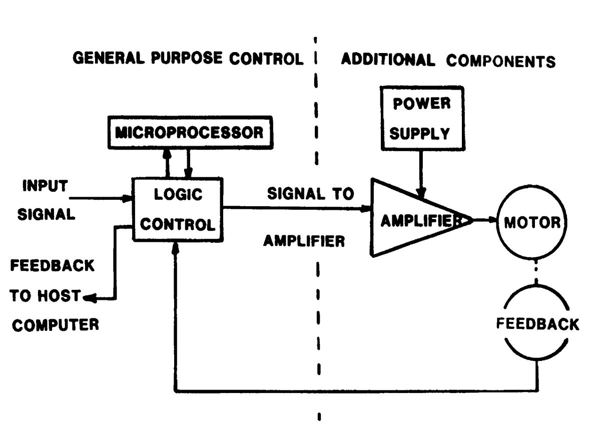


Fig. 8. A general-purpose microprocessor-based control system.

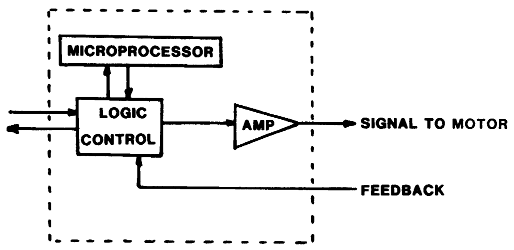


Fig. 9. Example of intelligent servo drive.

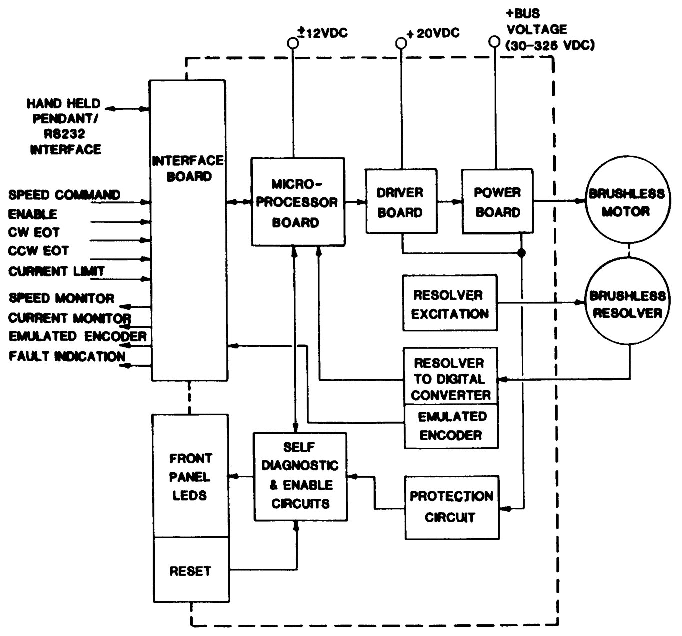


Fig. 10. Basic system operation of an intelligent brushless servo control. (CW = clockwise; CCW = counterclockwise; EDT = end of travel.) (*Pacific Scientific, Rockford, Illinois.*)

the engineer receives a stable system, thus reducing design, implementation, and setup time. Although most applications do not require further adjustments, the engineer may fine tune the system by way of a hand-held pendant, or any RS232C interface. The versatility of the microprocessor allows incorporation of self-protection, which makes the overall system virtually indestructable. Protection features include peak, RMS, and short-circuit protection, as well as thermal protection, hardware and software protection, loss of feedback and velocity error. A brushless resolver serves as the feedback device and replaces the Hall sensor feedback. This assures operation at the optimum phase angle. The entire unit controls the brushless motor by way of PWM sinusoidal driving function, thus allowing smoother operation even at very low speeds.

The "intelligent control" closes the velocity loop and directly interfaces with most readily available programmable motion controllers for position control. Basic system operation is shown in Fig. 10. The first function is to receive instruction from the motion controller and generate the speed command for the velocity loop. The algorithm looks at the *present command* of the motor, the *previous command*, and the *next instantaneous position*. This is a periodic *sample* and *compare* to a desired reference value. The difference between the speeds at two time periods serves as an indication for velocity errors and is used for velocity corrections. The feedback signals from the brushless resolver are conditioned and routed to the CPU (central processing unit) through the interrupt control. All real-time inputs have interrupt-driven software. A control manipulation is calculated and subsequently used to command the drive system. The algorithm generates a command that is the difference of a term proportional to the velocity error and a term that is proportional to the integral of the velocity error. This velocity feedforward technique, included in the microprocessor servo reference generation circuitry and applied directly to the velocity loop, allows the controller system to operate with minimal error, even during hard acceleration and deceleration. Velocity feedforwarding is valuable for maintaining wide dynamic system response. This technique stabilizes the loop and allows the system to drive the brushless motor in a smooth manner regardless of the trajectory.

The microprocessor must also determine the sign and magnitude of the current/voltage for each of the brushless motor's three phases in order to drive the motor at the appropriate torque/speed. The function of the waveform is described by a sine wave. Since the three phases are shifted by 120°, and there are four poles in the motor, or two electrical cycles per revolution, the commands to the windings are:

$$\begin{aligned}\text{Phase 1} &= \sin 2\omega t\\ \text{Phase 2} &= \sin(2\omega t + 120^\circ)\\ \text{Phase 3} &= \sin(2\omega t + 240^\circ)\end{aligned}$$

Since this calculation is accomplished only once per sample period, the appropriate weighting factor can vary considerably from the first computation time to the next. The solution is to base the weighting factors on a period base at one-half sample later. This is a velocity lead on the commutation. To optimize this, the sine functions are stored in memory. For this scheme to work, the microprocessor must "know" the relationship between the brushless motor phases and the internal drive scheme. This is accomplished by the resolver feedback. The resolver interface consists of sine and cosine reference waveforms and a phase-shifted feedback which contains absolute position data. This signal is sent directly to the microprocessor through a resolver-to-digital converter, returning a velocity per sampling period. This feedback has proven superior for improvement of dynamics and straightforwarding for controllability.

The intelligent digital microprocessor-based control system, coupled with brushless motors, constitutes an excellent solution for most applications. This combination has the ability to produce higher torque at higher speeds. The numerous specific advantages have been previously described in detail.

Solid-State Variable-Speed Drives[2]

The past two decades have seen rapid growth in the availability and usage of solid-state variable speed drives. Today, there is a profusion of types which are suitable for virtually every type of electrical machine from the subfractional to the multithousand horsepower rating. See Fig. 11. Despite the diversity, there are two common properties of these drives: (1) All of them accept commonly available AC input power of fixed voltage and frequency and through switching power conversion, create an output of suitable characteristics to operate a particular type of electric machine, i.e., they are *machine specific*. (2) All of them are based on solid-state switching devices. Even though many of the power conversion principles have been known as long as fifty years, when they

[2]By Richard H. Osman, Engineering Manager, AC Drives, Robicon Corporation (A Barber-Colman Company), Pittsburgh, Pennsylvania.

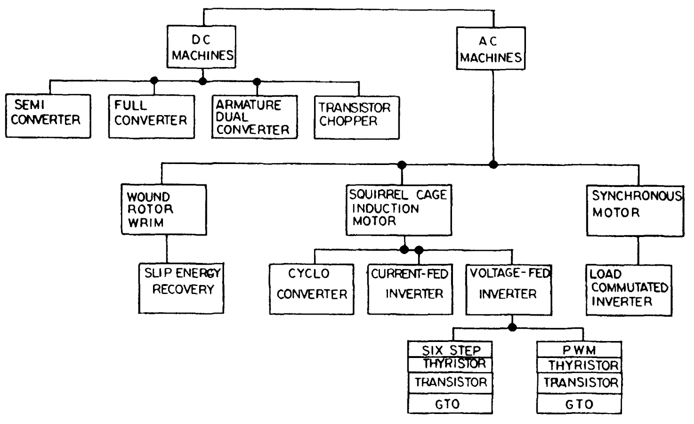


Fig. 11. General-purpose solid-state variable speed drive. (WRIM = wound rotor induction motor; GTO = gate-turn-off thyristors; PWM = pulse-width modulation.)

were developed using mercury arc rectifiers, it was not until the invention of the thyristor in 1957 that variable speed drives became practical.

The thyristor (SCR) is a four-layer semiconductor device which has many of the properties of an ideal switch. It has low leakage current in the off-state, a small voltage drop in the on-state, and takes only a small signal to initiate conduction (power gains of over 10^6 are common). When applied properly, the thyristor will last indefinitely. After its introduction, the current and voltage ratings increased rapidly. Today it has substantially higher power capability than any other solid-state device, and dominates power conversion in the medium and higher power ranges. The major drawback of the thyristor is that it cannot be turned off by a gate signal, but the anode current must be interrupted in order for it to regain the blocking state. The inconvenience of having to commutate the thyristor in its anode circuit at a rather high energy level has encouraged the development of other related devices as power switches.

Development of Power-Switching Devices

Transistors predate thyristors, but their use as high-power switches was relatively restricted (compared to thyristors) until the ratings reached 50 A and 1,000 V in the same device, since the early 1980s. These devices are three-layer semiconductors which exhibit linear behavior, but are used only in saturation or cut off. In order to reduce the base drive requirements, most transistors used in variable speed drives are Darlington types. Even so, they have higher conduction losses and greater drive power requirements than thyristors. Nevertheless, because they can be turned on or off quickly via base signals, they are attractive candidates for drives within the scope of the transistor ratings, particularly for pulse-width-modulated inverters.

Gate-Turn-Off Thyristors. More recently, successful attempts to modify thyristors to permit them to be turned off by a gate signal have been made. These devices are four-layer types and are called *gate-turn-off thyristors*, or simply GTO's. Power GTO's have been around since at least 1965, but only relatively recently (1981) have devices rated more than a few tens of amps become available. Present GTO's have about the same forward drop as a Darlington transistor (twice that of a conventional thyristor). GTO's require a much more powerful gate drive, particularly for turn-off, but the lack of external commutation circuit requirements makes them desirable for inverter use. GTO's are available at higher voltage and current ratings than power transistors. Unlike transistors, once a GTO has been turned on or off with a gate pulse, it is not necessary to continue the gate signal due to the internal positive feedback mechanism inherent in four-layer devices.

Technological Base. The three aforementioned devices (thyristor, transistor, and GTO) form the technological base on which the solid-state variable-speed drive industry rests today. There are other devices in various stages of development which may or may not become significant depending on their cost and availability in large current ($>$ 50 A) and high voltage (1,000 V) ratings. These include: (1) the metal oxide semiconductor field effect transistor (MOSFET); (2) the insulated gate transistor; and (3) the static induction thyristor.

It has not yet been possible to construct *power MOSFET's* which have acceptably low ON resistance while still having the 1,000 V rating necessary for reliable power conversion at the 500 VAC level. Therefore, their use has been limited to small drives (under 10 HP). Power MOSFET's are the fastest power switching devices (100 ns) of the lot, and they also have very high gate impedance, thus greatly reducing the cost of drive circuits.

The *insulated gate transistor* is a combination of a power bipolar transistor and a MOSFET which combines the best properties of both devices. A most attractive feature is the very high input impedance which permits them to be driven directly from lower power logic sources. Unfortunately, their ratings are not very impressive, being limited to a few tens of amperes and about 600 volts.

Static induction thyristors (SR's) are claimed to have the voltage and current ratings of GTO's, but with a much higher gate impedance so as to reduce driver requirements. The validity of these claims has not been proved in commercial use because SR's are just emerging from the development laboratory.

Variable-Speed Drive Hardware Development

Parallel to the development of power switching devices, there have been very significant advances in hardware for controlling variable speed drives. These controls are a mixture of analog and digital signal processing.

The advent of integrated circuit operational amplifiers and integrated circuit logic families made possible dramatic reductions in the size and cost of the drive control, while permitting more sophisticated and complex control algorithms without a reliability penalty. These developments occurred during the 1965–1975 period. Further consolidation of the control circuits occurred after that as large scale integrated circuits (LSI) became available. In fact, the pulse-width modulation (PMW) control technique was not practical until the appearance of LSI circuits because of the immense amount of combinational logic required. A significant trend has been the introduction of *microprocessors* into drive control circuits. While it is doubtful whether microprocessors will significantly reduce control circuit cost, there is general agreement that they are greatly expanding the capability of drive controls. The performance enhancements include: (1) more elaborate and detailed *diagnostics* due to the ability to store data relating to drive internal variables, such as current, speed, firing angle, and so on; (2) the ability to *communicate both ways* with user's central computers about drive status; (3) the ability to make *drive tuning adjustments* via keypads with parameters such as loop gains, ramp rates, current limit stored in memory rather than pot settings; (4) *self-tuning* drive controls; and (5) more adept techniques to overcome *power circuit nonlinearities*. The possibilities are large and are just beginning to become commercial realities.

DC Drives

The introduction of the thyristor had the most immediate impact in the DC drive area. Ward-Leonard (motor-generator) variable speed drives were quickly supplanted by thyristor DC drives of the type shown in Fig. 12 for reasons of lower cost, higher efficiency, and lower maintenance cost. This type of power circuit with three thyristors and three diodes in a three-phase bridge is generally referred to as a *semiconverter*. By phase control of the thyristors, it behaves as a programmable voltage source. Therefore, speed variation is obtained by adjustment of

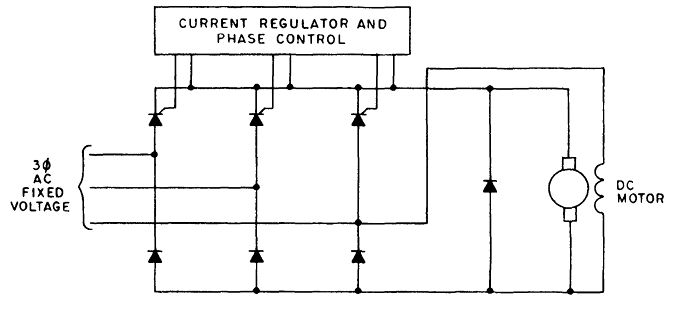


Fig. 12. Thyristor DC drive-3-phase semiconverter.

the armature voltage of the DC machine. Because the phase control is fast and precise, critical features like current limit are easily obtained. In fact, almost all thyristor DC drives today are configured as current regulators with a speed or voltage outer loop. The semiconverter is suitable for one-quadrant drives as it produces only one direction of current and output voltage. Input power factor is dependent on speed.

Six-Thyristor Full Converter. As the cost differential between thyristors and diodes narrowed, the semiconverter was largely displaced by the six-thyristor full converter as shown in Fig. 13. This circuit arrangement (the Graetz circuit) has become the workhorse of the electrical variable speed drive industry as will be pointed out. The control techniques are very much the same as in the semiconverter. However, the full converter offers lower output ripple and the ability to regenerate, or return energy to the AC line. The system can be made into a four-quadrant drive by the addition of a bidirectional field controller. Torque direction is determined by field current direction. Due to the large field inductance, torque reversals are fairly slow (100–500 ms) but adequate for many applications.

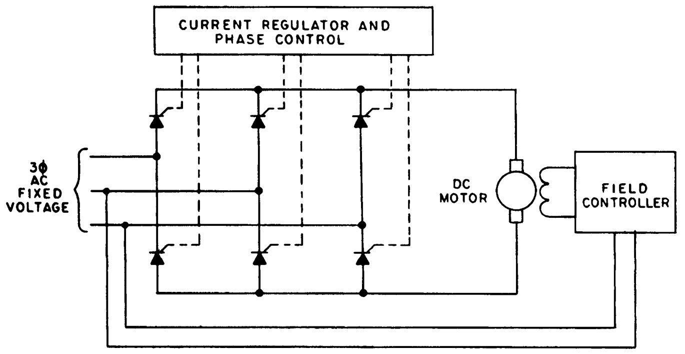


Fig. 13. Thyristor DC drive-3 phase full converter.

Dual Armature Converter. For the best response of thyristor DC drives, the dual armature converter of Fig. 14 is preferred. This is simply two converters (as shown in Fig. 13) connected back to back. Torque direction is determined by the direction of armature current, and since this is a low-inductance circuit, reversal can be accomplished in 10 ms (typically). Obviously, only one converter is conducting at one time with the other group of six thyristors not being gated. This is called "*bank selection.*"

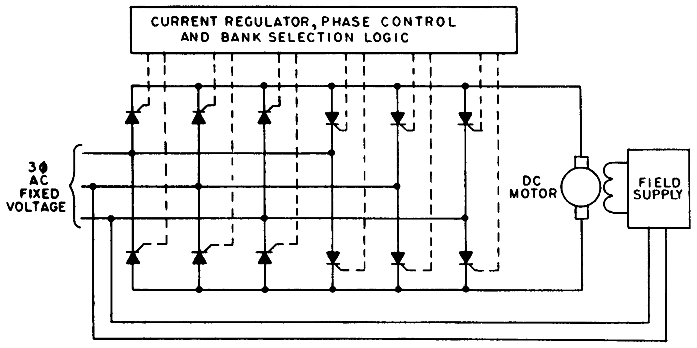


Fig. 14. Thyristor DC drive-armature dual converter.

Summary of Thyristor DC Drives. The three types of thyristor DC drives just described all share a common property in that the devices are turned off by the natural polarity reversal of the input line. This is called *natural* or *line commutation*. Thus, the inability to turn off a thyristor from the gate is no practical drawback in these circuits. Consequently, they are simple and very efficient (typically 98.5%) because the device forward drop is small compared to the operating voltage. These drives can be manufactured to match a DC machine of any voltage (commonly 500 V) or horsepower (0.5 to 2,500 HP, typically).

For certain types of applications, typically machine tool axis drives and tape transport drives, the response of phase controlled thyristor drives is not fast enough. A special class of DC drives has been developed. See Fig. 15. A fixed DC bus is developed from the line via a rectifier and capacitor filter. This voltage source is applied to the armature through power transistors. The voltage is modulated by duty cycle (or pulse width) control. The devices usually operate in the 1–5 kHz range. These specialty drives usually operate from 120 or 240 VAC and rarely exceed 10 HP. Frequently, they are applied with permanent magnet field DC machines.

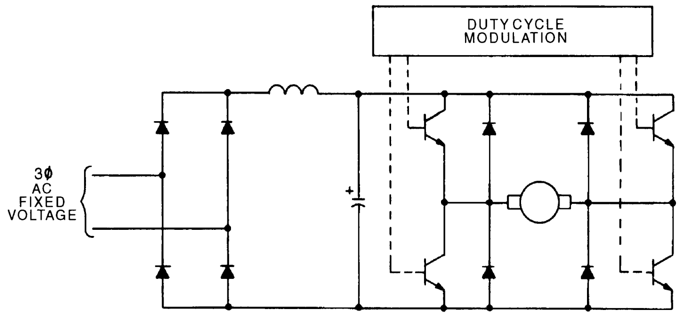


Fig. 15. Chopper DC drive-transistor bridge type.

AC Drives

The impact of the new solid-state switching devices was even larger on the AC variable speed drives, but it occurred somewhat later in time as compared to DC drives. AC drives are machine specific and more complex than DC drives. Solid-state variable-speed drives have been developed and marketed for wound-rotor induction motors (WRIM's), synchronous motors, and cage-type induction motors.

Historically, WRIM-based variable speed drives were commonly in use long before solid-state electronics. These drives operate on the principle of deliberately creating high-slip conditions in the machine and then disposing of the large rotor power which results. This is done by varying the resistance seen by the rotor windings.

Slip-Energy Recovery System. A more modern WRIM drive is shown in Fig. 16. This is called the *slip-energy recovery system* or *static Kramer drive*. The output of the rotor is rectified, and this DC voltage is coupled to the line via a thyristor converter. The line commutated converter is current regulated which effectively controls torque. Efficient, stepless speed control results. Very large (> 1,000 HP) drives can be built, as the stator may be wound for medium voltage, while the rotor operates at 400–400 V maximum. The power conversion equipment may be downsized if a narrow range (100% to 70%) of control is adequate, for example in fan drives. The performance drawbacks are a poor system power factor if not corrected; and no above-synchronous-speed operation.

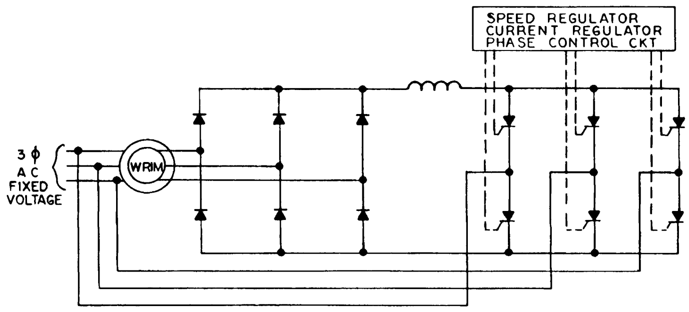


Fig. 16. Slip energy recovery system-wound rotor inductor motor. (WRIM = wound rotor induction motor; CKT = circuit.)

The WRIM is the most expensive AC machine. This has made WRIM-based variable speed drives noncompetitive as compared to cage induction motor (IM) drives or load commutated inverters using synchronous machines. It appears that the WRIM will become a casualty of the tremendous progress in AC variable speed drives as applied to cage induction motors.

Load Commutated Inverter. As shown in Fig. 17, the load commutated inverter is based on a synchronous machine. It uses two thyristor bridges, one on the line side and one on the machine side. All devices are naturally commutated, because the back EMF of the machine commutates the load side converter. This requires the machine to operate with a leading power factor and, therefore, it requires substantially more field excitation and a special exciter compared to a normally applied synchronous motor. This also results in a reduction in the torque for a given current. The machine side devices are fired in exact synchronism with the rotation of the machine, so as to maintain constant torque angle and constant commutation margin. This is done either by rotor position feedback, or by phase control circuits driven by the machine terminal voltage. The line side converter is current regulated to control torque. A choke is used between converters to smooth the link current. Load commutated inverters (LCI's) came into commercial use about 1980, and are used mainly on very large drives (1,000–10,000 HP). At these power levels, multiple series devices are employed (typically 2.4 to 4 kV input) and conversion takes place directly at 2.4 or 4 kV or higher. The efficiency is excellent and reliability has been very good. Although they are capable of regeneration, LCI's are rarely used in 4-quadrant applications because of the difficulty in commutating at very low speeds where the machine voltage is negligible. Operation above line frequency is straightforward.

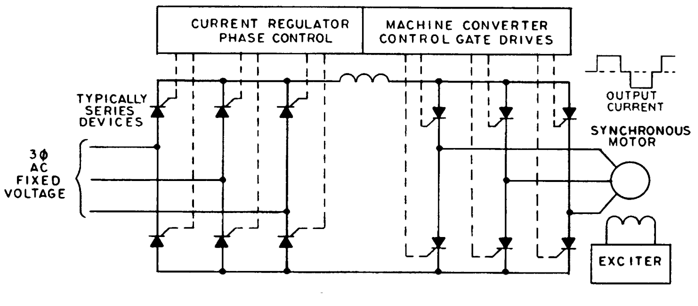


Fig. 17. Load-commutated inverter (LCI).

Induction Motor Variable Speed Drives

Induction motor variable speed drives have the greatest diversity of power circuits. (See Fig. 11.) Because the squirrel cage induction motor is the least expensive, least complex, and most rugged electric machine, great effort has gone into drive development to exploit the machine's superior qualities. Due to its very simplicity, it is the least amenable to variable speed operation. Since it has only one electrical input port, the drive must control flux and torque simultaneously through this single input. As there is no access to the rotor, the power dissipation there raises its temperature—so very low-slip operation is essential.

Cycloconverter. One approach in an IM drive is to "synthesize" an AC voltage waveform from sections of the input voltage. This can be done with three dual converters and the circuit is called a *cycloconverter*. See Fig. 18. The output voltage is rich in harmonics, but of sufficient quality for IM drives as long as the output frequency does not exceed $\frac{1}{3}$ to $\frac{1}{2}$ of the input. The thyristors are line commutated, but there are 36 of them. (Sometimes half-wave circuits are used which need only 18 devices, but more harmonics result.) The cycloconverter is capable of heavy overloads and 4-quadrant operation, but it has a limited output frequency and poor input power factor. For special low-speed high horsepower (> 1,000) applications, such as cement-kiln drives, the cycloconverter has been used.

Autosequentially Commutated Current-Fed Inverter. Still another approach to an IM drive is to generate a smooth DC current and

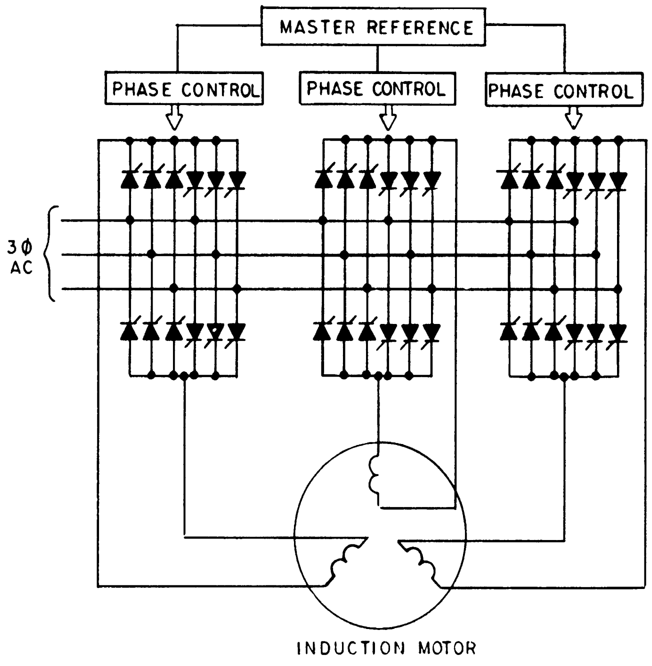


Fig. 18. Cycloconverter induction motor drive.

feed that into different parts of windings of the machine so as to create a discretely rotating magnetomotive force (MMF). This type of inverter is called the *autosequentially commutated current fed-inverter* (ASCI). See Fig. 19. This circuit was invented later than other inverters and is much more popular in Europe and Japan than in the United States. Once again, the input stage is a three-phase thyristor bridge which is current regulated. A link choke smooths the current going to the output stage. There, a thyristor bridge distributes the current into the motor windings with the same switching function as the input bridge, except at variable frequency. (Notice the similarity to the LCI.) The current waveform is a quasi-square wave whose frequency is set by the output switching rate and whose amplitude is controlled by the current regulator. The capacitors and rectifiers are used to store energy to commutate the thyristors, since the induction motor cannot provide this energy and remain magnetized, in contrast to the synchronous motor. This type of drive has simplicity, good efficiency (95%), excellent reliability, and four-quadrant operation up to about 120 Hz. Harmonics in the output current are reasonably low, giving a form factor of 1.05 (same as the LCI).

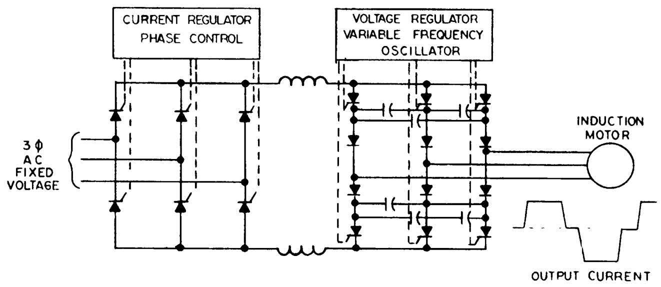

CURRENT REGULATOR PHASE CONTROL
VOLTAGE REGULATOR VARIABLE FREQUENCY OSCILLATOR
3ϕ AC FIXED VOLTAGE
INDUCTION MOTOR
OUTPUT CURRENT

Fig. 19. Autosequentially commutated current-fed inverter (ASCI).

Moreover, harmonic currents are not machine dependent and decrease at light load. The input power factor is load and speed dependent, but much better than the cycloconverter. Above 100 HP, the ASCI is very cost effective. Because they are constructed with SCR's, they have recently (1984) become available at 2.4 and 4 kV direct conversion for very large (> 1,000 HP) units. It is almost always possible to retrofit an existing motor with this type of drive. Due to the controlled current properties, this drive is virtually immune to damage from ground faults, load shorts and commutation failures.

Since MMF (current) is directly controlled and the drive is regenerative, ASCI's can readily be equipped with field oriented controls for the most demanding four-quadrant operation.

Field-Oriented Controls. Special mention should be made of AC induction motor drives with field oriented controls. They are the state of the art. The control technique keeps track of the flux and MMF vectors inside the machine in order to provide a fast and precise torque response to an external reference. In addition, they are 4-quadrant drives, capable of producing either direction of torque in either direction of rotation. They are the AC drives of choice for the most demanding applications such as traction drives and machine tool axis drives.

Voltage-Source Inverter. The third approach to IM drives is to generate a smooth DC voltage and apply that to different combinations of the machine windings so as to create a rotating flux. This circuit is called a *voltage-source inverter* (VSI). An implementation using thyristors is shown in Fig. 20. This circuit was the first application of thyristors to IM drives. The input is a 3-phase thyristor bridge which feeds a capacitor filter bank forming a controlled low-impedance voltage source. The output stage consists of six main thyristors (1–6) in a bridge with antiparallel diodes. There are six auxiliary or commutating thyristors (11–16) which together with the *L-C* circuits, impulse commutate the main devices. The output waveform is a quasi-square wave of voltage whose amplitude is set by the DC link voltage. Here the output frequency is determined by the output switching rate, and the output voltage is set by the voltage regulator on the input converter. In order to reduce the size of the commutating *L-C*, special thyristors with fast turn-off times are required—in contrast to the ordinary phase control types used in DC drives, LCI's, and ASCI's. Despite the complexity, these drives have had a reasonably good reliability record. They have good efficiency (typically 95% at full speed, full load), speed dependent power factor, and can operate at very high frequencies (180 Hz and up). Regeneration to the line is not possible. Many of these units are in service, and they are still available from 50 to 55 HP, typically at 460 VAC. They are not available at over 600 V.

Transistors and GTO's in Voltage-Source Inverter. In order to reduce the cost and complexity of the VSI shown in Fig. 20, the thyristors and their commutation circuits have been replaced with transistors or with GTO's. The resulting circuit is shown in Fig. 21. The performance features are about the same as the thyristor version, but size and weight are substantially reduced. Although the conduction losses are higher due to the higher "on" voltages, commutation losses are reduced substantially—so efficiencies remain in the 95% range. The transistor version has been available since about 1982 at 460 VAC input; 230-volt units have been available since the mid-1970s. Presently, 100 HP transistor drives with single output devices are available; up to 300 HP with parallel output devices can be obtained.

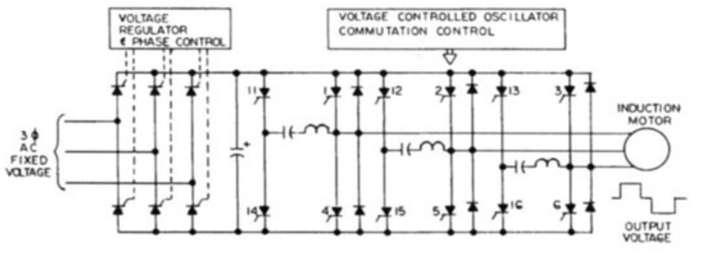

VOLTAGE REGULATOR & PHASE CONTROL
VOLTAGE CONTROLLED OSCILLATOR COMMUTATION CONTROL
3φ AC FIXED VOLTAGE
11
1
12
2
13
3
14
4
15
5
16
6
INDUCTION MOTOR
OUTPUT VOLTAGE

Fig. 20. Voltage-source inverter; six-step thyristor, impulse commutated.

Since GTO's have somewhat higher ratings, drives from 50 to 500 HP at 460 VAC are now available. It is difficult to forecast which device will be more successful, but transistors are much more widely used up to 100 HP than GTO's. There will be a three-way competition in the future among GTO, transistor, and thyristor drives in the 100–500 HP range to capture the bulk of the market. GTO and transistor costs will have to be substantially reduced to challenge thyristor designs at the upper end of this range.

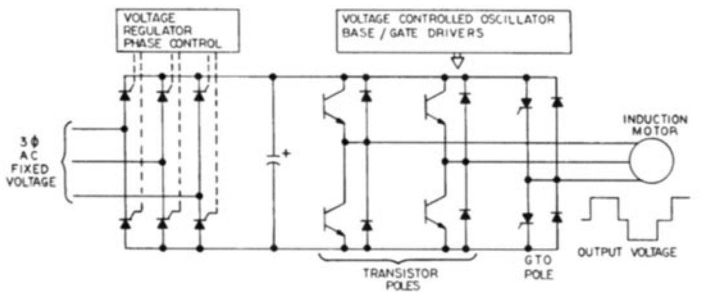

VOLTAGE REGULATOR PHASE CONTROL
VOLTAGE CONTROLLED OSCILLATOR BASE / GATE DRIVERS
3φ AC FIXED VOLTAGE
INDUCTION MOTOR
OUTPUT VOLTAGE
TRANSISTOR POLES
GTO POLE

Fig. 21. Voltage-source inverter, six-step transistor or gage-turn-off thyristor (GTO).

Pulse-Width Modulated (PWM) Drives. The induction motor drives discussed thus far are all similar in that the amplitude of the output is controlled by the input converter. Another category of voltage source inverters controls both the frequency and amplitude by the output switches alone. A representative circuit based on transistors is shown in Fig. 22. Note that the input converter is replaced with a diode bridge so that the DC link operates at a fixed, unregulated voltage. The diode front end gives virtually unity power factor, inde-

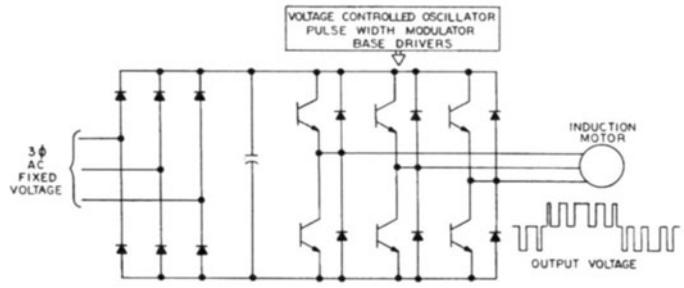

VOLTAGE CONTROLLED OSCILLATOR PULSE WIDTH MODULATOR BASE DRIVERS
3φ AC FIXED VOLTAGE
INDUCTION MOTOR
OUTPUT VOLTAGE

Fig. 22. Pulse-width modulated (PWM) inverter-transistor implementation.

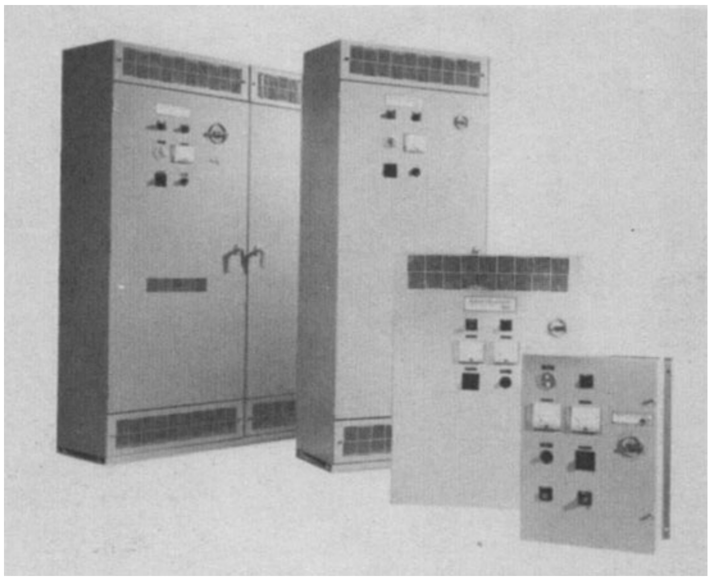


Fig. 23. AC variable-frequency drives ranging from 5 to 1000 HP. (*Robicon Corporation, Pittsburgh, Pennsylvania.*)

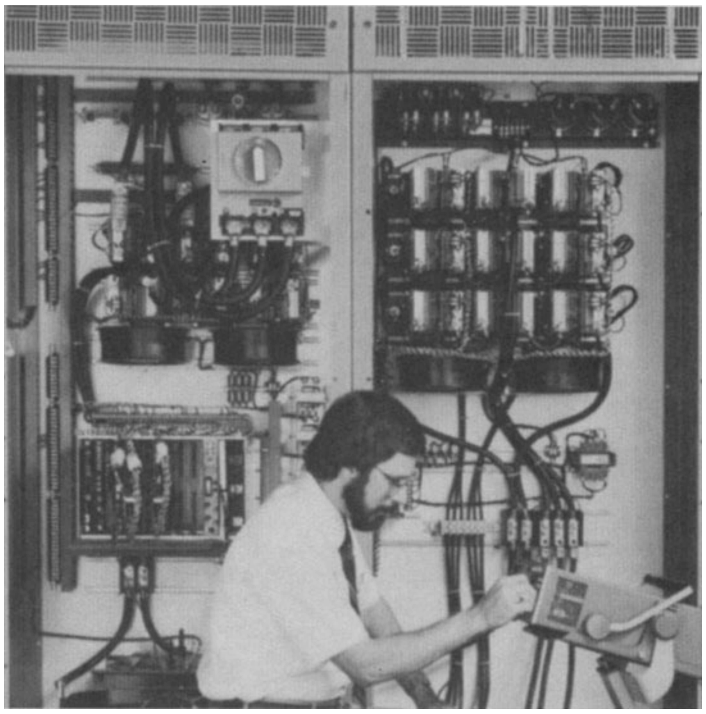


Fig. 24. Representative 800HP, 480V variable-speed frequency drive. (*Robicon Corporation, Pittsburgh, Pennsylvania.*)

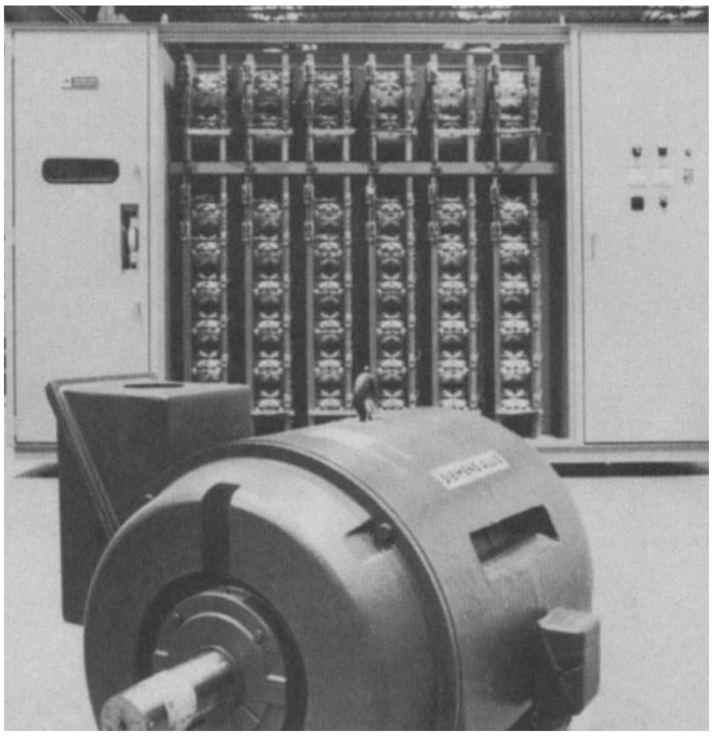


Fig. 25. Representative 2500HP, AC variable-frequency drive. (*Robicon Corporation, Pittsburgh, Pennsylvania.*)

pendent of load and speed. This type of drive is called *pulse-width-modulated* (PWM).

An output voltage waveform is synthesized from constant amplitude, variable-width pulses at a high (1–3 kHz) frequency so that a sinusoidal output is simulated; the lower harmonics (5,7,11,13,17,19, . . .) in six-step waveforms are not present in sophisticated PWM's. One advantage is smooth torque, low harmonic currents, and no cogging. Some PWM designs merely encode the six-step square wave, but this results in having the worst features of both PWM and six-step designs. Although this approach eliminates the phase control requirements and cuts the front end losses somewhat, there are offsetting drawbacks. Since every switching causes an energy loss in the output devices and their suppressors, the total losses at high speed go up considerably over six-step (six switches per cycle) if the same devices are used. To overcome this, many versions revert to six-step at 60 Hz. The output devices are stressed much more severely than in six-step. Many PWM designs do not have a voltage regulator; at any given output frequency, they deliver a preset fraction of the input voltage. If the input fluctuates, so does the output. Finally, the high-frequency switching may cause objectionable acoustic noises in the motor. There are both transistor and GTO PWM units on the market today in the range of 1–3,000 HP at 460 VAC. The transistor versions have a better reliability record. As with all voltage source inverters, regeneration to the line is not inherent.

Views of representative industrial solid-state variable speed drives are given in Figs. 23, 24, and 25. See also **Stepper, Linear, and Planar Motors.**

Additional Reading

Bailey, S.J.: "Servo Design Today: Hardware Elements Fade as Software Closes Feedback Loop," *Cont. Eng.*, 57–61, February 1985.

Bailey, S. J.: "Servomotor and Stepper: Key Elements of Motion Control," *Cont. Eng.*, 55–59 (February 1986).

Bailey, S.J.: "AC Motor Drives Selection," *Cont. Eng.*, 101–105 (April 1986).

Bailey, S. J.: "Servo Vs Stepper—Motion Control Design Decisions with Dynamic Overtones," *Cont. Eng.*, 68–72 (May 1987).

Bailey, S. J.: "Lessening the Gap Between Incremental and Continuous Motion Control," *Cont. Eng.*, 72–76 (February 1987).

Bailey, S. J.: "AC Motor Drives Use Microprocessors to Set Top Specs for Motion Control," *Cont. Eng.*, 98–102 (April 1987).

Bedford, B. D., and R. G. Hoft: "Principles of Inverter Circuits," Wiley, New York, 1964.

Bose, B. K.: "Adjustable Speed AC Drive Systems," Wiley, New York, 1981.

Brichant, F.: "Force-Commutated Inverters," Macmillan, New York, 1984.

Ghandi, S. K.: "Semiconductor Power Devices," Wiley, New York, 1977.

Kosow, I. L.: "Control of Electric Machines," Prentice-Hall, Englewood Cliffs, New Jersey, 1973.

Mazurkiewicz, J.: "Brushless Motors Coming on Strong," *Electronic Products*, 61–75, September 1984.

Mazurkiewicz, J.: "Advances in Microprocessor Control for Brushless Motors," *Electronic Motion and Control Association* Conf., San Diego, Calif., November 1984.

Murphy, H.: "Star-Modulated, Variable Frequency AC Drives Bring Higher Performance," *Cont. Eng.*, 104–108 (April 1987).

Pelly, B. R.: "Thyristor Phase-Controlled Converters and Cycloconverters," Wiley, New York, 1971.

Schaefer, J.: "Rectifier Circuits: Theory and Design," Wiley, New York, 1965.

Scoles, G. J.: "Handbook of Rectifier Circuits," Wiley, New York, 1980.

Sen, P. C.: "Thyristor DC Drives," Wiley, New York, 1981.

SESAME SEED OIL. See **Vegetable Oils (Edible).**

SET (Mathematics). A collection of numbers or symbols considered as a whole. For example, the set of all prime numbers or the set of all matrices with determinant equal to unity. Geometrically, the symbols in a set determine a *domain*.

SET (Permanent). When a solid has been strained beyond the elastic limit and the deforming stress is completely removed, in general the strain does not decrease ultimately to zero but to some nonvanishing value, known as a permanent set.

SETTLING TIME. The time required, following the initiation of a specified stimulus to a system, for the output to enter and remain within a specified narrow band centered on its steady-state value. See diagram that accompanies entry on **Response (Instrument).** The stimulus may be a step impulse, ramp, parabola, or sinusoid. For a step or impulse, the band is often specified as $\pm 2\%$. For nonlinear behavior, both magnitude and pattern of the stimulus should be specified.

SEWELLEL. See **Rodentia.**

SEX. The state of an individual as determined by its adaptations for a special part in biparental reproduction and modifications of the process. Also a category of individuals adapted for a special part in reproduction. The usual sexes are male and female. Neuter individuals exist among colonial invertebrates, including both sexless forms and abortive females whose limited reproductive powers are not exercised under normal conditions. Sex is also expressed in hermaphrodite animals which carry on the usual processes of sexual reproduction but have male and female organs in the same individual.

The differentiation of the sexes is associated with the development of two kinds of gametes in the process of sexual reproduction. Organs and ducts capable of producing the larger egg cells of the female and providing them with quantities of food material make up a reproductive system much different from that which produces the minute male spermatozoa and the seminal fluid in which they are discharged. The development of the external genitalia for internal insemination also results in conspicuous differences since the male has a projecting penis or other intromittent organ while the female has the terminal portion of the genital ducts specialized for the reception of this organ. In the mammals the mammary glands of the female also constitute a conspicuous sexual distinction. These organs may be classed as essential and accessory organs of reproduction, the former category including the gonads and ducts and the latter such parts as the external genitalia and the mammary glands.

Sexes also differ in more or less conspicuous secondary sexual characters such as the beard of man, which are definitely associated with sex but have no direct part in reproduction.

The sexes of most animals are specialized in behavior as well as in structure for the performance of reproductive acts, for accessory functions such as the building of nests, and for subsequent duties of parental care. All of these phases of sexual differentiation are intricately variable among the many species of animals.

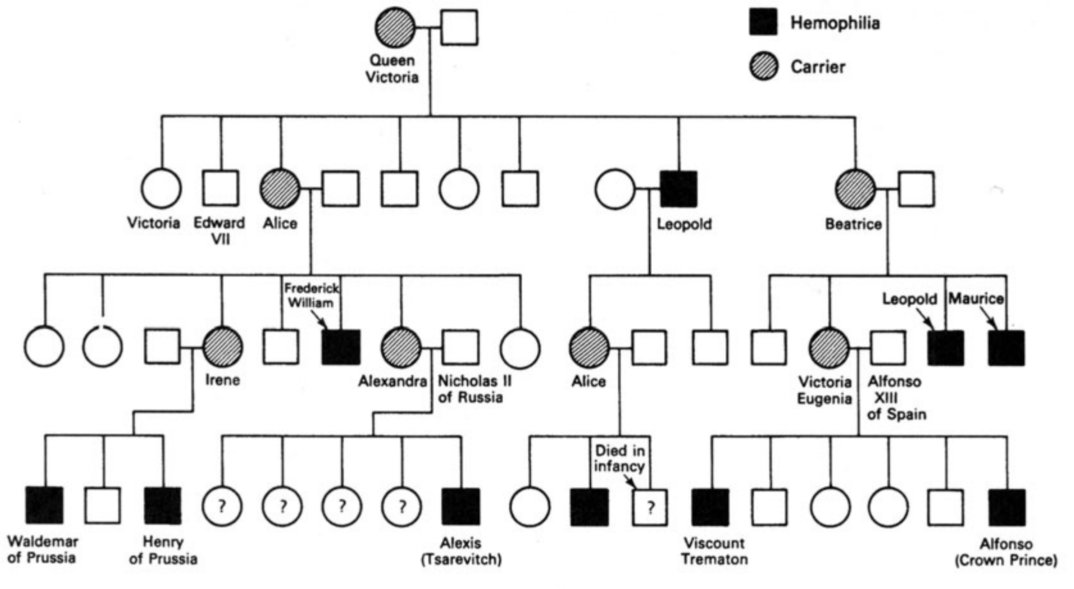
SEX (Flower). See **Flower.**

SEX-INFLUENCED INHERITANCE. Inheritance which is influenced by sex, but not restricted to one sex or the other. Baldness in man is often given as a possible case of sex-influenced inheritance. Baldness is supposed to be dominant in men and recessive in women. Hence, a man will be bald if he carries only one gene for baldness, but a woman must carry two such genes before she will be bald. The horns of sheep are also good examples. The horned condition is due to a gene which is dominant in the males and recessive in the females. The Dorset sheep are also homozygous for the gene for horns, so both sexes have horns. The Suffolk sheep are homozygous for the gene for hornless. When the two breeds are crossed, the male offspring all have horns and the females are all hornless. When these first generation offspring are crossed, the offspring show a ratio of 3 horned to 1 hornless in the males and 1 horned and 3 hornless in the females. This is to be expected on the basis of the method of inheritance of a sex-influenced gene.

SEX-LIMITED INHERITANCE. A condition where certain genes are expressed in only one sex, although they may be carried by both sexes. The human beard is a good example. It normally shows only in males, although a man inherits the characteristics of his beard just as much from his mother as from his father. Breast development in a woman and not in a man is another example. In general, the development of these traits in mammals is conditioned by the sex hormones. A woman who takes male sex hormones will develop a beard as a result of the action of the latent genes she has carried all her life. In cattle, the qualities of milk production are transmitted just as much from the bull as from the cow. The most outstanding cases of sexual dimorphism resulting from sex-limited genes is to be found in the birds. The brilliant plumage of the male pheasant or peacock stands out in great contrast to the drab pattern of feathers found in the female.

In invertebrate animals, sex-limited genes respond to the chromosomal make-up of the individual cells. Butterflies often show such great differences in the two sexes that they would never be taken for the same species had they not been seen mating. When a gynandromorph is found of the bilateral type, all of the tissue on one side of the body will show the typical male pattern, but that on the other side will show the female pattern. There is no mixing of traits as would be the case if sex hormones flowed to all parts of the body and influenced the expression of the genes.

SEX-LINKED INHERITANCE. A pattern of inheritance resulting from genes on the nonhomologous portion of the X-chromosome of most animals, but it can also be the genes on the nonhomologous portion of the W-chromosome of the animals with WZ sex determination. In the former case, the female carries two of each of the sex-linked genes, whereas the male carries only one of each. The males, therefore, express all of their sex-linked genes, and the females express those genes which are dominant or homozygous recessive, or intermediate in their expression. A male receives all of his sex-linked genes from the female parent; a female receives a sex-linked gene of each kind from both parents. Hemophilia, bleeder's disease, is a typical sex-linked gene in man. It is recessive, but shows when present on a man's single X-chromosome. Hence, half of the sons of a heterozygous woman will have hemophilia, but all of the daughters will be normal because they will receive a gene for normal blood clotting from their father. The gene for hemophilia was prevalent in the royal families of Europe. See accompanying diagram. Color blindness is another human trait resulting from a sex-linked gene.



Pedigree of hemophilia in the royal families of Europe. The gene apparently arose as a mutation in the immediate ancestry of Queen Victoria of England and spread through the royal families of Europe, with a great impact on history.

The first experimental evidence of sex linkage was found in 1910 when Morgan discovered a white-eyed mutant of *Drosophila melanogaster*. When white-eyed males were mated with normal red-eyed females, the F_1 generation displayed all red eyes. However, the F_2 generation included both red and white eyes in the proportion of 3 red to 1 white. This ratio suggested that the mutant gene for white eyes was recessive. The important observation was that the only males of the F_2 generation had white eyes. Therefore, the recessive gene for white eyes expressed itself only in males. Sex-linked genes in males are referred to as *hemizygous* since only one member of an allelic pair of genes is required for expression.

In birds the condition is reversed: the male has the two W-chromosomes and carries two of each sex-linked gene, whereas the female has a single W-chromosome and will express all the sex-linked genes. See **Genes and Genetics;** and **Heredity.**

Ann C. Vickery, Ph.D., Assoc. Prof., College of Public Health, University of South Florida, Tampa, Florida.

SEXTANS. A minor southern constellation located near the equator.

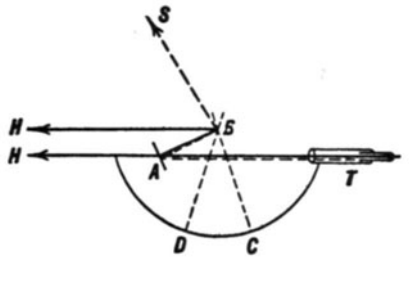
SEXTANT. The sextant is a light, portable instrument designed for the purpose of measuring the angular distance between the two objects. It represents the most recent stage in a succession of portable devices for angular measurement advancing from the astrolabe through the cross-staff down to the modern instrument. The immediate predecessor of the sextant was the quadrant (the "hog-yoke" of the sailing ship era), but the design of this instrument is very similar to that of the sextant used today. Since the sextant is used by navigators for the purpose of measuring the apparent altitude of celestial objects, the opinion is prevalent that it is purely a navigational instrument. Such is far from the case, and the instrument is of value to explorers, surveyors, or any person who desires to measure angular distance.

The optical system of the sextant (and quadrant) is shown in the accompanying diagram. The mirror *A* (called the horizon glass) is divided into two sections by a line parallel to the plane of the instrument. The upper section is unsilvered so that an observer looking through the telescope *T* can see directly through *A* to an object in the direction *H*. The mirror *B* (called the index mirror) may be rotated about an axis perpendicular to the plane of the instrument and coincident with the center of the graduated arc *CD*. Attached to the index mirror is a vernier arm which sweeps along the arc.

With the mirror *B* strictly parallel to *A* the observer will see two superimposed images of the object in the direction *H*. One of these is the direct image observed through the upper section of the horizon glass, the other is a reflected image with light traveling along the path *HBAT*. Under these conditions the index *C* should read zero; in case this is not so, an "index correction" must be applied to all observations with the instrument.

In case the angular distance between an object in the direction *H* and another in the direction *S* (angle *SBH*) is desired, the observer moves the index arm along the arc until an image of *S* (light path *SBAT*) is superimposed on the direct image of *H*. Application of the laws of reflection of light will show that the angle *CBD* through which the mirror is turned in one-half the angle *SBH*. To obviate the necessity of dividing each reading of the instrument by two, the arc is so graduated that when *CBD* is actually 60° the index will read 120°.

To obtain the altitude of a celestial object at sea, the instrument is directed toward the visible horizon and the index arm moved until the desired object is in the field of view. Now the sextant is rotated back and forth about the optic axis of the telescope (the line *HAT* in the figure),

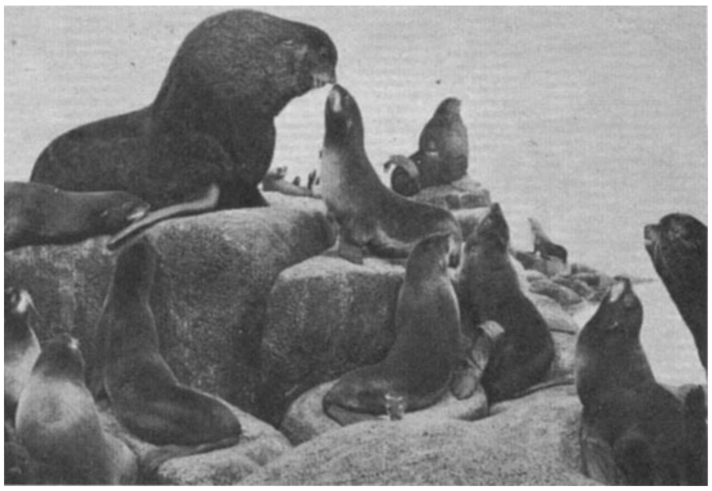


Optical system of sextant.

and the image of the celestial object will apparently swing back and forth along a short arc. The index arm is now carefully adjusted until the celestial object is just tangent to the horizon at the lowest point of its apparent swing. Under these conditions the reading of the sextant will give the sextant altitude measured along the vertical circle through the object.

To overcome situations where it is difficult or impossible to see the horizon clearly, a small bubble made visible in the instrument's optical system is used as an indication of vertical. A gyro vertical also has been used as a horizon reference.

SEXUAL SELECTION. A form of natural selection in which the sex of the individual plays an important part in the selection. In birds, we often find an extreme sexual dimorphism because of the effects of such selection. The male may have brilliant plumage, whereas the female may be much more drab in her plumage. This is supposed to have arisen because of the elaborate courtship of many birds in which the males display their plumage, and it is thought there is some degree of selection of mates by the females. In other species of animals, where the males have battles for possession of the females, as in the fur seals and the moose, selection favors strength and fighting ability in the males.


Sexual selection in the fur seals. The bull in the upper left has collected his harem of devoted females through countless victorious battles with other males. At the extreme right, an envious bachelor male looks on.

SEYCHELLES PALM. See **Palm Trees.**

SEYFERT GALAXY. Any galaxy having a very bright nucleus showing a high excitation spectrum with rather broad emission lines. Recently, some Seyfert Galaxies have been shown to be radio galaxies. See also **Galaxy.**

SHACKLE. A shackle is a piece used for connecting together two parts. The parts so connected can have some relative motion which is permitted by the shackle, but at the same time the extent of their freedom is limited by the restraining action of the shackle.

SHAD. See **Gizzard Shad; Herring.**

SHADBUSH. See **Rose Family.**

SHALE. A fine-grained sedimentary rock whose original constituents were clays or muds. It is characterized by thin laminae breaking with an irregular curving fracture, often splintery, and parallel to the often indistinguishable bedding planes.

SHANNON. In information theory, a unit of logarithmic measures of information equal to the decision content of a set of two mutually exclusive events expressed by the logarithm to base two; e.g., the decision content of a character set of 8 characters equals 3 Shannons. Synonymous with information content binary unit. (*American National Dictionary for Information Processing.*) See also **Information Theory.**

SHANNON FORMULA. A theorem in information theory which states that a method of coding exists whereby C binary digits per second may be transmitted with arbitrarily small frequency of error where C is given by

$$C = B \log_2\left(1 + \frac{S}{N}\right)$$

and no higher rate can be transmitted; B is the bandwidth, and S/N is the signal-to-noise ratio. See also **Information Theory.**

SHARED TIME CONTROL. In an automatic control system, control action in which one controller divides its computation or control time among several control loops rather than acting on all loops simultaneously.

SHARKS (*Chondrichthyes*). Of the order *Selachii*, there are several families of sharks. A shark is a carnivorous fish with a cartilaginous skeleton. The mouth opens on the ventral surface of the head and is armed with many rows of sharp teeth attached to the skin and similar in structure to the placoid scales of the body. The tail is of the heterocercal form, having two lobes with the backbone extending into the upper lobe. The openings of the gill slits are separate. Of the principal families of sharks, there are: (1) the frilled shark (*Chlamydoselachidae*); (2) the sixgill and sevengill cowsharks (*Hexanchidae*); (3) sand sharks (*Carchariidae*); (4) goblin sharks (*Scapanorhynchidae*); (5) mackerel sharks (*Isuridae*); (6) thresher sharks (*Alopiidae*); (7) the basking shark (*Cetorhinidae*); (8) the whale shark (*Rhincodontidae*); (9) catsharks (*Scyliorhinidae*); (10) false catsharks (*Pseudotriakidae*); (11) smooth dogfishes (*Triakidae*); (12) the requiem sharks (*Carcharhinidae*); (13) hammerhead sharks (*Sphynidae*); (14) hornsharks (*Heterodontidae*); (15) saw sharks (*Pristiophoridae*); (16) spiny dogfishes (*Squalidae*); (17) spineless dogfishes (*Dalatiidae*); (18) the alligator dogfish (*Echinorhinidae*); (19) the angel sharks (*Squatinidae*); and (20) carpet and nurse sharks (*Orectolobidae*).

Because of their danger and threat of danger, sharks have received much attention—over the many hundreds of years that people have been swimming in the seas, whether for recreation or as the result of accidents at sea. However, the interest in the dangers of sharks increased markedly with the advent several years ago of diving (for military or constructional purposes), and particularly with the great expansion of skin diving as a hobby and sport. The lore pertaining to the true danger and the habits of sharks with relation to the presence of people obviously would consume many books. Sharks are well known for their curiosity. The problem, of course, is to differentiate between their curiosity and their possible more serious, aggressive intentions.

Many sharks are docile by nature and require certain stimuli to alter their temperament from one of curiosity to one of aggression and attack. An almost certain stimulus is the presence of blood or fish juices in the water. Thus, in shark-infested waters, a diver never should retain any speared or otherwise injured fishes in the vicinity of skin-diving operations. Records indicate that sharks may be interested only in the injured fishes and not in the persons nearby, but in seeking out the source of fish juices and blood, the sharks may make mistakes and thus cause injury to persons in the area. Also, sharks do not like to be disturbed—they do not like to have objects in their way and they do not like to be touched, as for example a skin diver attempting to catch a shark by the tail. A shark may suddenly reverse direction in such a situation and attack the diver. Because of the sensitive and unpredictable temperaments of many sharks, the diver never should consider a shark "friendly" or become careless as the result of past interesting and uneventful encounters with sharks. For example, it has been postulated by some authorities that just as among people, there will be found the occasional psychotic shark which does not conform to normal habits and patterns of behavior. In all situations, of course, the size of the shark as an index of potential danger should always be considered. It is a good rule to consider any shark that is over 10 feet (3 meters) in length as a potential source of trouble.

Sharks are also known to be rather indiscriminate when they are feeding. Experiments with sharks in large aquariums, for example, have indicated that during a feeding "frenzy," a shark may easily mistake an undesirable food item for the items that it really wants most. Feeding sharks in captivity always must be done with precaution. The sharks, obviously, are not aware of where their food ends and where the feeder's hand begins. Further, during a period of feeding excitement, a captive shark may attack and eat other creatures in the aquarium that it normally would not bother.

Some authorities have suggested that there may be as many persons injured by shark bumps as by shark bites. There is considerable postulation in attempts to describe the reasons behind such statistics. ("Shark Attack" by V. M. Coppleson, published by Angus and Robertson, Sydney, Australia, 1959).

There appears to be a correlation between the temperature of the water and the feeding habits of some sharks. Statistics tend to show that fewer shark attacks occur in waters below 60°F ($\sim$ 16°C) and that the number goes up in waters above 70°F ($\sim$ 20°C). This correlates with the fact that the sharks most dangerous to people are found in tropical waters. Also, shark attacks in temperate waters are usually limited to summer months. This fact, of course, could be easily explained because there are more people in the water during the summer than at other times. In the study by Dr. Coppleson, it is also pointed out that shark attacks in Australian waters tend to occur in mid-afternoon. Statistics also show that sharks prefer male to female human victims (a ratio of 20:1).

Much investigation has gone into the development of shark repellents. This topic has always been of interest to the military (abandonment of aircraft over shark-infested waters; activities of naval divers, etc.). One relatively successful repellent developed during World War II was copper acetate containing a nigrosene dye (ratio of 20–80%). This was found quite effective with Atlantic sharks, but quite ineffective with Pacific sharks. This difference still requires further investigation. It has been noted that captive sharks often will not feed if the aquarium water contains copper compounds introduced for controlling parasites. Apparently, these compounds cause an irritation of the sharks' nostrils, thus disturbing their sense of smell and consequent desire for food.

Because of its preference for deep waters, the data on the frilled shark are not extensive. The number of specimens caught has been limited. The adult attains a length of about $6\frac{1}{2}$ feet (2 meters), the young are born alive, the coloring is brown with no particular markings.

The more primitive sixgill and sevengill cowsharks are identified by counting the gill openings on each side of the head. The main groups of sharks always possess five gill openings per side. Coloring of cowsharks varies from gray to brown with no particular markings. A giant sixgill shark was caught in English waters over a hundred years ago. It measured over 26 feet (7.9 meters) in length. This is the largest sixgill shark on record. They are widely distributed in Atlantic and Pacific waters. The young are born alive, seldom exceeding 16 inches in length. The sixgills apparently prefer deep water during daytime hours, but frequent the surface after dark for feeding. The Pacific sevengill is the broad-headed *Notorhynchus maculatum*. It prefers offshore waters and is rarely seen in shallow water. The shark is dark gray with black spots. There is a nursery ground of these sharks in the southern end of San Francisco Bay, an exception to their preference for deeper water. Generally, the sevengills are not considered edible.

Although much remains to be learned about sharks, a number of past misconceptions have been exposed in recent years. For example, sharks are not ravenous eaters. Grey reef adults feed only once in six or even twelve days. Species living on the sea bottoms go without food for weeks, and the big basking sharks found in British waters appear to starve all winter. Nor, as once believed, do sharks depend entirely upon their sense of smell. This sense, of course, is extremely acute. It has been estimated that sharks can detect (by smell) dilutions of one part in a million. Their sense of hearing also registers sounds hundreds of meters away. Although the shark's vision is lacking in resolution of details, it does perform well under poor lighting conditions. Sharks also are able to sense electric fields in surrounding waters and may use this sense as one means of navigation. Although sharks must continue to swim to avoid sinking, only the very slightest movements are required by some species. It has been observed that the Australian gray nurse shark can virtually hover. It has been noted that the great white man-eating shark can nearly pivot on its nose. The outstanding buoyancy of some shark species is attributed to very large livers containing light oils and fats, which act as swim bladders. At one time, it was believed that sharks had to keep moving in order to breathe, requiring a continuous flow of water past the gills, but it has been shown that many species, including some of the largest, such as the tiger and bull sharks, can achieve temporary breathing through muscular pumping.

The sand shark is a dangerous-looking creature with large mouth and wicked teeth. The sand tiger shark typifies one's visions of a dangerous, wicked shark. There are no records of attack on people on the American side of the Atlantic, but it is highly feared in South African waters. The largest recorded specimen measured about $10\frac{1}{2}$ feet (3.2 meters) in length, with a weight of over 300 pounds (136 kilograms).

Of the mackerel sharks, the most famous is the great white shark (*Carcharodon carchiarias*), also known as the maneater or maneating-shark. According to records, one of the largest maneaters was caught off Port Fairey, Australia in 1870. The jaws of this shark are preserved and on display at the British Museum. The shark has been reported as measuring about $36\frac{1}{2}$ feet (11.1 meters) in length. However, the majority of maneaters caught have ranged between 20 and 25 feet (6.1 and 7.6 meters) in length. The bodies of these sharks are massive and hence a shark of only 17 feet (5.2 meters) in length may weigh up to nearly 3,000 pounds (1361 kilograms). The weight record may be held by a 21-foot (6.4 meters) shark caught in Cuban waters. It is estimated to have weighed about 7,000 pounds (3175 kilograms). J. E. Randall (*Science*, **181** (4095), 169–170 (1973)) casts some doubts over these previously reported dimensions.

The temperament of the maneater is considered very bad—with the viscious habit of considering just about anything it sees as edible. Some of the contents found in the stomachs of captured maneaters have included the remains (in some cases, the intact bodies) of large dogs, seals and sea lions themselves weighing in excess of 100 pounds (45.3 kilograms), and, interestingly, other sharks that may measure up to 7 feet (2.1 meters) in length.

The mako or sharp-nosed mackerel shark is known for its great activity once hooked, displaying marlin-like maneuvers and swimming much faster than its relative, the maneater. The *Isurus oxyrhinchus* is the Atlantic mako. The largest specimen on record attained a length of about 12 feet (3.6 meters) and weighed around 1,200 pounds (544 kilograms). The *Isurus glauca* is the Indo-Pacific mako. Both species of makos prefer tropical seas. The makos tend to gulp their food, as evidenced by finding large intact items in their stomachs. In one instance, a 120-pound (54.4-kilogram) swordfish was found in the stomach contents of a Bahaman mako which weighed over 700 pounds (317.5 kilograms). Other species of mako sharks include the *Lamna nasus* (common Atlantic mackerel shark, sometimes called porbeagle); and the *Lamna ditropis* (Pacific mackerel shark).

The thresher shark is noted for a very long whiplike tail. This may equal the length of the body, providing the fish with power and maneuverability. They are of offshore, tropical distribution. Large specimens run about 20 feet (6.1 meters) in length and weigh up to about 1,000 pounds (453.6 kilograms). The number of species of thresher sharks is quite limited.

The basking shark is well named because of its apparent preference to spend much time simply floating along the surface or cruising at very slow speeds. The second-largest known shark is *Cetorhinus maximus*, the giant basking shark, with recorded lengths up to 45 feet (13.7 meters). Very much like a mackerel shark, the basking shark differs in its preference for plankton rather than carnivorous food. It is believed that the young are born alive, probably up to 6 feet (1.8 meters) in length. Although found worldwide in temperate waters, there are significant concentrations reported in waters off southern California and Europe. Fishing for basking sharks is considered important commercially, notably because of the value placed upon their large livers (yielding oil). For example, a liver weighing over 850 pounds (386 kilograms) may be found in an 8,500-pound (3856-kilogram) shark about 30 feet (9.1 meters) in length. Unfortunately, the liver oil contains no vitamins, but it is used for various tanning processes.

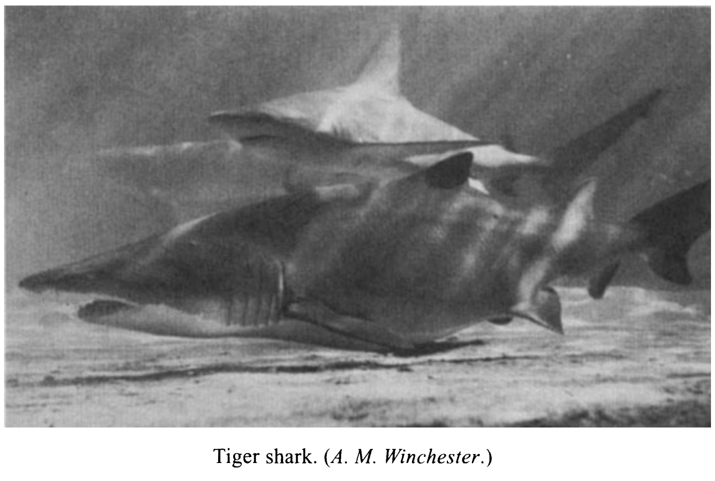
Regarded as the largest of fish, the whale shark attains a length of 45 feet (13.7 meters) at minimum. There have been numerous reports and writings over the years pertaining to the docile nature of these large creatures. Although the whale shark feeds on very small substances for food, such as small fishes, squid, and crustaceans, it nevertheless possesses numerous small teeth. The sharks feed in a vertical position. In American Atlantic waters, the population center is around the Caribbean. There is a concentration in the Pacific off the Gulf of California. Whale sharks also have been seen in the Red Sea and off the Philippines.

Catsharks have beautiful stripes, bars, and mottling, prefer inshore waters, and reach a maximum length of about 3 feet (1 meter). They are elasmobranches (cartilaginous skeleton, platelike scales, lack of air bladder). The common European spotted dogfishes (*Scyliorhinus caniculus* and *S. stellaris*) are catsharks (even though named dogfishes). The *S. retifer*, also called the chain dogfish, is found in the waters off New Jersey. The South African "skaamoog" (*Holohalaelurus regani*) has markings suggestive of Egyptian hieroglyphics. The *Cephaloscyllium*, which attains a length of about 4 feet (1.2 meters), ranges the eastern Pacific, but is absent from Atlantic waters. The *Cephaloscyllium uter* is also known as a swell shark and occurs from Monterey Bay in California to Lower California waters. Its name is derived from the fact that, upon being pulled from the water, it imbibes large quantities of air, sometimes swelling to twice its normal size. If released to the water, the air is expelled and the fish returns to normal dimensions.

Only two species of the false catsharks are known. These sharks are quite large and prefer deep water. They are rarely seen. The identifying marking of these sharks is the long base of the dorsal fin. The largest specimen from the Atlantic measured somewhat less than 10 feet (3 meters). Of the specimens captured, all have been taken from waters ranging from 1,000 to 5,000 feet (305 to 1524 meters) in depth.

Small in size, usually less than 5 feet (1.5 meters) in length, there are relatively few species of the smooth dogfishes, which are more or less intermediate between nurse sharks and catsharks. Very abundant in American Atlantic coastal waters is the *Mustelus canis*, ranging from as far south as Uruguay northward to Cape Cod. They have been carefully studied and their biological traits and life history are well known. The smooth dogfish is known for its ability to change its coloration. It is not considered edible. *Rhinotriacis henlei* is the most abundant of sharks in American Pacific coastal waters. Other members of the family of smooth dogfishes (*Triakidae*) include the spotted shark (*Mustelus punctulatus*) found in the Mediterranean and the waters of South Africa. Achieving a length of about 6 feet (1.8 meters), this shark is covered with tiny black spots, hence its name. The *Triakis semifasciata* (leopard shark) is found in American Pacific coastal waters, ranging from Lower California northward as far as Oregon. San Francisco Bay is a nursery ground for this species.

There are well over 60 species of requiem sharks (*Carcharhinidae*). They enjoy the characteristics of typical sharks. The *Galeocerdo cuvieri* (tiger shark), shown in the accompanying diagram, is omnivorous, eating birds, fishes, animals, garbage, coal, turtles and so on. Its base color is a grayish-brown with a lighter-colored undersurface. There usually is


Tiger shark. (*A. M. Winchester.*)

some mottling along the upper surfaces. The tiger shark attains a length of about 18 feet (5.4 meters) in American waters; probably about 14 feet (4.2 meters) in Australian waters. The weight may be as much as 1,400 pounds (635 kilograms). Preferring deep water, the tiger shark on occasion will follow prey into shallow waters. Commonly found in the Caribbean and the waters around Florida, the species is not known along the northern European coasts or in the Mediterranean.

During World War II, shark livers became an important source of vitamins as the result of the blockading of Norway, since cod-liver oil was previously obtained from Norway. Thus, so-called soupfin sharks were commercially sought, even on the California coast, until about 1950 when low-cost synthetic vitamins appeared. The best sharks for liver were the soupfin (*Galeorhinus zyopterus*) and the dogfish (*Squalus acanthias*). The *Galeorhinus australis* (Australian school sharks) bears resemblance to the American Pacific soupfin. However, the latter is not considered a food fish, whereas the school shark is important commercially.

The great blue shark is also a requiem shark, well named because of the indigo blue coloration of its upper surfaces. It attains a length of nearly 13 feet (3.9 meters) and is long, slender, and streamlined, with a sharp nose. It is considered a sporting fish, but not edible. Common along the American Atlantic coast is the lemon shark (*Negaprion brevirostiris*), frequenting the waters from Brazil as far northward as North Carolina. These sharks reach a length of from 7 to 11 feet (2.1 to 3.3 meters). They prefer inshore shallow waters. Biological studies of the lemon shark have been carried out at Florida's Cape Haze Marine Laboratory. The silky shark (*Carcharhinus floridanus*) is also a requiem shark and is abundant along both Atlantic and Pacific coasts. Of interest, despite its abundance, is the fact that it was not officially identified and catalogued until 1953. The white-tip shark (*Carcharhinus longimanus*) with a gray coloration is probably the most common of sharks in the Atlantic and Pacific coastal waters. It reaches a length of about 13 feet (3.9 meters).

A number of sharks can tolerate brackish water and may live for awhile in fresh water, but the only known species that lives permanently in fresh water is the maneating *Carcharhinus nicaraguenis* which inhabits Lake Nicaragua. This shark is gray with a very heavy body, reaching a maximum length of about 8 to 10 feet (2.4 to 3 meters).

Because of the most unusual and well-named head, the hammer-head shark (*Sphyrnidae*) is readily identified. It is interesting to note that in a large specimen of, say, 15 feet (4.5 meters) in length and weight up to 1,500 pounds (680 kilograms) the eyes will be separated from each other by as much as 3 feet (1 meter). Because of its peculiarly shaped head, another member of this family of sharks is called the "shovelhead." Hammerheads prefer tropical seas, but move north into temperate waters in the summer. In 1959, there was an epidemic of shark attacks on the west coast of the United States, attributed in the main to hammerheads. The unusual shape of the hammer-type head has aroused much biological interest. Some experts postulate that the hammer may serve as a balancing mechanism. Because hammerheads do not survive long in captivity, studies have been difficult.

Additional excellent information on the complex and extensive topic of sharks can be obtained from a number of the references listed under the entry on **Fishes.** Of particular note is the special coverage of sharks in *Oceanus*, **24**(4), Winter 1981–1982.

SHASTA FIR. See **Fir Trees.**

SHAULA (λ Scorpii). Ranking twenty-third in apparent brightness among the stars, Shaula has a true brightness value of 1,700 as compared with unity for the sun. Shaula is a blue-white, spectral type B star and is located in the constellation Scorpius, a zodiacal constellation. Estimated distance from the earth is 300 light years. See also **Constellations.**

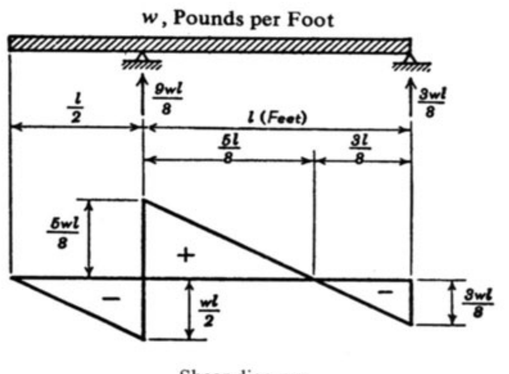
SHEAR. A force that lies in the plane of an area or a parallel plane is called a shearing force. It is the force which tends to cause the plane of the area to slide on the adjacent planes.

The vertical shear for any section of a simple beam is the magnitude of the resultant of the transverse loads on either side of the section. Transverse loads are those which are at right angles to the length of the beam. If the loads are inclined, the vertical components, only, should be used in computing the vertical shear. The resisting shear at any section is the internal force which opposes the shearing action of the external loads. It is numerically equal to the external shear but in the opposite direction. Vertical shear, which is always accompanied by bending movement at a section of a beam, is numerically equal to the rate of change of this moment with respect to distance along the beam. This shear is arbitrarily assumed to be positive if the resultant of the vertical loads to the left of a section acts in an upward direction. In a symmetrically loaded simple beam the shear is equal to zero at the center of the beam. See **Elasticity.**

In addition to vertical shear in a beam, there is always a horizontal shear which is a result of the difference in the flexural stresses (see **Flexure**) between any two vertical planes. The tendency of adjacent horizontal planes to slide upon each other is caused by horizontal shear. The effect may be better understood by visualizing a beam composed of flat planks laid one on top of the other. As the beam bends, due to the applied loads, the bottom of one plank will slide upon the top surface of the one beneath it unless this effect is restrained by friction, nails, bolts, or other fastenings.

The unit stresses resulting from vertical shear are called vertical shearing stresses. At any point in a beam, these stresses are numerically equal to the horizontal shearing stresses. The variation of the unit shearing stresses over the cross section of a rectangular beam is parabolic, being equal to zero at the top and bottom surfaces and a maximum at the neutral axis. When horizontal and vertical shear, only, act at a point in a body, the body is said to be in a state of pure shear at the point.

Shear is not restricted to beams. It occurs wherever there is bending. Columns, which are subjected to bending caused by eccentric loads, or by inclined or lateral loads, must be designed to withstand the shearing stresses. Rivets and welds are also subjected to shear. If the riveted connection is made so that the shear occurs between two plates only, it is called a single shear. When the type of connection is such that the shearing force is opposed by resisting shears acting on two planes, as in the case of three plates riveted together, the condition is called double shear.



Shear diagram.

A shear diagram is a graphical representation of the variation of vertical shear on a beam. An illustration of a shear diagram is given above for an overhanging beam with a uniformly distributed load covering the entire length of the beam. The points where the shear changes sign are points of maximum bending moment. The area of the shear diagram between any two points is equal to the change of bending moment between these points.

SHEAR CENTER. The point through which the external shear must act at any cross section of a beam in order to eliminate torsional stresses is called the shear center of the particular cross section in question. The shear center coincides with the centroid of the internal shearing forces on the cross-section. If the beam is to bend without twisting, the loads must be applied in such a manner that the external shear at any section will pass through the shear center. The shear center has no meaning for sections where pure bending occurs because there is no shear and, therefore, no torsional stress can exist under these conditions.

SHEATHBILL. See **Shorebirds and Gulls.**

SHEEP. See **Goats and Sheep.**

SHEEPSHEAD. See **Porgies.**

SHEEP TICK (*Insecta, Diptera*). A wingless parasitic fly, *Melophagus ovinus* (Linne), that resembles a true tick only in its flattened body and leathery texture. Like the other flies, it has sucking mouth parts and hence draws blood from the skin of the host. Some scientists have described the insect as a degenerate, louselike fly that has completely lost its wings. The insect is especially harmful to lambs. The tick is irritating to the sheep, causing it to rub and scratch, sometimes injuring the fleece and reducing its marketable value. It migrates from ewes to lambs at shearing time. It also attacks goats. Distribution is worldwide in all sheep-producing regions.

The adult sheep tick is about $\frac{1}{4}$ inch (6 millimeters) in length and has a covering of spiny hairs. The tick is also present on sheep in another stage of development, when it is somewhat like a brown seed, tenaciously fastened to the hair of the animal, especially inside the thighs, along the belly, and around the neck. This is the pupal stage of the fly. The sheep tick does not lay eggs, but is live-borne. The young are about $\frac{1}{8}$ inch (3 millimeters) in length, oval in shape, and off-white in color.

Numerous commercial preparations are available as sprays and dips for controlling the sheep tick. Such formulations may include lindane, malathion, rotenone, or toxaphene. The natural pyrethrins or allethrin are also effective. Care must be taken in applying control chemicals and to make certain that such applications are not made just prior to shearing or marketing. Sheared wool should not be kept near flocks for fear of infestation traveling from one to the other and, after dipping or spraying, the animals should be turned into an area that is considered free of ticks. Previous areas of infestation should be vacated for up to 6 weeks.

SHELL. A hard external covering secreted by folds of the body wall of many animals. The term applies properly to the shells of Brachipoda and Mollusca, although it is sometimes used in reference to the hard exoskeleton of crustaceans. Also a hard covering of eggs.

Brachiopod shells consist of two valves, one upper and one lower. They have an outer layer of organic matter known as the periostracum, under it a thin layer of calcium carbonate, and a thick inner layer of mixed organic and calcareous matter, deposited in prismatic form. The valves are opened and closed by a complex system of muscles.

Molluscan shells are usually spiral in form, like many common snail shells, or bivalve like those of mussels and the oyster. The valves of such shells are lateral in position. A third rarer form is the shell of nautilus, which is spirally coiled in one plane. Internal shells of slugs, chitons, and some cephalopods are in the form of plates of calcareous matter.

The external shells of mollusks have an external horny layer, the periostracum, a smooth lining of nacre or mother-of-pearl, and a thick calcareous middle layer.

Because of their permanence and beauty the shells of many marine mollusks have attracted the attention of collectors, and many have received common names. Among them the spiral staircase or wentletrap, periwinkle, conch, finger shell, cowry, coffee-bean, apple-seed, oyster drill, whelk, papal miter, cockle, gem, and others are to be found on the coasts of the United States.

The shells of eggs are calcareous or chitinous coverings secreted by a portion of the female reproductive ducts known as the shell gland. In the birds they are characteristically and often beautifully colored, and in the insects they may be beautifully sculptured. They are sometimes perforated by a minute opening or group of openings called the micropyle for the entry of the sperm. The egg shell of insects is also called the chorion, a term not to be confused with the chorion of vertebrate embryos. See also **Mollusca.**

SHELLAC. A secretion or excretion of the lac insect, *Coccus lacca*, found in the forests of Assam and Siam. Freed from wood it is called "seed lac." It is soluble in alkaline solutions such as ammonia, sodium borate, sodium carbonate and sodium hydroxide, and also in various organic chemicals. When dissolved in acetone or alcohol, shellac yields the familiar shellac varnish of superior gloss and hardness. Orange shellac is bleached with sodium hypochlorite solution to form white shellac.

SHELL (Atomic). See **Chemical Elements.**

SHELLFISH. Any sea animal that is protected by a hard shell, such as crabs, lobsters, crawfish and crayfish, and shrimp of the class *Crustacea*; and clams, quahogs, oysters, and snails of the phylum *Mollusca*. See also **Crustaceans (Edible);** and **Mollusks.**

SHELLFISH POISONING. See **Foodborne Diseases.**

SHELL MOLDING. The forming of a mold cavity for casting metals by placing sand bonded with a thermosetting resin against a preheated metal pattern (150 to 300°C).

SHEPPARD'S CORRECTIONS. If a continuous frequency distribution is grouped, the moments derived by replacing each observation by the central value of the group into which it falls differ from those of the original distribution. Provided the distribution tapers smoothly to zero in both directions, average corrections to the grouped moments, known as Sheppard's corrections, may be applied as follows:

$$\begin{aligned} \mu_1 \text{ (corrected)} &= \mu_1 \\ \mu_2 \text{ (corrected)} &= \mu_2 - h^2/12 \\ \mu_3 \text{ (corrected)} &= \mu_3 \\ \mu_4 \text{ (corrected)} &= \mu_4 - \tfrac{1}{2}h^2\mu_2 + 7h^4/240 \end{aligned}$$

SHERARDIZING. The process for applying an adherent protective coating of zinc to steel parts by heating at 700°F (371°C) in contact with zinc dust in a rotating-drum container.

SHIELD (Geology). From a tectonics viewpoint, a shield is a large area of exposed basement rocks in a craton commonly with a very gentle convex surface, surrounded by sediment-covered platforms; e.g., Canadian Shield, Baltic Shield. The rocks of virtually all shield areas are Precambrian. See also **Craton.** From a paleontological viewpoint, a shield is (a) a protective cover or structure on an animal, likened to or resembling a shield; e.g., an ossicle of an ophiuroid, or the carapace of a crustacean, or a large scale on the head of a lizard; (b) a float or curved, lateral outgrowth at one or more levels of a tangential rod or needle in the skeleton of an acantharian radiolarian and forming by fusion the lattice shell; (c) one of the discoidal elements of the placolith coccolith.

SHIGELLOSIS. See **Foodborne Diseases.**

SHINGLES. See **Dermatitis and Dermatosis.**

SHIPWORM (*Mollusca, Lamellibranchiata*). A peculiar marine mollusk which bores into submerged wood and apparently is among the few animals that can digest cellulose and related materials. The shipworms belong to several genera of which *Teredo* is most often cited. All are slender wormlike creatures but they have the characteristic structures of the bivalves. The valves of the shell are small separate parts, located at the anterior end of the worm, and are used for excavating the burrow.

Shipworms do great damage to wooden hulls and marine piling, consequently they have been subject to detailed studies to determine methods of avoiding their destructive attacks.

SHITTIMWOOD. See **Acacia Trees.**

SHOCK SYNDROME. In the simplest of terms, shock is the inadequate delivery of blood to the major organs of the body. Unless immediately treated, deprivation of blood supply causes disturbance of the metabolism (sometimes a shift from aerobic to anaerobic metabolism at the cellular level) of the organs with resultant damage. Because of these profound consequences, the treatment of shock is considered an emergency procedure. Cellular damage may be reversed with very prompt treatment, but is otherwise irreversible, leading to the ultimate death of the patient. Recovery from shock depends upon promptness of treatment and the age and general underlying health of the patient. Authorities suggest five broad categories of shock, based upon causation.

(1) *Hypovolemic shock.* Several conditions cause a massive loss of blood, plasma, or extracellular fluid from the body or intravascular compartment. The latter may be lost from the gastrointestinal tract from vomiting or diarrhea, abusive use of diuretics, in extensive burns, as well as acute pancreatitis. The most common loss of blood and plasma is encountered in hemorrhagic shock, as that which may result from serious trauma or severe gastrointestinal bleeding. The arterial blood pressure is lowered, causing a deficient supply of blood to tissues, as the result of loss of fluid volume from the vascular space.

(2) *Septic shock.* This may occur as the result of septicemia caused by a gram-negative bacterial infection. Bacterial endotoxin (a complex lipopolysaccharide) is released into the bloodstream. Septic shock is only infrequently a consequence of infections by gram-positive organisms, viruses, fungi, or rickettsias. Sequestration and pooling of blood in various vascular compartments lowers the availability of blood for the perfusion of other vital organs.

(3) *Cardiogenic shock.* This may result from a massive myocardial infarction caused by extensive damage to the myocardium. This sometimes occurs in connection with cardiac surgery; less commonly by severe myocarditis. Cardiogenic shock also may be caused by an arrhythmia in a patient with serious heart disease. In essence, the heart fails to pump, causing a reduction in cardiac output and arterial pressure.

(4) *Obstruction to cardiac filling.* Where cardiac filling is prevented or lessened, as by a massive pulmonary embolism or tumors or other space-occupying lesions, this type of shock may be precipitated.

(5) *Neurogenic shock.* In this type of shock, there is loss of vasomotor tone. This may arise from general or spinal anesthesia, an injury to the spinal cord, or from the massive intake of depressant drugs, such as certain narcotics and barbiturates. Respiratory arrest causing sustained hypoxia also may be a factor in this syndrome.

Depending upon the type of shock, various drugs are administered, including the catecholamines, such as norepinephrine, epinephrine, metraminol, dopamine, and isoproterenol. These drugs increase the arterial perfusion pressure. Vasodilating agents also may be used, particularly in connection with the treatment of septic shock and in some cases of hemorrhagic shock. These drugs, which include sodium nitroprusside, nitroglycerin, isosorbide dinitrate, phentolamine, and adrenal corticosteroids, must be administered with extreme care while the patient is constantly monitored. In cases of cardiogenic shock, circulatory assistance may be provided by the use of intra-aortic balloon counterpulsation. The balloon is inserted through a femoral artery and positioned, usually with the assistance of fluoroscopy, in the descending thoracic aorta. The balloon is programmed from the electrocardiogram such that it will deflate just prior to systole and inflate in diastole. Even though the use of intra-aortic balloon counterpulsation has been successful in many cases, it is estimated that to date it has increased the survival rate from cardiogenic shock by only 5 to 10% of cases. Where possible, if the patient has survived for 2 days with balloon counterpulsation and if other conditions are favorable, angiography and coronary artery revascularization will be attempted. See also list of entries given in the entry on **Heart and Circulatory System (Human).**

SHOCK TUBE. A relatively long tube or pipe in which very brief high-speed gas flows are produced by the sudden release of gas at very high pressure into a low-pressure portion of the tube; the high-speed flow moves into the region of low pressure behind a shock wave.

SHOCK WAVE. Infinitesimal disturbances in a fluid medium are propagated with a characteristic speed known as the sound speed. When the restriction on the amplitude of the disturbance is lifted, the linear approximation breaks down and the velocity of propagation becomes dependent on the amplitude of the disturbance. Another feature of this phenomenon is that the forward gradient of the disturbance rapidly steepens until it becomes a discontinuity and propagates as such. A *shock wave* is then a discontinuity in the physical properties of a fluid medium which propagates through the medium at supersonic velocity without further change. The strength of the shock is defined by the Mach number, the ratio of its velocity to the undisturbed sound speed. Such waves are generated by the detonation of explosive material, by high-speed aircraft and missiles, and by earthquakes.

Since all media are necessarily discrete, a true discontinuity is inconceivable, but, as the thickness of the shock transition corresponds to only a few mean free paths in a gas (or internuclear distances in a solid), the transition can be treated as a discontinuity to the same extent that the medium can be regarded as continuous. In comparison with an adiabatic or isentropic change, the shock wave is an irreversible process and hence leads to an increase in the entropy of the material. The pressure, density, and temperature of the medium are all raised on passage through the shock and the flow velocity is reduced. The latter is easily understood by observing that, with respect to the moving front, the molecules enter with an ordered flow motion at supersonic speed and the transport processes in the front transform a major fraction of this ordered flow into the random temperature or kinetic motions of the molecules.

The extent to which the various properties change through the transition depends on the magnitude or strength of the shock and on the thermodynamic properties of the fluid. For an essentially incompressible material, such as a liquid or solid, the major change normally occurs in the pressure variable, whereas for a gaseous medium the most significant change is in the temperature. Although shock waves in solid and liquid materials have been used to study physical properties at high pressures, the method is rather limited by the small test times available before the interaction of other wave phenomena which prevent the attainment of thermodynamic equilibrium, and it is in gases that shock waves have proved of most interest.

The detailed behavior of the shock transition is in itself a most important subject for study, since shock waves are associated with the flight of supersonic aircraft and with the re-entry of ballistic missiles into the earth's atmosphere. In addition to their own intrinsic interest, shock waves are important for other reasons. Since the transition involves the translational motions of the molecules, energy must eventually be transferred into other modes before the system reaches equilibrium. These subsequent relaxation processes involve the rotation, vibration, chemical reaction, electronic excitation, and even ionization of the molecules if the shock is sufficiently strong. The shock phenomenon thus provides an excellent method for studying energy transfer processes. For chemical reactions in particular, the shock wave provides a source of heat which is essentially instantaneous and is completely homogeneous. Also, provided the thermodynamic properties of the medium are known, the temperature is completely defined by a determination of the shock velocity.

Although shock waves can be created in many ways, including the detonation of high explosives and in wind tunnels, the simplest technique makes use of the *shock tube*, discovered in 1899 by Vieille. A long tube of uniform cross section is divided into two parts by a thin diaphragm and gas is admitted to these at different pressures. If the diaphragm is ruptured in some way, a shock wave is generated in the low-pressure gas and a corresponding rarefaction, or expansion, wave in the driver gas. Because the motion is restricted to a single dimension by the containing walls, the strength of the shock does not decrease with distance as it would in a three-dimensional expansion and the relaxation processes become simple functions of distances behind the front. This extremely simple piece of equipment can generate temperatures up to 20,000 K, since the strength (or velocity) of the shock depends only on the pressure ratio across the diaphragm immediately prior to rupture and on the thermodynamic properties of the gases in the two sections.

The disadvantage of all shock tube work is that the front moves so rapidly and subsequent wave interactions follow so soon afterwards that the available testing time is very short, often as little as 100 microseconds. Shock waves can also be created in highly ionized media where the forces are Coulombic in origin and the shocks are termed "collisionless." In this situation, shock waves lie more properly in the realm of plasma physics.

See also **Aerodynamics;** and **Supersonic Aerodynamics.**

SHOOTING STAR. The popular term used to designate meteors. These objects bear little if any relation to the stars other than that they are seen as bright, rapidly moving objects against the dark sky.

SHORE EFFECT. The change in the characteristics of an electromagnetic wave as it passes along a land-sea boundary, due to a difference in the propagation characteristics of the two regions. A source of error in radio direction-finders.

SHORT CIRCUIT. An electrical circuit is considered to be shorted when the terminals are connected directly together with only the impedance of the short connecting leads between them, thus for all practical purposes there is no resistance between them, hence no voltage can exist between them. While shorting a circuit which does not contain and is not connected to any source of voltage will produce no harmful effects, shorting a set of terminals across which a voltage normally exists will produce in many instances disastrous current flows. In power circuits, protection is often provided by circuit breakers or fuses which open the circuit under the high values of current which will flow on short-circuits. Even then the transient effects which result from short circuits may cause generators to arc over. A short obviously puts the circuit out of use.

See also **Circuit Breaker;** and **Fuse (Electric).**

SHORT-RANGE FORCE. A force between two particles which is essentially ineffective when the interparticle separation exceeds a certain distance; usually applied to nuclear forces which have a range of several times 10^{-13} centimeter.

SHOVELLER. See **Waterfowl.**

SHOWER. See **Precipitation and Hydrometeors.**

SHREWS. See **Moles and Shrews.**

SHRIKE (*Aves, Passeriformes*). This bird (*Lanius*) is chiefly known for its habit of catching other birds and small animals and impaling uneaten remnants on thorns. It is also known as the butcher bird. The beak is notched and in some species hooked. The numerous species occur on all continents but South America. See accompanying illustration. The minivet is a brightly-colored shrike (cuckoo-shrike), several species of which inhabit eastern Asia and India. The bird is about 6 inches (15 centimeters) long, is black along the back, with bright orange underneath and on the tail and wings. The female tends to be dark gray with a dull yellow coloration. Nests are cup-shaped and made of roots, pine needles, spider webs, and twigs. There are three to four green eggs with pale pink spots.

Shrike.

SHUNT (Electrical). An electrical bypath so arranged that an electric current divides and flows partially through a second path (termed *shunt*). Shunts are employed in instrument circuits and other electrical equipment. For example, in the shunt-wound generator, the field coils are shunted across the armature circuit. The shunt resistance is very high and consequently only a very small portion of the current flows through the shunt winding.

A current by-pass often is used with permanent-magnet moving-coil electrical instruments because only currents up to about 50 milliamperes can be taken into a moving-coil through the springs. Direct current instruments in excess of this range require the use of a parallel resistance circuit formed by one or several shunts. Where these ranges are moderate, the shunts usually are self-contained. On higher ranges, the shunts become physically large and convert more than a few watts into heat. One kilowatt for 50 millivolts, 20,000 amperes. Thus, they are used as accessories external to the instrument. Small or large, these devices consist of one or several manganin conductors terminating in copper blocks which are provided with separate terminals for connection to the instrument to avoid errors.

The manganin sections are soldered into the copper blocks. Shunt construction is such that heat is carried off at a rate sufficient to keep the operating temperature below the softening point of the solder. Adequate conductors tightly fastened, clean contact surfaces, and free air circulation are important.

Ammeters for use with external shunts are provided with special leads for shunt connection. Shunts are usually made to produce a standard potential drop, such as 50 millivolts at rated current and the associated dc mechanism is then built to give fullscale deflection on a slightly smaller potential so as to allow for lead resistance. As the leads form part of the mechanism circuit, shown in Fig. 1, their resistance must not be altered.

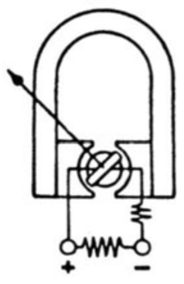


Fig. 1. Use of a shunt with a permanent-magnet moving-coil ammeter.

Usually the current through the mechanism is a neglible portion of the total and the potential drop across the shunt terminals is nearly the same with or without the mechanism in the circuit. In instruments of moderate precision and fairly high current range, this difference can be neglected and shunts and instruments made interchangeable. In instruments of high precision and in instruments of low current range, the shunt adjustment must take into account the instrument current. Such combinations are usually not interchangeable.

Multirange shunts should be connected as shown in Fig. 2 to avoid the use of a switch with its variable contact resistance between shunts and mechanism.

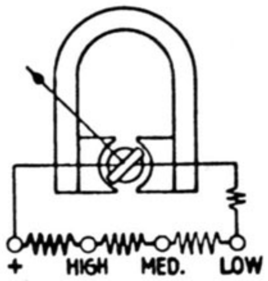


Fig. 2. Preferable manner for connecting a multirange shunt to avoid use of switch.

Because of the relatively high current consumption of alternating current instruments, shunts are not used for obtaining either multiple ranges or extending the base range. The division of current between mechanism and shunt would become unfavorable and would invite inaccuracies due to the difficulty of obtaining good pressure contacts, as well as the fact that the current division would be a function not merely of the relative resistances of the parallel circuits, but also of their relative impedances. Moreover, ac mechanisms are not suitable for low millivolt ranges so that shunts would have to have large potential drops and would generate excessive heat. The ranges of alternating current ammeters are, therefore, extended with the aid of current transformers.

SIAMANG. See **Anthropoids.**

SIBERIAN HIGH. See **Atmosphere (Earth).**

SICKLE CELL ANEMIA. See **Anemias.**

SIDEREAL PERIOD. The sidereal period of any object is its period of revolution around its primary. In general, sidereal period may be defined as the time required for an object to move from a particular position among the stars back to the same longitude again, as seen from the sun.

SIDEREAL TIME. See **Time.**

SIDERITE. This mineral is a carbonate of iron, $FeCO_3$. It is hexagonal with rhombohedral crystals, and also occurs in various massive forms. It has a rhombohedral cleavage; uneven fracture; is brittle; hardness, 3.75—4.25; specific gravity, 3.96; luster, vitreous to pearly; color, gray, yellowish- or greenish-gray, green, reddish-brown and brown. Siderite is found as concretionary masses in the sedimentary rocks; as a replacement mineral from the action of iron solutions upon limestones; and in metalliferous veins as a gangue mineral. It is relatively common. Siderite is found in Austria, Saxony, the Czech Republic and Slovakia, France, England, Italy, Greenland, Australia, Brazil and Bolivia. In the United States important localities are in Connecticut, Pennsylvania, New Jersey, Ohio, and Washington. It is an iron ore. The mineral was at one time called chalybite.

SIDEWINDER. See **Snakes.**

SIFAKA. See **Lemur.**

SIGMA$_T$ (symbol σ_t). A conveniently abbreviated value of the density of a seawater sample of temperature t and salinity S:

$$\sigma_t = [\rho(S, t) - 1] \times 10^3$$

where $\rho(S,t)$ is the value of the seawater density in c.g.s. units at standard atmospheric pressure. If, for example, $\rho(S,t) = 1.02648$, then $\sigma_t = 26.48$.

SIGMA PARTICLE. A hyperon with a rest-mass energy of about 1193.4 MeV, an isospin quantum number 1, an angular momentum spin quantum number $\frac{1}{2}$, and a strangeness quantum number 1. Symbol, Σ.

SIGMOIDOSCOPY. An instrumental technique for examining the rectum and sigmoid colon. The patient is placed in the knee-chest position, preferably on a motorized table that allows support of the knees below the level of the abdomen. After introduction of the sigmoidoscope, any liquid colonic contents are aspirated. A cotton swab may be pressed against the mucosa and rotated 360 degrees to remove mucus and debris. The physician then searches for ulcerations, a granular mucosal surface, polyps, friability (bleeding), and other conditions, depending upon the exact purpose of the examination. Where information beyond the reach of the sigmoidoscope is required and when this cannot be fully obtained through a barium-contrast x-ray, colonoscopy may be used. This involves an instrument employing fiber optics that permits examination of the colonic surface from the anus to the ileocecal valve.

SIGNAL. 1. An independent input variable. 2. A visual, audible, or other indication used to convey information. 3. The intelligence, message, or effect to be conveyed over a communication system. 4. A signal wave.

SIGNAL CONDITIONING. A process for modifying an input signal prior to introduction into an electronic system, such as a digital-data acquisition or instrumentation operation. The meaning of the term varies from one type of application to the next. Modification of an analog input signal prior to amplification may include: attenuation (scaling), filtering, conversion (current to voltage or voltage to current), impedance-level transformation, bridge or signal compensation, and in numerous instances specialized operations such as cold-junction compensation in the case of thermocouple inputs. Commercially available signal conditioning apparatus may include both amplification and conversion from analog to digital form. Because of the rather nebulous nature of this term, its use in procurement always should be accompanied by detailed specification of functions and operating parameters.

SIGNAL GENERATOR. In the development, calibration, and testing of electronic hardware, instruments, and systems, frequently it is required to simulate certain electrical signals. A signal source provides the stimulus that creates the response to be measured. This source usually is an oscillator or a standard-signal generator of known characteristics which can be adjusted to establish a known set of conditions. These characteristics include: (1) the frequency, (2) the output voltage and impedance, (3) the carrier-signal waveform, which typically may be sine wave, square wave, pulse, or random noise, and (4) the modulation which carries the system information through variation of phase, frequency, amplitude, or timing of the carrier waveform.

Signal sources can be classified functionally as to whether the information that they yield is readily usable in *frequency domain*, or in *time-domain* analysis. Sine wave techniques form the basis of power generation and transmission systems and most communication systems, leading to ready frequency domain analysis. Many developments in information transmission and data handling, such as radar systems, digital computers, telemetry, and wire telegraphy, are based upon pulse techniques which yield most easily to time domain analysis.

Common to all of these systems are ultimate performance limitations, determined by system bandwidths and noise. Bandwidth and transient performance are closely related. They convert one into the other and can be measured as phenomena in either the frequency or the time domain at the convenience of the analyst. Noise is most easily measured by comparison with a noise source of known characteristics.

Types of Signal Sources. These fall into five major categories:

1. *Oscillators*, or sine wave generators which embrace frequencies from 0.01 Hz to 7 GHz, with maximum output levels from a few milliwatts to 200 watts.
2. *Standard signal generators*, which are sine wave oscillators with accurately calibrated output voltage behind a standard impedance and with calibrated modulation capabilities. For wideband measurements, a sweep frequency instrument provides calibrated sweep bands as well as calibrated output. Mechanical sweep devices also are available for converting conventional signal generators and oscillators to sweep generators.
3. *Frequency synthesizers* generate output frequencies continuously adjustable over wide ranges, all coherently derived from a single quartz crystal oscillator. Commercial instruments are available with optional degrees of resolution from one part in 10^3 to one part in 10^9, with either manual or programmable control.
4. *Pulse generators* for time domain measurements. See **Pulse Generator.**
5. *Random-noise generators* that produce wideband noise of known spectrum and energy distribution for noise and vibration testing in mechanical systems, noise measurements in communication circuits, and applications is psychological, probability, and information theory research. These devices are described under **Noise Generator.**

Oscillators

The variable frequency, sine wave oscillator is the basic general purpose signal source. With it one can make a series of measurements at specified frequencies which can be combined to specify performance in the frequency domain. These measurements may be made by manual settings, point-by-point, or by a frequency swept automatically over the desired range to display the system response on a chart recorder or a cathode ray oscilloscope.

Oscillators are of four basic types:

1. *LC Oscillators*. At radio frequencies where tuning can be accomplished with air capacitors, the LC circuit is the best and most economical frequency-determining system. They cover frequencies from 500 kHz to 1,050 MHz. At frequencies above 1,000 MHz, circuits with distributed constants are used.

2. *RC Oscillators*. The frequency is determined by resistive and capacitive elements in this type of instrument. These devices in various configurations cover a frequency range of 0.01 to over 1 MHz.

3. *Beat-frequency Oscillators*. In these devices, the output frequency is the difference between the frequencies of a variable-frequency and a fixed-frequency oscillator. Several decades of frequency can be covered in one band with a single control.

4. *Klystron Oscillators*. In these devices, the frequency is determined by a velocity-modulated electron stream which excites a resonant cavity. In one form, a reflex klystron in a coaxial cavity with a noncontacting plunger is used to cover frequencies from 1.7 to 4.1 GHz. Internal square wave and frequency modulation are provided.

An automatic, audio frequency measuring system that combines a beat-frequency audio generator and a graphic level recorder for automatic plotting of frequency response data is shown in Fig. 1. This type of instrument is used widely for measuring the response of filters, attenuators, networks, loudspeakers, amplifiers, microphones, transducers, and complete acoustic systems.

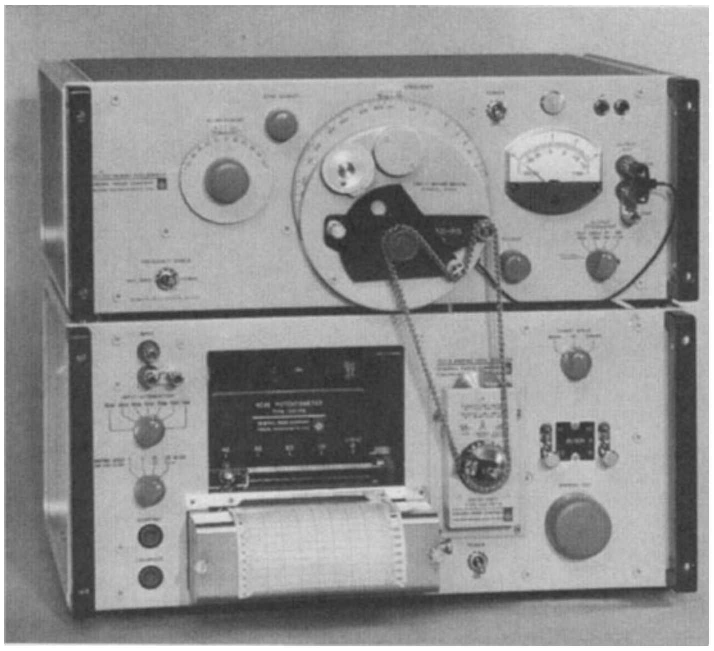


Fig. 1. Automatic audio-frequency measuring system.

Standard Signal Generators

These instruments provide a source of alternating current energy of accurately known characteristics. The carrier, or center frequency, is indicated by a dial setting, the output voltage by a meter reading and associated attenuator setting, and the modulation by a meter reading that is set by appropriate knobs. See Fig. 2.

Common types of modulation signals are sine wave, square wave, and pulse. The output signal either may be frequency or amplitude modulated by these signals. When the frequency modulation system produces a considerable excursion in frequency at a relatively low cyclical rate, the instrument is known as a sweep frequency generator and is particularly useful for automatic data display. Standard signal generators are used for testing radio receivers, as voltage standards over the range from a few microvolts to about one volt, and generally as power

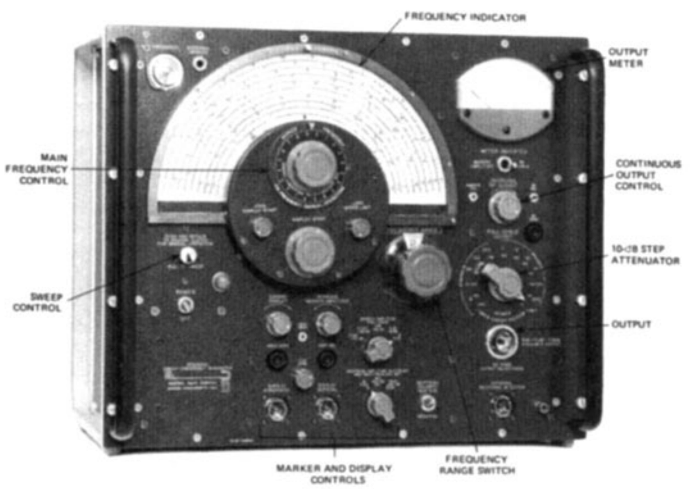


Fig. 2. Standard sweep frequency generator.

sources in measurement of gain, bandwidth, signal-to-noise ratio, standing wave ratio, and other circuit properties.

For use as a standard signal generator, the oscillator must be stable, have reasonably constant output over any one frequency range, have a good waveform, and have no appreciable hum or noise modulation. Careful overall shielding of the generator is essential to minimize stray fields.

The elements of an amplitude modulated standard signal generator are shown in Fig. 3. An amplifier may be added readily at lower frequencies, as shown in Fig. 4, to isolate the oscillator from the load and to minimize the incidental frequency modulation that usually results from amplitude modulation. The elements of a standard sweep frequency generator are shown in Fig. 5.

The specifications for three standard signal generators and for a standard sweep frequency generator are given in Table 1 to illustrate the overall performance parameters of these instruments.

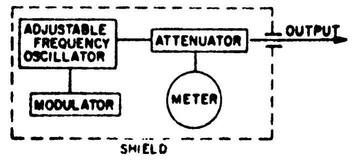


Fig. 3. Elements of a standard signal generator.

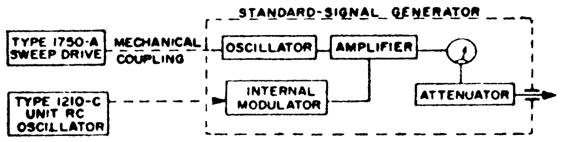


Fig. 4. Elements of an amplitude modulated standard signal generator with amplifier.

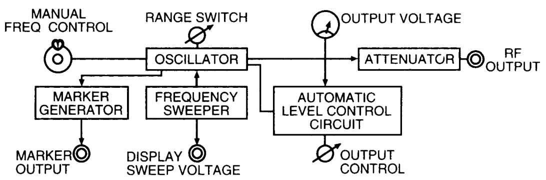


Fig. 5. Elements of a standard sweep frequency generator.

TABLE 1. PERFORMANCE PARAMETERS OF SIGNAL GENERATORS

Instrument Type	Frequency Range	Open-circuit Voltage	Impedance Ohms	Output Modulation %
Standard signal generators	5 kHz-50 MHz	0.1μV-200 mV	10. 50	0-80
	67 kHz-80 MHz	0.1 μV-6 V	50	95
	9.5 MHz-500 MHz	0.1 μV-10 V	50	95
Standard sweep frequency generators	0.7 MHz-230 MHz	0.3 μV-1 V	50	Sweep all bands
	0.45 MHz-10.7 MHz	0.3 μV-1 V	50	

Frequency Synthesizers

These devices are well suited for repeatable, high precision frequency response measurements on amplifiers, filters, transducers, and similar electronic devices. Either point-by-point or sweep techniques may be used. An instrument of this type is shown in Fig. 6 and the overall specifications of four configurations of it are given in Table 2 to illustrate the overall performance parameters obtainable in these kinds of instruments. The output frequency is synthesized directly by repetitive arithmetic manipulations of frequency in a series of identical modules. The synthesizers are equipped for direct, front-panel, digit selection.

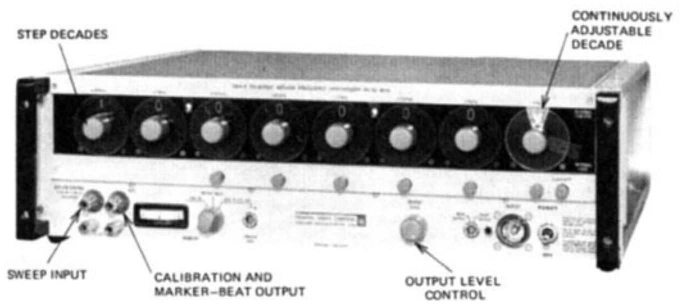

STEP DECADES
CONTINUOUSLY ADJUSTABLE DECADE
SWEEP INPUT
CALIBRATION AND MARKER-BEAT OUTPUT
OUTPUT LEVEL CONTROL

Fig. 6. Coherent decade frequency synthesizer.

Frequency synthesizers combine the advantages of tunable oscillators, which are not usually highly stable, with those of frequency standards, which though very stable, are not tunable. In addition, frequency synthesizers can be tuned like a decade box, i.e., in discrete steps for precision and repeatability; or they can be swept over bands as wide as the instrument's full frequency range, or as narrow as 0.001 Hz.

SIGNAL (Instrument). With reference to industrial and scientific instruments, the Instrument Society of America defines *signal* as information about a variable that can be transmitted.

Actuating-Error Signal. The reference-input signal minus the feedback signal.

Error Signal. In a closed loop, the signal resulting from subtracting a particular return signal from its corresponding input signal.

Feedback Signal. That return signal which results from a measurement of the directly-controlled variable.

Input Signal. A signal applied to a device, element, or system.

Measured Signal. The electrical, mechanical, pneumatic, or other variable applied to the input of a device. It is the analog of the measured variable produced by a transducer (when such is used).

In a thermocouple thermometer, the measured signal is an emf which is the electrical analog of the temperature applied to the thermocouple. In a flowmeter, the measured signal may be a differential pressure which is the analog of the rate of flow through an orifice. In an electric tachometer system, the measured signal may be a voltage which is the electrical analog of the speed of rotation of the part coupled to the tachometer generator.

Output Signal. A signal delivered by a device element, or system.

Reference-Input Signal. A signal external to a control loop which serves as the standard of comparison for the directly-controlled variable.

Return Signal. In a closed loop, the signal resulting from a particular input signal, and transmitted by the loop and to be subtracted from the input signal.

Signal Transducer. A transducer which converts one standardized transmission signal to another.

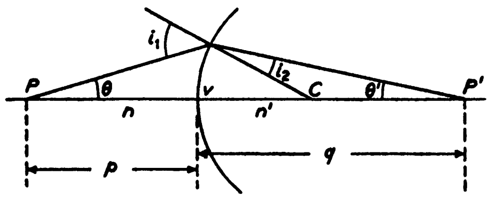
Signal-to-Noise Ratio. Ratio of signal amplitude to noise amplitude.

TABLE 2. PERFORMANCE PARAMETERS OF FREQUENCY SYNTHESIZERS

Characteristic	Example #1	Example #2	Example #3	Example #4
Frequency range	0-100 KHz	0-1 MHz	30 Hz-12 MHz	10 kHz-70 MHz
Smallest digital step	0.01 Hz	0.1 Hz	1 Hz	10 Hz
Smallest direct-calibrated continuously-adjustable decade increments	0.0001 Hz	0.001 Hz	0.01 Hz	0.1 Hz
Maximum bandwidth controllable by continuously-adjustable decade module	100 kHz	1 MHz	1 MHz	1.2 MHz
Spurious frequency outputs:				
Harmonic (a^+ maximum output)	< -40 dB	< -40 dB	< -30 dB	< -30 dB
Nonharmonic	< -80 dB	< -60 dB	< -60 dB	< -60 dB
Output	(Coupling switch at ac) Adjustable, 0 to 2 V, rms (Coupling switch at dc) Adjustable, 0 to 0.8 V, rms 0 to 2 V, rms		(Output impedance switch at 50 ohms) 0 to 2 V, rms (Output impedance switch at zero)	0.2 to 2 V, rms, metered and leveled behind 50 ohms ± 5%

SIGNAL LEVEL. At any point in a transmission system, the difference of the measure of the signal at that point from the measure of an arbitrarily-specified signal chosen as a reference. In audio techniques, the measures of the signal are often expressed in decibels, thus their difference is conveniently expressed as a ratio.

SIGN CONVENTION (Lens and Mirror). Since every distance involved in lens computations must be measured from some origin, a convention of signs should be adopted to insure consistency in the derivation and use of formulas. Unfortunately, this has not been done by all authors. The following probability has the largest following: (1) Draw all figures with the light incident on the surface of reflection or refraction from the left. (2) Consider the object distance $p = PV$ positive when P is at the left of the vertex. (3) Consider the image distance $q = VP'$ positive when P' is at the right of the vertex. (4) Consider the radius of curvature $R = CV$ positive when the center of curvature lies to the right of the vertex. (5) Consider the slope angles positive when the axis must be rotated counterclockwise through less than $\pi/2$ to bring it into coincidence with the ray. (6) Consider angles of incidence and refraction positive when the radius of curvature must be rotated counterclockwise through less than $\pi/2$ to bring it into coincidence with the ray. (7) Consider distances normal to the axis positive when measured upward.



Demonstration of sign convention.

In the accompanying diagram, only θ' is negative. When the convention is followed, the simple thin-lens formula results.

SIGNIFICANCE TESTS. Suppose that a sample provides an estimate t of a parameter θ, and that a certain hypothesis specifies a certain value for θ, $\theta = \theta_0$ say. t will differ from θ_0 by a discrepancy $d = \theta_0 - t$, and it may be possible to deduce from the sampling distribution of t the probability P that a discrepancy as large as d would have arisen if the hypothesis $\theta = \theta_0$ were true. If this probability is small, the sample may be taken to provide evidence against the truth of the hypothesis. This procedure is called a test of significance; if P is less than some value α (commonly chosen to be 0.05 or 0.01), we say that t is significantly different from θ_0 at the level α, or simply that d is significant at this level. The hypothesis $\theta = \theta_0$ (i.e., $d = 0$) is called the null hypothesis.

Alternative tests of the same hypothesis are often available, based on different statistics. To choose between them, we introduce the notion of an alternative hypothesis, $\theta = \theta_1$ say. We may now calculate, for a given α, the probability β of obtaining a significant result at this level when the alternative hypothesis is true. This probability (a function of α and θ_1) is called the power of the test. A few tests can be shown to be at least as powerful as any alternative test for all values of θ_1. Such tests are said to be uniformly most powerful and are clearly to be preferred in cases where they exist.

If we imagine some action being taken based on the result of a significance test, this action may be referred to as accepting or rejecting the null hypothesis. Sampling fluctuations may then lead to two types of incorrect action:

1. We may reject the null hypothesis when it is true.
2. We may accept the null hypothesis when it is false.

These are referred to as errors of the first and second kind; their probabilities are respectively α and $(1 + \beta)$. A test for which $(1 + \beta) \geq \alpha$ for all values of θ_1 is said to be unbiased.

Sir Maurice Kendall, International Statistical Institute, London.

SIGNIFICANT DIGITS. 1. The digits that determine the mantissa of the logarithm of the number beginning with the first digit on the left that is not zero and ending with the last digit on the right that is not zero.

2. The digits of a number that have a significance; the digits of a number beginning with the first nonzero digit on the left side of the decimal point, or with the first digit after the decimal point if there is no nonzero digit to the left of the decimal point, and ending with the last digit to the right. Note that the use of the final zero in the number 0.230 implies that the number is known to third place accuracy.

SIGNIFICANT WAVE. In ocean wave forecasting, a fictitious wave whose height and period are equal to the average height and period of the highest one-third of the actual waves that pass a fixed point.

SIKAS DEER. See **Deer.**

SIKES SCALE. See **Specific Gravity.**

SILAGE. A feedstuff resulting from the anaerobic preservation of moist forage crops or crop residues by acidification. This definition (M. E. McCullough, Georgia Experiment Station, University of Georgia) aptly confines silage to the production of a feedstuff and eliminates the confusion of nonagricultural uses. The definition confines the process to anaerobic conditions and excludes decomposition under aerobic fermentations in which the final result is the disposition of a waste product, such as sewage. The two important inclusions in the definition are the terms *forage* and *acidification.* This confines the definition to harvested farm products that are being stored for feed and limits the method of preservation to acidification. The latter limitation does permit the acidification process to include either or both the direct application of acids or the development of acids through fermentation within the silage mass. Silage must not be confused with simply wet grains in storage. McCullough points out that there are many terms in common use that have no scientific significance in silage terminology (such as *haylage, cornlege, oatledge*, etc.—all coined names used to designate a particular method for harvesting and storing forage). They are like brand names in the feed industry, such as *dairy feed* or *hog supplement*, which have no meaning by themselves. Other terms used in connection with silage, such as *wilted, direct cut*, or *recut*, have value only as being descriptive of the methods used in harvesting or storing the crop.

Advantages and Limitations of Silage. In terms of the dairy farm, there are several advantages to making and feeding silage, including:

1. Silage may be fed at any time of the year or year-round.
2. Silage may be harvested in adverse weather as well as in fair weather.
3. There is very little waste in silage feeding processes.
4. The silage crop can be readily harvested at the most favorable stage of maturity.
5. Silage requires less storage space than hay.
6. Silage crops are removed from the land earlier, thus facilitating a double cropping system.
7. The use of silage increases the carrying capacity of cows per unit of land and per farm by saving more feed per unit of land.
8. The fire hazard in storing silage is minimal.
9. Forage stored as silage is the most economical method of preserving the greatest amounts of nutrients (TDN, total digestible nutrients) per unit of land.
10. Silage is an excellent substitute for pasture.
11. The making and feeding of silage can be completely mechanized.
12. A few weeds will not seriously affect the quality of silage, since most weeds seem to lose off-flavors in the ensiling process. Wild onion is an exception; its flavor is usually retained in the silage.
13. Silage may be kept for many years, almost indefinitely, if it is properly made and well stored.
14. A silo saves feed that might otherwise be wasted.
15. When forage is stored in the form of silage, it has some values that other stored roughages do not possess, including palatability, laxative effect, and the fact that it can be fed free-choice (self-fed) without danger to dairy cattle.

Excellent stand of oats and peas being cut for silage. This crop had been started under irrigation in early spring. On this dairy farm south of Fairbanks, Alaska, silage is cut in the field and hauled to pit silos. Holstein cows are wintered on the silage. (*USDA photo.*)

Limitations of silage include:

1. Initial costs of silo and silage-making machinery are high.
2. After a silo is opened, some silage should be fed daily to prevent spoilage.
3. Sale of silage is generally impractical; hence it must usually be fed on the farm where it is stored.
4. Silage requires some skill and knowledge to be handled successfully, thus keeping spoilage, other losses, and reduction in quality to a minimum.

The accompanying illustration shows cutting of an excellent stand of oats and peas for use as silage.

SILANES. See **Silicon.**

SILICA GLASS. See **Glass.**

SILICATES (Soluble). The most common and commercially used soluble silicates are those of sodium and potassium. Soluble silicates are systems containing varying proportions of an alkali metal or quarternary ammonium ion and silica. The soluble silicates can be produced over a wide range of stoichiometric and nonstoichiometric composition and are distinguished by the ratio of *silica to alkali*. This ratio is generally expressed as the *weight percent ratio* of silica to alkali-metal oxide (SiO_2/M_2O). Particularly with lithium and quaternary ammonium silicates, the molar ratio is used.

Sodium silicates find wide application in many types of detergents and cleaning compounds and have been used for many years as adhesives and cements. Both sodium and potassium silicates are important bonding agents in a large variety of ceramic cement and refractory applications, notably because of their heat stability and resistance to chemicals. Alkalimetal silicate bonds are used in high-temperature ceramic products in the fabrication of electrical components. Soluble silicates find wide application for pelletizing, granulating, and briquetting finely divided particles, such as clays, fertilizers, and ores. Sodium silicates also are used as bonding materials for foundry mold and core compositions. Because of their adherence properties, soluble sodium and potassium silicates are widely used as coatings. Frequently, sodium silicates are used to protect against water-line corrosion in tanks. The ability to form sols and gels is an interesting and very useful characteristic of soluble silicates. Silica gels are used in a major way as desiccants and as carriers for the production of petroleum-cracking catalysts, as well as raw materials in the manufacture of zeolites. Activated sols are used in water clarification.

Generally, sodium and potassium silicates are made by fusion of pure sand with alkali-metal carbonate or alkali-metal sulfate and carbon, this operation carried out in large open-hearth furnaces heated to a temperature range of 1300–1500°C. The resulting glasses may be used in this form, or dissolved in water to produce silicate solutions. Sodium and potassium silicate solutions also can be made by dissolving sand in sodium or potassium hydroxide solution at elevated temperatures and pressures. Lithium silicate glasses, although insoluble in water, can be made by dissolving silica gel in, or mixing silica sols with lithium hydroxide solutions. Anhydrous sodium metasilicate is made from the anhydrous melt. This salt crystallizes rapidly from its aqueous solution at temperatures in the range of 80–85°C.

The most important property of sodium and potassium silicate glasses and hydrated amorphous powders is their solubility in water. The dissolution of vitreous alkali is a two-stage process. In an ion-exchange process between the alkali-metal ions in the glass and the hydrogen ions in the aqueous phase, the aqueous phase becomes alkaline due to the excess of hydroxyl ions produced while a protective layer of silanol groups is formed in the surface of the glass. In the second phase, a nucleophilic depolymerization similar to the base-catalyzed depolymerization of silicate micelles in water takes place.

When sodium silicate solutions of intermediate ratios are concentrated to a thick gum, they become very sticky and tacky. This property is important to many of the adhesive applications. It is related to high cohesion and low surface tension rather than primarily to viscosity.

The stability of soluble silicate solutions depends strongly on pH and concentration. The addition of acids and acid-forming compounds gives rise to the formation of silica gels. Soluble alkali-metal silicate solutions are not compatible with most organic water-miscible solvents. The addition of alcohols and ketones causes phase separation into liquid layers. A few organic systems, however, particularly polyols, such as glycols, glycerins, sugars, and polyethylene glycols, are compatible and miscible with alkali-metal silicate solutions. See also **Adhesives and Glass.**

SILICEOUS. Containing silicon dioxide or one of its compounds.

SILICIC. 1. Containing or pertaining to silicon. 2. Containing silicic acid (ortho) H_4SiO_4; or silicic acid (meta) H_2SiO_3; or silicic acids of a higher degree of hydration (disilicic acids, trisilicic acids, etc.).

SILICIFICATION. An important geochemical process by which certain sedimentary rocks such as limestones and dolomites, or calcareous fossils are partially or entirely replaced by silica, SiO_2. See also **Chert;** and **Flint.**

SILICON. Chemical element, symbol Si, at. no. 14, at. wt. 28.086, periodic table group 14, mp 1408–1,412°C, bp 2,355°C, density 2.242 g/cm^3 (solid crystalline, 20°C), 2.32 g/cm^3 (single crystal, 20°C). Elemental silicon has a face-centered cubic crystal structure (diamond structure). The existence of a hexagonal form of silicon with a wurtzite-type structure and with lattice parameters $a = 3.80$Å and $c = 6.28$Å was established in 1963 (Wentorf-Kasper). Claims to different parameters were made by Jennings-Richman (1976). These differences are discussed by Kasper-Wentorf (1977). Much new knowledge concerning the crystalline structures and phase transitions of silicon has been gained during the mid-1980s, notably from research under immensely high pressures and investigations involving the tunneling microscope, as described shortly.

The common form of silicon is a dark-gray, hard solid. It can be obtained as a brown microcrystalline powder, which is not an allotrope of the gray form. Both forms are unaffected by air at ordinary temperatures, but when heated in air to high temperatures a protective layer of oxide is formed. Silicon reacts with nitrogen at high temperatures to form the nitride; with chlorine to form the chloride, with several metals to form silicides. Crystalline silicon is unattacked by HCl or HNO_3, or H_2SO_4, but is attacked by hydrofluoric acid to form silicon tetrafluoride gas. Silicon is soluble in NaOH solution forming sodium silicate and hydrogen gas. Silicon reacts with dry chlorine to form silicon tetrachloride.

There are three naturally occurring isotopes, ^{28}Si through ^{30}Si, and three radioactive isotopes have been identified, ^{27}Si, ^{31}Si, and ^{32}Si. The latter isotope has a half-life of approximately 700 years, while the half-lives of the other two are short, measured in terms of seconds and hours.

Lavoisier showed in 1787 that SiO_2 was not a single element and indicated that it was the oxide of a hitherto unknown element. In the early 1800s, Scheele, Davy, Gay-Lussac, and Thénard attempted to isolate the element, but were not successful. In 1871, Berzelius discovered silicon in a cast-iron melt and, in 1823, succeeded in isolating the element by reduction of potassium fluorosilicate with potassium. Small laboratory amounts were produced by H.E. Sainte-Claire Deville in 1854 and by C. Winkler in 1864. It was not until 1900 that the effective properties of silicon as a deoxidizing agent for steel production were observed. Shortly thereafter, ferrosilicon alloys, using quartzite, coke, and iron pellets, were produced in electric refining furnaces of the type already in use for making calcium carbide. With this technique, it was possible to produce silicon of about 98% purity. It remained for the rigid purity requirements of semiconductors many years later before silicon of higher purities was produced.

Silicon is ranked second in the order of chemical elements appearing in the earth's crust, an average of 27.72% occurring in igneous rocks. In terms of seawater, it is estimated that a cubic mile of seawater contains about 15,000 tons of silicon (3240 metric tons per cubic kilometer). In terms of abundance throughout the universe, silicon is ranked seventh. First ionization potential 8.149 eV; second, 16.27 eV; third, 33.30 eV; fourth, 44.95 eV. Oxidation potentials $Si + 2H_2O \rightarrow SiO_2 + 4H^+ + 4e^-$, 0.86 V; $Si + 6OH^- \rightarrow SiO_3^{2-} + 3H_2O + 4e^-$, 1.73 V.

Other important physical properties of silicon are given under **Chemical Elements.**

Because of its chemical reactivity, silicon does not occur in elemental form in nature. The element is present in igneous rocks and clays as alumino-silicate; as the oxide SiO_2 in quartz, sand. (Fig 1), flint, and the gems amethyst, jasper, chalcedony, agate, onyx, tridymite, opal, crystobalite; as silicates in zircon (zirconium silicate, $ZrSiO_4$), in willemite (zinc silicate, Zn_2SiO_4), in wollastinite (calcium silicate, $CaSiO_3$), in serpentine (magnesium silicate, $Mg_3Si_2O_7$). Impure (up to 98% Si) silicon is obtained from the oxide (1) by igniting with aluminum powder, or (2) by reduction with carbon in an electric furnace. See also **Cancrinite.**

Silicon Production for Alloys: The production of raw steel requires about 1.6–1.7 kilograms of silicon per metric ton of steel. The silicon is used in the form of ferrosilicon, which contains about 20% silicon. It is estimated that about 3 million metric tons of ferrosilicon are consumed annually in steelmaking. The 20%-silicon-content ferrosilicon can be made in a conventional blast furnace. Ferrosilicons with higher silicon contents (45, 75, 90, and 98%) must be produced in electric furnaces. The raw materials are pure quartzites. The presence of impurities, such as Al_2O_3 and CaO, interfere with the melting process because of the formation of dross. The reducing agent used is chemical coke. For the very high concentrations of silicon (90–98%), ash-free petroleum coke or charcoal are used. Iron is added in the form of small pellets or chips in the production of the 45–75%-silicon alloys.

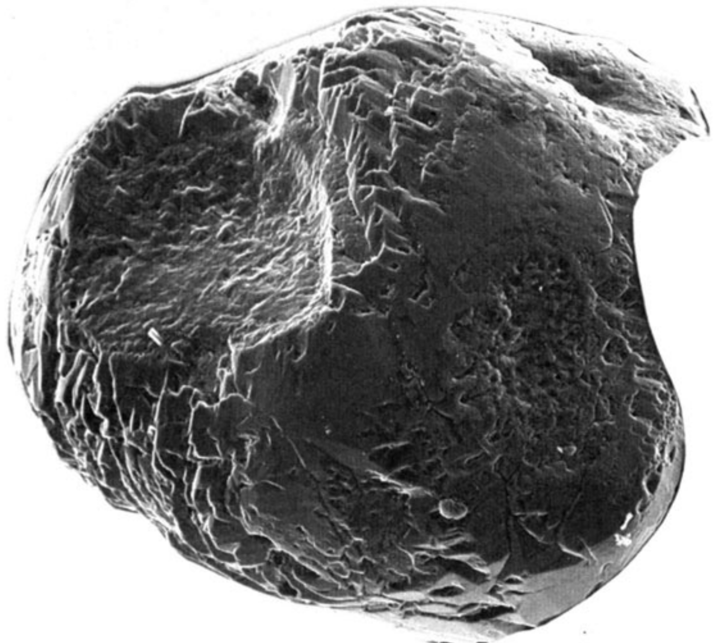


Fig. 1. Grain of sand, originally magnified 100 ×.

For certain metal alloys, a calcium silicon alloy is required. This alloy also is used as a steel deoxidizer and is favored because it forms a low-melting-point calcium silicate product. A representative composition of the alloy is: 30–33% Ca, 60–64% Si, 3–5% Fe, 1–2% Al, 0.3–0.6% C, and less than 0.15% S and P.

Silicon is used in the primary and secondary aluminum industry. The purity of silicon for metallurgical purposes ranges from 96.7 to 98.5% silicon; 0.10 to 0.75% aluminum; 0.03 to 0.04% calcium; with the remainder being principally iron.

Silicon Carbide: This compound is an important industrial abrasive, having a hardness of 9.5 on the Mohs scale. In this compound, each silicon atom is surrounded tetrahedrally by four carbon atoms, and similarly, each carbon atom is surrounded by four silicon atoms. Silicon carbide is made by reducing pure quartz (glass-sand) with petroleum coke in an electric-resistance type furnace, known as the Acheson process. The product is hexagonal crystals ranging from light-green to black. It is used as a ceramic raw material for dross-repellent linings as well as for many abrasive applications.

Silicon carbide also has been recognized for many years because of its having a unique set of electronic material advantages over silicon and gallium arsenide. Not only can SiC withstand higher device operating temperatures (approximately 650°C compared with silicon's 150°C), but SiC devices can operate with ten times the voltage capability and three times the thermal conductance capability and are mechanically much more robust than traditional semiconductors.

The foregoing characteristics enable the configuration of a whole new family of high-power microwave and high-temperature electronics that can withstand high radiation for military and commercial systems.

Only recently has it been possible to produce uniform, centimeter-sized crystals. A high-purity vapor transport growth process has been developed (Westinghouse) to produce 1.5-inch (3.8-cm) device-grade SiC crystals and wafers. These comprised the building blocks for a demonstration of microwave transistors in the early 1990s. See Fig. 2.

Fig. 2. Researcher Dan Barrett (Westinghouse Science & Technology Center) checks the hot (2400°C) crystal growth furnace that he designed for physical vapor transport growth of single crystals of silicon carbide.

Super- or Hyperpure Silicon: For semiconductor use, there can be only one atom of impurity for every 100,000 silicon atoms! The starting material for the manufacture of hyperpure silicon is silicon tetrachloride, $SiCl_4$, or trichlorosilane, $SiCl_3H$. Both of these materials can be reduced with hydrogen to yield a compact deposition of silicon on hot surfaces, ranging from 800–1,200°C. The starting compounds are purified of boron and phosphorus by fractional distillation and absorption techniques. The process hydrogen is purified by passing it through molecular sieves under high pressure, followed by absorption techniques at a low temperature (−190°C). With the highly purified starting ingredients an excess of hydrogen is circulated through heated quartz tubes. Or, the gas mixture may be blown into quartz bell jars, whereupon the silicon is deposited on filaments of tantalum or tungsten or on thin rods of hyperpure silicon, which may be heated by electrical resistance or radiofrequency energy. This process yields polycrystalline rods of silicon which range up to about one meter in length and 150 millimeters in diameter.

To be used in seimconductor devices, the polycrystalline silicon must be converted to single crystals of a defined, predetermined type of conductivity (*n* or *p* type). The crystals must be rigidly controlled as regards their resistivity, and possess the highest degree of crystallographic perfection. The two crystal-growing techniques used are: (1) crucible free vertical float zoning which removes all residual impurities, including phosphorus, arsenic, and oxygen, but boron is essentially irremovable by floating zoning, or (2) crucible pulling in which the crystals, particularly those of lower resistivity, are drawn out of a melt in a process known as the Czochralski technique. Both processes must be conducted under helium or argon, often under a vacuum of 10^{-5} torr.

Production of Ultrapure Silicon Crystal. In 1990, Westinghouse engineers reported the production of the purest crystal of silicon ever made—namely, four times purer than previously reported material. The crystal also is significantly larger, adding to its practicality in the manufacture of microelectronic circuits and devices.

The cylindrical structure, shown in Fig. 3, called a *boule,* weighs 22 pounds (10 kg) and is over a yard (meter) long, with a diameter of just over 3 inches (8 cm). Impurities are a few parts in 100 billion, compared with more than 10 parts in 100 billion previously reported for 1-inch (2.5-cm) diameter ultrapure crystals.

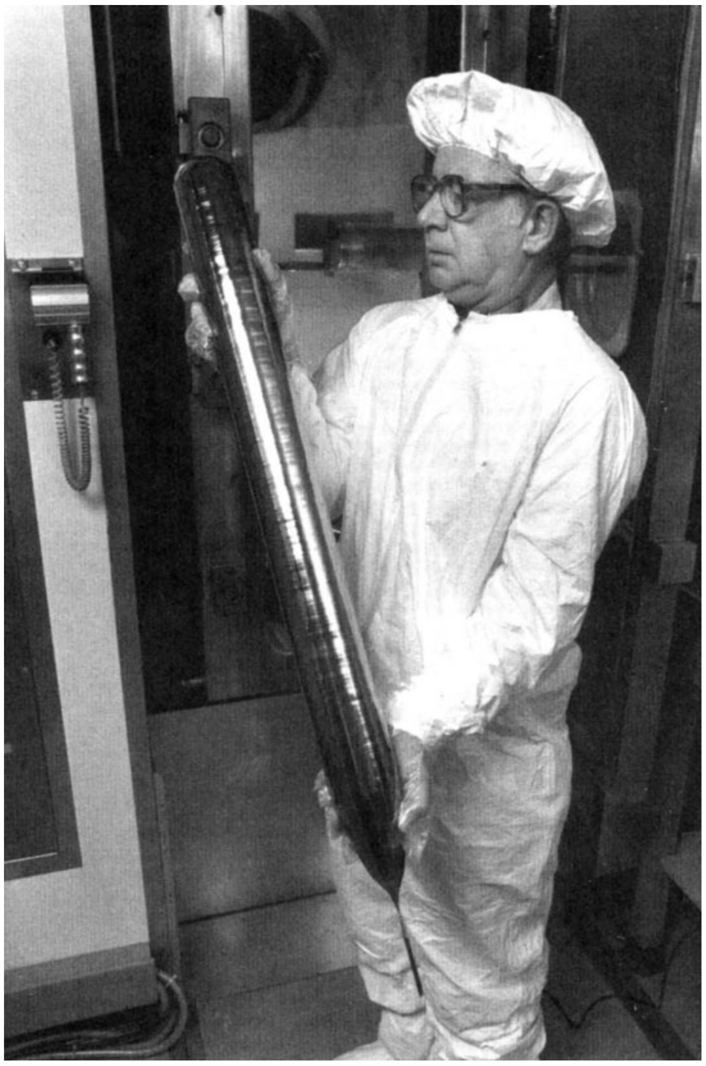


Fig. 3. Ultrapure silicon crystal grown in a float-zone crystal-growth facility, the largest installation of its type in the United States. The furnace can be adapted to produce boules of silicon up to 4 feet (1.2 meters) long and 5 inches (13 cm) in diameter. The boule shown has an impurity content of less than a few parts in a total of 100 billion. (*Source: Westinghouse Technology*.)

Crystal boules are sliced into wafers on which microelectronic circuits and power semiconductor devices are fabricated. An important use of the wafers is for infrared dectors for space, defense, and environmental applications.

Liquid-Solution Synthesis of Silicon Crystals. In late 1992, J. R. Heath (IBM Watson Research Laboratory) reported on a liquid-solution phase technique for preparing submicrometer-sized silicon single crystals. The synthesis is based on the reduction of $SiCl_4$ and $RSiCl_3$ (R = H, octyl) by sodium metal in a nonpolar organic solvent at high temperatures (385°C) and high pressure (above 100 atmospheres). For R = H, the synthesis produces hexagonal silicon single crystals ranging from 5 to 2000 nanometers. For R = octyl, the synthesis also produces hexagonal-shaped silicon single crystals.

Light Emission from Silicon. Because of silcon's successes in the electronic components field, research has been going on to find a form of Si that will produce luminous radiation. Because of former failures, numbers of scientists have given up this research. However, independently in French and British laboratories during 1990, some success has been achieved. These researchers have found that, if one etches Si into structures so tiny that the electronic behavior of the material is transformed, full-color emission from what are termed "silicon quantum wires." A British researcher L. Canham (Royal Signals and Radar Establishment, Malvern, England) has observed that, to make silicon quantum wires, a process for sculpting silicon, known for some 30 years, is the basis. A silicon wafer is immersed in an acid electrochemical bath, which bores into the disk to produce extremely small so-called "wormholes." The latter are etched chemically, enlarging them until they meet one another. The result is a columnar structure of silicon. The latter are about a micron high, and 50 of them, stacked end to end, would span an area about equivalent to a cross-section of a human hair and are only a few nanometers thick ($\frac{1}{15,000}$), smaller than a hair. The researchers have found that, when such a structure is bathed in ultraviolet light, light emission occurs. The emitted wavelength is determined by the porosity of the Si layer.

J. P. Harbison (Bellcore, Redbank, New Jersey) observes, "This is not the moment of the breakthrough for light-emitting silicon, but it is the moment when a lot of people are realizing its potential."

See also **Crystal;** and **Semiconductors.**

Research on Silicon Structure and Surface Properties. The rather unusual properties of silicon have intrigued scientists for many years. It possesses the physical properties of a *metalloid* (exhibits properties of both a metal and nonmetal). In several ways, silicon resembles germanium and, to a lesser extent, it resembles arsenic and boron. Silicon is a semiconductor of electricity, the conductivity rising with temperature. Silicon, in pure form, is intrinsically a semiconductor. The presence of impurities in very minute amounts markedly increases its conductivity. By introducing elements of group 13, such as boron, which have a deficiency of electrons, the *p*-type silicon results, in which electricity is conducted by migration of electron vacancies or holes. On the other hand, introduction of elements of group 15, such as arsenic or phosphorus, in which there is no deficiency of electrons, the *n*-type silicon results, in which extra electrons carry an increased current because of their migration. Scientists have not been satisfied with oversimplified explanations such as that just given. As a key material in the microelectronics field, where the processing of silicon into chips and other configurations for electronic components is essentially effected at the surface of the silicon, particular interest concerns those crystalline structural details that play a role in the electronic nature of the element. However, prior to the emergence of solid-state technology, scientists were puzzled by what appeared to be crystal structure and surface anomalies and, consequently, research dates back many years, with progress largely determined by the instrumentation available to investigators. The *tunneling microscope*, the invention of which is accredited to G. K. Binnig and H. Rohrer and partially to E. W. Müller (who also invented the field-ion microscope in the 1950s), has contributed much (as of the mid-1980s) toward an understanding of the surface of the silicon crystal, an understanding which is expected to be translated ultimately in manufacturing improvements and better final properties of silicon-based electronic components. See **Scanning Tunneling Microscope.**

The relatively recent availability of means to create extremely high pressures (See **Diamond Anvil High Pressure Cell**) has made it possible to gain further insights into the character of silicon. It has been learned from such experimentation that at a pressure of 110 kilobars (about 1.6 million pounds per square inch), silicon enters *truly metallic* phases. At the pressure stated, silicon abruptly assumes a structure which is similar to the beta form of tin. At this pressure and at a temperature of 6 degrees Kelvin (six Celsius degrees above absolute zero) the metal becomes superconducting, that is, it offers no resistance to the passage of electrons. At a pressure of 130 kilobars, the beta-tin form of silicon transforms into what has been designated as the *primitive hexagonal* phase, a phase first discovered in 1984. This research was conducted by Cohen, Chang, and Dacorogna (University of California, Berkeley). Prior to this experiment, it was not thought that such a phase would exist in the crystal of any chemical element. The researchers entered into theoretical calculations after the experiment to better understand the properties of the primitive hexagonal phase. The calculations were lengthy and required a CRAY/X-MP computer. A major finding was that the bonds linking the atoms in each of the planes defined by the hexagons should be weaker than the bonds linking the atoms in adjacent planes—this indicating that the electronic charge distribution should be inhomogeneous along one dimension. This inhomogeneity is an indication of a good superconductor because, in effect, it provides corridors through which electrons can move.

In their investigation, the researchers turned back to the much earlier hypotheses of quantum mechanics, including the properties of phonons (in quantum mechanics, a phonon can be treated as a particle, one that interacts with electrons). A strong interaction improves the opportunities for superconductivity. However, where the coupling is too strong, the integrity of the lattice may collapse and a structural phase transition may occur. Testing for superconductivity at such high pressures will be difficult. The Berkeley group predicts that the superconducting temperature will rise to a value greater than 10 degrees as the pressure nears the value required for the transition from the simple hexagonal phase to the hexagonal closed-packed phase. If expectations are proved, silicon could be the best superconductor of all chemical elements. Translating this to practical application, of course, may or may not be feasible at some future date. See also **Superconductivity.**

Silicon Chemistry and Compounds

Like carbon, silicon forms chiefly covalent bonds, but its greater atomic radius enables it to form positive ions more readily. Unlike carbon and tin, silicon is not allotropic, having only one elemental form, the diamond structure in which each atom is surrounded tetrahedrally by four others to which it is covalently bonded. An apparently amorphous brown powder, produced by combustion of silane, SiH_4, has been found to be a microcrystalline variety of this covalently-bonded structure. Much research has been conducted during the 1970s and 1980s pertaining to the more exotic silicon compounds, such as the *disilenes*, and to silicon-mediated organic synthesis. These topics are discussed later in this article. The following several paragraphs are devoted to the large number of traditional silicon compounds whose constitution and characteristics have been well established over the years.

Silicon dioxide, (SiO_2): This compound exists in at least eleven distinct crystalline forms. Several of them are obtained by heating α-quartz, which has a number of transition points, to produce β-quartz, and to give various forms of tridymite and crystobalite. The unit of structure is the tetrahedron in which each silicon atom is covalently bonded to four oxygen atoms, and the variation is in the ways these tetrahedra are interconnected (by oxygen atoms) to form a three-dimensional system.

Silicon dioxide is converted by hydrofluoric acid into silicon tetrafluoride, SiF_4, a gas. SiF_4 can also be produced directly from the elements, as can the other tetrahalides, silicon tetrachloride, $SiCl_4$ (a liquid), silicon tetrabromide, $SiBr_4$ (a liquid), and silicon tetraiodide, SiI_4 (a solid). The silicon halides hydrolyze much more readily than the carbon halides, because the unoccupied silicon 3*d* orbitals are energetically not far above its 3*s* and 3*p* orbitals. This fact also permits the formation of the sp^3d^2 hybrid bonds of the fluorosilicate ion, SiF_6^{2-}, and additional compounds of the halides, e.g., $SiX_4 \cdot 2$ pyridine. Silicon also is intermediate between carbon and the higher members of main group 4 of the periodic table in forming a dichloride, $SiCl_3$, by strong heating of silicon with silicon tetrachloride.

Quartz and other forms of silica react very slightly with water to form monosilicic acid, $(SiO_2)_n + 2nH_2O \rightarrow nSi(OH)_4$. As shown, this reaction is a depolymerization followed by a hydrolysis, and proceeds rapidly with hot alkalis or fused alkali metal carbonates, yielding soluble silicates containing the SiO_4^{4-} and $(SiO_3^{2-})_n$ ions. The hydrolysis reaction is geologically important, because it is considered to be the starting point in the formation of the innumerable silicate minerals that occur so widely in nature, just as many of the silica minerals may have originated by the reverse reaction. Many of the more complex silicic acids are considered to form by polymerization of $Si(OH)_4$ molecules by sharing of —OH ions between two silicon ions (octahedrally coordinated by six hydroxyl ions) followed by condensation with the loss of water to produce $-\overset{|}{\underset{|}{Si}}-O-\overset{|}{\underset{|}{Si}}-$ linkages. The polymerization of silicic acid is carried out industrially to produce silica gel, a stable sol of colloidal particles. The various methods involve careful removal of H_2O, the catalytic effect of acid or alkali (or fluoride ion) and controlled pH. Many varieties of silica gel have been made, including the zerogels and aerogels, in which the aqueous phase is displaced by a gaseous one.

In 1992, Yeganeh-Haeri, Weidner, and Parise (Center for High Pressure Research, State University of New York, Stony Brook) used laser Brillouin spectroscopy to determine the adiabatic single-crystal elastic stiffness coefficients of silicon dioxide in the alpha-cristobalite structure. This SiO_2 polymorph, unlike other silicas and silicates, was found to exhibit a negative Poisson's ratio. Alpha-cristobalite contracts laterally when compressed and expands laterally when stretched. Tensorial analysis of the elastic coefficients showed that Poisson's ratio reached a maximum value of -0.5 in some directions, whereas averaged values for the single-phased aggregate yielded a Poisson's ratio of -0.16.

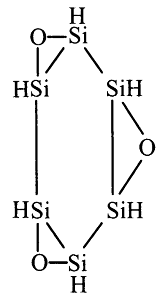
Silicon Dioxide as a Chemical Intermediate. In 1992, R. M. Laine (University of Michigan, Ann Arbor) announced the development of a process that transforms sand and other forms of silica into reactive silicates that can be used to synthesize unusual silicon-based chemicals, polymers, glasses, and ceramics. The Laine procedure produces pentacoordinate silicates directly from low-cost raw materials—silicon dioxide, ethylene glycol, and an alkali base. The mixture is approximately a 60:1 ratio of silica gel, fused silica (or sand) to metal hydroxide and ethylene glycol. Heating the mixture slowly, the ethylene glycol and water (used to put the materials in solution) boils off. The resulting glycolatosilicates, unlike the hexa- and tetracoordinate forms, are reactive and offer potential for synthesizing a wide range of materials. Laine observes, "The new silicon chemistry could produce alternatives to many petrochemical-based products and could be competitive with or superior to present carbon-based materials." Thus far, a number of materials have been produced by the process:

1. A clear polymer capable of conducting electric current when spread in a thin layer across a flat surface. Potential includes applications in batteries, heated windshields, and electrochromic windshields.
2. A fire retardant polymer that is easily impregnated into wood to "petrify" the material, making it stronger and nonflammable.
3. Liquid-crystal polymers stable to about 425°C (800°F), with potential for uses in watch displays and aerospace instrumentation.
4. Silicate glasses capable of withstanding high temperatures.

Silicates: The great number of naturally occurring silicates result, as just indicated, from the polymerization and dehydration of monosilicic acid to form, ultimately, such groups and ions as $(Si_2O_7)^{6-}$, $(Si_3O_9)^{6-}$, $(Si_4O_{12})^{8-}$, and $(Si_6O_{18})^{12-}$. Various cations, such as those of boron, B^{3+}, aluminum, Al^{3+}, etc., in the structure lie at the centers of anionic polyhedra having as anions the O^{2-} ions of neighboring SiO_4 tetrahedra, in which each Si—O bond has an electrostatic bond strength of 1. Cations of lower charge density, on the other hand, like sodium, Na^+, potassium, K^+, calcium, Ca^{2+}, etc., are located interstitially. The great variety of the silicates is due to the considerable degree of isomorphism, exhibited not only by elements of the same group, but by elements of different groups, whereby they partly replace each other in the complex silicates, and by no means necessarily in stoichiometric proportions. Thus troosite may be represented by the formula $(Zn, Mn)_2SiO_4$, chrysolite by $6Mg_2SiO_4 \cdot Fe_2SiO_4$, and vermiculite by $(Mg, Fe)_3(AlSi)_4O_{10} \cdot (OH)_2 \cdot 4H_2O$, even the silicon in vermiculite being partly replaced (by aluminum). One plane of classification of the silicates is upon the basis of the linking of the SiO_4 tetrahedra:

A. Discrete silicate radicals.
   1. Single tetrahedral (SiO_4^{4-}), e.g., phenacite, Be_2SiO_4.
   2. Two tetrahedra $(Si_2O_7)^{6-}$, e.g., hardystonite, $Ca_2ZnSi_2O_7$.
   3. Three tetrahedra $(Si_3O_9)^{6-}$, e.g., benitoite, $BaTiSi_3O_9$.
   4. Four tetrahedra $(Si_4O_{12})^{18-}$, e.g., axinite, (Fe, Mn) $Ca_2Al_2BO_3Si_4O_{12}$.
   5. Six tetrahedra $(Si_6O_{18})^{12-}$, e.g., beryl, $Be_3Al_2Si_6O_{18}$.
B. Silicon-oxygen chains of indefinite length.
   1. Single chains with one silicon atom to three oxygen atoms, e.g., diopside, $CaMg(SiO_3)_2$.
   2. Double chains (Si:O = 4:11), e.g., tremolite, $Ca_2Mg_5(Si_4O_{11})_2(OH)_2$.
C. Silicon-oxygen sheets. (Si:O = 2:5), e.g., talc, $Mg_3(SiO_5)_2(OH)_2$.
D. Silicon-oxygen spatial networks.
   1. Composition SiO_2 (composed of interlinked SiO_4 tetrahedral), e.g., quartz, SiO_2.
   2. Composition $M_n(Si, Al)_nO_{2n}$, e.g., feldspar, KSi_3AlO_8. These are probably based upon silicon and aluminum tetrahedra, variously linked.

Silanes: The increasingly large number of silicon compounds produced by industrial processes may be systematized about the silanes and their substitution products, just as the silicates are about the SiO_4 tetrahedron. Silicon, like carbon, forms a number of hydrides, though their number is much more limited. The silane series, analogous to the paraffin hydrocarbons, has at least six members, silane (SiH_4), disilane ($H_3Si—SiH_3$), . . . hexasilane ($H_3Si—SiH_2—SiH_2—SiH_2—SiH_2—SiH_3$). They are increasingly unstable, hexasilane dissociating at room temperature. They are halogenated with free halogens to form substituted silanes, and catalytically with the hydrogen halides. The halosilanes react with NH_3 to form silylamines or silazanes and are hydrolyzed by water to form siloxanes. Prosiloxane, H_2SiO, polymerizes readily but disiloxane, $H_3Si—O—SiO_3$, and the higher siloxanes, although they polymerize, can readily be studied. They have properties like the ethers and other analogous carbon compounds. Hydrogen-containing siloxanes, such as $HO_2Si—SiO_2H$ are also known and polymerize readily. There are also ring siloxanes, such as siloxen, which has a polymerized structure of epoxy form (a powerful reducing agent)



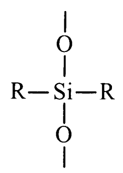
Silyl and polysilyl radicals also combine with nitrogen, arsenic, and other main group 5 elements, as with sulfur and selenium.

The silazanes are such compounds of silicon, nitrogen and hydrogen of the general formula $H_3Si(NHSiH_2)nNHSiH_3$, being called disilazane, trisilazane, etc., according to the number of silicon atoms present. (In disilazane, n in the above formula has a value of 0, in trisilazane it is 1, etc.)

The silthianes are sulfur compounds having the general formula $H_3Si(SSiH_2)nSSiH_3$ which are called disilthiane, trisilthiane, etc., according to the number of silicon atoms present. They have the generic name *silthianes*. (In disilthiane, n in the above formula has a value of 0, in trisilthiane, a value of 1, etc.)

Silicones: These are semiorganic polymers with a quartzlike structure in which various organic groups are attached to the silicon atom. By varying the kind and number of organic groups, a variety of materials ranging from liquids through gels and elastomers to rigid solids (resins) can be produced.

The organosilicon compounds may be regarded as substituted silanes, although of course their preparation is not usually in this way. Thus, ethyl silicate, $Si(OC_2H_5)_4$, is prepared from silicon tetrachloride and ethyl alcohol, and tetraethyl silane, $Si(C_2H_5)_4$, is prepared from silicon tetrachloride and diethylzinc. The silicon-carbon bond, unlike the carbon-carbon bond, has about 12% of ionic character, varying somewhat with the atoms or groups attached to the two atoms. Other types of organosilicon compounds include the esters, the alkoxyhalosilanes, the higher tetra-alkylsilanes (prepared from silicon tetrachloride and Grignard reagents), the alkylsilanes (H partly replaced by R), the alkylhalosilanes, the alkylalkoxysilanes, the alkylsilylamines, some aryl compounds of the foregoing types, and many related derivatives of disilane and the polysilanes. Other types of compounds are those having silicon-carbon chains and the organosiloxane compounds, for which the name "silicones" is often used. These are essentially chains or networks of groups



joined by oxygen atoms attached to the silicon atoms as shown. There are many other groups of silicon compounds, as well as individual ones.

Aluminates: Many complex silico-aluminates or aluminosilicates are formed in nature. Of these, clay in more or less pure form (pure clay, kaolinite; kaolin, china clay, $H_4Si_2Al_2O_9$ or $Al_2O_3 \cdot 2SiO_2 \cdot 2H_2O$) is of great importance. Clay is formed by the weathering of igneous rocks, and is used in the manufacture of bricks, pottery, porcelain, Portland cement. Sodium aluminosilicate is used in water purification to remove dissolved calcium compounds.

Fluosilicate: Sodium fluosilicate, Na_2SiF_6, white solid slightly soluble; magnesium fluosilicate, $MgSiF_6$, white solid, soluble.

Sulfides: Silicon monosulfide, SiS, yellow solid, somewhat volatile, formed by heating to redness crystalline silicon in sulfur vapor, reactive with water; silicon disulfide, SiS_2, white crystals, formed by heating amorphous silicon and sulfur, and then subliming, reactive with water.

Nitrides: Trisilicon tetranitride, Si_3N_4, by heating silicon oxide plus carbon to 1,500°C in a current of nitrogen gas.

Silicates. See **Adhesives**.

Silicon-Silicon Double Bond: Of the chemical elements, Si is closest to carbon in terms of its chemical properties. Multiple bonds pervade carbon chemistry and thus it is no surprise that investigators, over a period of many years, have been seeking evidence of multiple bonding in silicon. As early as 1911, Kipping reported compounds exhibiting this bonding, but these substances were later shown to be polymers or cyclic oligomers. It was not until the 1960s that good evidence was reported for the existence of Si=C (silene), Si=Si (disilene), and Si=O (silanone) compounds. The full reality of such compounds, however, was not not reported until 1981. At that time, a silene and a disilene, each of which is stable at room temperature, were reported by two separate groups. Brook, et al. reported on a silene; West, et al. reported on a disilene. It has since been concluded that many disilenes can be prepared, including compounds that are unexpectedly stable. Molecules containing Si=Si bonds can be synthesized by several routes. The key to stabilization of these compounds is to provide large substituents bonded to the Si atoms so that polymerization is blocked. It has been determined that disilenes react chemically by addition across the double bond, as do alkenes. Tetramesityldisilene, as reported by West, also undergoes a wide variety of addition reactions previously unestablished in organic chemistry. The result is several "new" and unusual types of molecules, the details of which are reported in the West paper listed under references.

Silicon-Mediated Organic Synthesis: As reported by Paquette (reference listed), since the late 1960s, organic chemists have used the chemical properties of tetracovalent silicon to achieve a variety of new synthetic transformations. Paquette (Ohio State University) summarizes, "In carbon-functional silanes, exceptional stabilization is provided to a carbocation center in the beta position when the carbon-silicon bond lies in plane. This phenomenon directs electrophilic attack to the silicon-substituted carbon in aryl-, vinyl-, and alkynylsilanes and to carbon-3 in allylsilanes. For different reasons, silicon also stabilizes a carbon-metal bond in the alpha position. Consequently, access to many silicon-containing organometallics is readily available. The exceptional strength of silicon-oxygen and silicon-fluorine bonds is yet another factor that controls the chemical reactivity of silicon reagents. In recent developments, preparative chemists have taken advantage of these properties in imaginative and useful ways." In the Paquette paper, these observations are developed in exceptional and illustrated detail.

Reactions of Elemental Silicon. In the late 1980s, E. A. Pugar and P. E. D. Morgan and a team of researchers at the Rockwell International Science Center, Thousand Oaks, California, conducted a thorough effort to understand "Low Temperature Direct Reactions Between Elemental Silicon and Liquid Ammonia or Amines for Ceramics and Chemical Intermediates." Details are given in reference cited.

Because of the important potential applications of silicon nitride, the use of low-cost starting materials, such as elemental silicon and liquid ammonia or amines, may be more effective than the existing chloride method. In earlier work, this process was found to form silicon diimide ($Si(NH)_2$), but required purification steps to remove chloride.

Pugar and Morgan elucidate their research and include a summary of the work of other researchers over the years. The report concludes: "Through the use of modern sensitive probes, direct elemental silicon reactions with liquid ammonia, silicon-hydrazine and silicon-organic amines have been discovered. The reaction of elemental silicon with nitrogen-containing reagents, under rather benign conditions, can produce ceramic precursors and with further chemical treatments can produce fibers, films, and other commercial and industrial products."

Nomenclature of Silicon Compounds: The name of the compound SiH_4 is *silane.* Compounds having the general formula $H_3Si \cdot [SiH_2]n \cdot SiH_3$ are called disilane, trisilane, etc., according to the number of silicon atoms present. Compounds of the general formula Si_nH_{2n+2} have the generic name *silanes*. Example: Trisilane, $H_3Si \cdot SiH_2 \cdot SiH_3$

Compounds having the formula $H_3Si \cdot [NH \cdot SiH_2]_n \cdot NH \cdot SiH_3$ are called disilazane, trisilazane, etc., according to the number of silicon atoms present; they have the generic name *silazanes*. Example: Trisilazaine, $H_3Si \cdot NH \cdot SiH_2 \cdot NH \cdot SiH_3$

Compounds having the formula $H_3Si \cdot [S \cdot SiH_2]_n \cdot S \cdot SiH_3$ are called disilthiane, trisilthiane, etc., according to the number of silicon atoms present; they have the generic name *silthianes*. Example: Trisilthiane, $H_3Si \cdot S \cdot SiH_2 \cdot S \cdot SiH_3$

Compounds having the formula $H_3Si \cdot [O \cdot SiH_2]_nO \cdot SiH_3$ are called disiloxane, trisiloxane, etc., according to the number of silicon atoms present; they have the generic name *siloxanes*. Example: Trisiloxane, $H_3Si \cdot O \cdot SiH_2 \cdot O \cdot SiH_3$

For designating the positions of substituents on compounds named as silanes, silazanes, silthianes, and siloxanes, each member of the fundamental chain is numbered from one terminal silicon atom to the other. When two or more possibilities for numbering occur, the same principles are followed as for carbon compounds. Examples:

1-Butyl-2,3-dichloro-2pentyltrisilane
$Cl \cdot SiH_2 \cdot SiCl(C_5H_{11}) \cdot SiH_2 \cdot C_4H_9$
2-Methyl-3-pentyloxytrisilazane
$SiH_3 \cdot N(CH_3) \cdot SiH(OC_5H_{11}) \cdot$
$NH \cdot SiH_3$
1-Methoxytrisiloxane
$CH_3O \cdot SiH_2 \cdot O \cdot SiH_2 \cdot O \cdot SiH_3$

The names of representative radicals containing silicon are shown below. These illustrate the principles on which any further radical names should be formed.

Silicon, hydrogen

silyl	H_3Si—
silylene	H_3Si=
silylidyne	$HSi \equiv$
disilanyl	$H_3Si \cdot SiH_2$—
trisilanyl	$H_3Si \cdot SiH_2 \cdot SiH_2$—
disilanylene	—$SiH_2 \cdot SiH_2$—
trisilanylene	—$SiH_2 \cdot SiH_2 \cdot SiH_2$—
cyclohexasilanyl	$SiH_2 \cdot SiH_2 \cdot SiH_2$ | | $SiH_2 \cdot SiH_2 \cdot SiH_2$

Silicon, hydrogen, oxygen

siloxy	$H_3Si \cdot O$—
disiloxanyl	$H_3Si \cdot O \cdot SiH_2$—
disilanoxy	$H_3Si \cdot SiH_2 \cdot O$—
disiloxanoxy	$H_3Si \cdot O \cdot SiH_2 \cdot O$—

Silicon, hydrogen, sulfur

silylthio	$H_3Si \cdot S$—
disilanylthio	$H_3Si \cdot SiH_2 \cdot S$—

disilthianyl	$H_3Si{\cdot}S{\cdot}SiH_2$—
disilthianylthio	$H_3Si{\cdot}S{\cdot}SiH_2{\cdot}S$—

Silicon, hydrogen, sulfur, oxygen

disilthianoxy	$H_3Si{\cdot}S{\cdot}SiH_2{\cdot}O$—
disiloxanylthio	$H_3Si{\cdot}O{\cdot}SiH_2{\cdot}S$—

Silicon, hydrogen, nitrogen

silylamino	$H_3Si{\cdot}NH$—
disilanylamino	$H_3Si{\cdot}SiH_2{\cdot}NH$—
disilazanyl	$H_3Si{\cdot}NH{\cdot}SiH_2$—
disilazanylamino	$H_3Si{\cdot}NH{\cdot}SiH_2{\cdot}NH$—

Silicon, hydrogen, nitrogen, oxygen

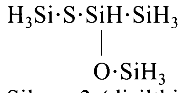
disilazanoxy	$H_3Si{\cdot}NH{\cdot}SiH_2{\cdot}O$—
disiloxanylamino	$H_3Si{\cdot}O{\cdot}SiH_2{\cdot}NH$—

Compound radical names may be formed in the usual manner. Examples:

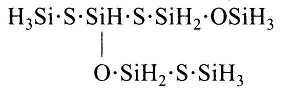
Silyldisilanyl	$(H_2Si)_2SiH$—
Disilyldisilanyl	$(H_3Si)_3Si$—
Triphenylsilyl	$(C_6H_5)_3Si$—

Open-chain compounds which have the requirements for more than one of the structures already defined are named, if possible, in terms of silane, silazane, silthiane, or siloxane containing the largest number of silicon atoms. Examples:

3-Siloxytrisilthiane



1-Siloxy-3-(disilthianoxy)trisilthiane

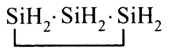


When there is a choice between two parent compounds possessing the same number of silicon atoms, the order of precedence is siloxanes, silthianes, silazanes, and silanes. Examples:

1-Silylthiodisiloxane
$SiH_3{\cdot}O{\cdot}SiH_2{\cdot}S{\cdot}SiH_3$
1-Silylaminodisilthiane
$SiH_3{\cdot}S{\cdot}SiH_2{\cdot}NH{\cdot}SiH_3$
1-Phenyl-3-silyldisiloxane
$SiH_3{\cdot}SiH_2{\cdot}O{\cdot}SiH_2{\cdot}C_6H_5$

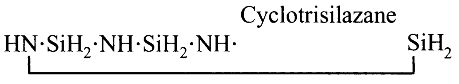
Cyclic silicon compounds having the formula $[SiH_2]_n$ are called cyclotrisilane, cyclotetrasilane, etc., according to the number of members in the ring; they have the generic name *cyclosilanes*. Example:

Cyclotrisilane



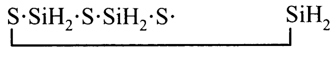
Cyclic compounds having the formula $[SiH_2{\cdot}NH]_n$ are called cyclodisilazane, cyclotrisilazane, etc., according to the number of silicon atoms in the ring. They have the generic name *cyclosilazanes*. Example:

Cyclotrisilazane



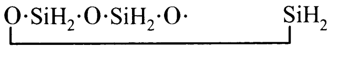
Cyclic compounds having the formula $[SiH_2{\cdot}S]_n$ have the generic name *cyclosilthianes* and are named similarly to the cyclosilazanes. Example:

Cyclotrisilthiane



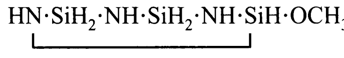
Cyclic compounds having the formula $[SiH_2{\cdot}O]_n$ have the generic name *cyclosiloxanes* and are named similarly to the cyclosilazanes. Example:

Cyclotrisiloxane

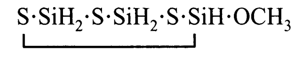


Cyclosilanes, cyclosilazanes, cyclosilthianes, and cyclosiloxanes are numbered in the same way as carbon compounds of similar nature. Examples:

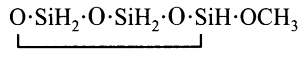
2-Methoxycyclotrisilazane



2-Methoxycyclotrisilthiane

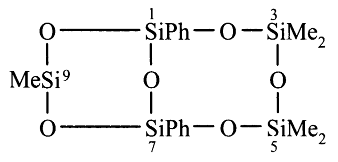


2-Methoxycyclotrisiloxane

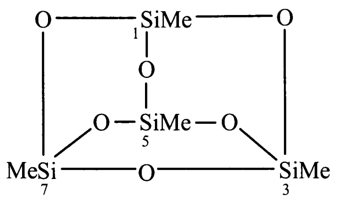


Polycyclic siloxanes (polycyclic compounds whose members consist entirely of alternating silicon and oxygen atoms) are named as bicyclosiloxanes, tricyclosiloxanes, etc., or as spirosiloxanes, and are numbered according to methods in use for carbon compounds of similar nature. Polycyclic silthianes, silazanes, and silanes are treated similarly. Examples:

3,3,5,5,9,9-Hexamethyl-1,7-diphenylbicyclo[5,3,1]pentasiloxane



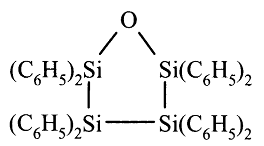
Tetramethyltricyclo[3,3,1,1]tetrasiloxane



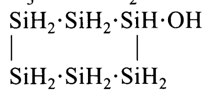
The names of compounds containing silicon atoms as heteromembers (with or without other heteromembers) but not classifiable as (linear or cyclic) silanes, silazanes, silthianes or siloxanes are derived with the aid of the oxa-aza convention. Examples:

2,2,4,4,6,6-Hexamethyl-2,4,6-trisilaheptane
$(CH_3)_3Si{\cdot}CH_2{\cdot}Si(CH_3)_2{\cdot}CH_2{\cdot}Si(CH_3)_3$
2,4,6,8,-Tetraoxa-5-carbanosilane
$SiH_3{\cdot}O{\cdot}SiH_2{\cdot}O{\cdot}CH_2{\cdot}O{\cdot}SiH_2{\cdot}O{\cdot}SiH_3$

Octaphenyloxacyclopentasilane

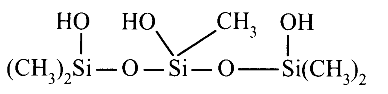


Hydroxy-derivatives in which the hydroxyl groups are attached to a silicon atom are named by adding the suffixes *ol*, *diol*, *triol*, etc., to the name of the parent compound. Examples:

Silanol	$H_3Si{\cdot}OH$
Silanediol	$H_2Si(OH)_2$
Silanetriol	$HSi(OH)_3$
Disilanehexaol	$(HO)_3Si{\cdot}Si(OH)_3$
Disiloxanol	$H_3Si{\cdot}O{\cdot}SiH_2{\cdot}OH$
Cyclohexasilanol	

Polyhydroxy-derivatives in which hydroxyl group is attached to a silicon atom are named wherever possible in accordance with the principle of treating like things alike. Example:

1,1,3,5,5-Pentamethyltrisiloxane-1,3,5-triol



Otherwise they are named in accordance with the principle of the largest parent compound. Example:

2-Hydroxysilyltetrasilane-1,4-diol

$$\begin{array}{c} \quad\quad\quad SiH_2\cdot OH \\ \quad\quad\quad | \\ HO\cdot SiH_2\cdot SiH_2\cdot SiH\cdot SiH_2\cdot OH \end{array}$$

Substituents other than hydroxyl groups (functional atoms or groups and hydrocarbon radicals) attached to silicon are expressed by appropriate prefixes or suffixes. Examples:

Ethyldisilane
$CH_3\cdot CH_2\cdot SiH_2\cdot SiH_3$
Hexachlorodisiloxane
$Cl_3Si\cdot O\cdot SiCl_3$
Dibutyldichlorosilane
$(CH_3\cdot CH_2\cdot CH_2\cdot CH_2)_2SiCl_2$
Silylamine
$H_3Si\cdot NH_2$
Silanediamine
$H_2Si(NH_2)_2$
Silanetriamine
$HSi(NH_2)_3$
N-Methylsilylamine
$H_3Si\cdot NH\cdot CH_3$
NN-Dimethylsilylamine
$H_3Si\cdot N(CH_3)_2$
NN′-Dimethylsilanediamine
$H_2Si(NH\cdot CH_3)_2$
NN′N″-Trimethylsilanetriamine
$HSi(NH\cdot CH_3)_3$
Acetoxytrimethylsilane
$(CH_3)_3Si\cdot O\cdot OC\cdot CH_3$
Diacetoxydimethylsilane
$(CH_3)_2Si(O\cdot OC\cdot CH_3)_2$

Compounds containing carbon as well as silicon and in which there is a "reactive group" in the carbon-containing portion of the molecule not shared by a silicon atom are named in terms of the organic parent compound wherever feasible. Examples:

α-Trimethylsilylacetanilide
$(CH_3)_3Si\cdot CH_2\cdot NH\cdot C_6H_5$
1-Trichlorosilylethanol
$Cl_3Si\cdot CH(OH)\cdot CH_3$
2-Trimethylsilylethanol
$(CH_3)_3Si\cdot CH_2\cdot CH_2\cdot OH$
(Hydroxydimethylsilyl)methanol
$$\begin{array}{l} (CH_3)_2Si\cdot CH_2\cdot OH \\ \quad\quad | \\ \quad\quad OH \end{array}$$
α-(Hydroxydimethylsilyl)acetanilide
$$\begin{array}{l} (CH_3)_2Si\cdot CH_2\cdot\ CO\cdot NH\cdot C_6H_3 \\ \quad\quad | \\ \quad\quad OH \end{array}$$
(Silylmethyl)amine
$H_3Si\cdot CH_2\cdot NH_2$
But by rules 70.16 and 70.17:
(Methoxymethyl)silanol
$CH_3O\cdot CH_2\cdot SiH_2\cdot OH$
N-Methylsilylamine
$H_3Si\cdot NH\cdot CH_3$

Compounds in which metals are combined directly with silicon are, in general, named as derivatives of the metal. Example:

(Triphenylsilyl)lithium
$(C_6H_5)_3SiLi$

However, in exceptional cases, the metal may be named as a substituent. Example:

Sodium *p*-(sodiosilyl)benzoate
p-$NaO_2C\cdot C_6H_4\cdot SiH_2Na$

Metallic salts of hydroxy-derivatives may be named in the customary manner. Example:

Sodium salt of triphenylsilanol
$(C_6H_5)_3Si\cdot ONa$

Essentially Nonchemical Properties of Silicon

Were it not for the firm establishment of silicon as an indispensable material for modern electronics, the other exceptionally attractive properties of Si may have been overlooked for many years. Over the past few decades, electronics components manufacturers have mastered the skills required for manufacturing microminiature components, and this experience has given Si a headstart for use in other subminiature structures. Si has been recognized as an outstanding material for making micromachined subminiature structures essentially just within the past decade, and it has become one of the key materials in the comparatively new field of *nanotechnology*.

As a mechanical material, silicon is stronger than steel, it does not show mechanical hysteresis, and it is highly sensitive to stress. This combination of properties qualifies Si as an excellent sensor for detecting acceleration, pressure, force, and other variables encountered in processing and manufacturing. One method of measuring fluid flow, for example, traditionally has depended upon sensing pressure differentials, as in the case of an orifice-type flowmeter. Thus, silicon sensors can be used. Silicon accelerometers can employ the same piezoresistive sensing technique used in pressure sensors.

In addition to sensors, Si can be used at the subminiature scale for the production of tiny pipe, nozzles, and valves required by automatic control systems. Thus, the weight and bulk of future control systems may be reduced by several orders of magnitude.

As of the early 1990s, engineers are working at the "edge" of a new kind of robotics—that is, subminiature handling devices that can master the handling requirements of the new nanomanufacturing technology. Such robots would be miniature, fully integrated silicon systems drawing heavily on the technologies of silicon-integrated electronics and micromachining. Semi-intelligent robots could be used in many manufacturing and control tasks. According to some researchers, such robots could have intelligence at the lowest possible system level, thus allowing them to function semiautonomously, with occasional input from a central control system.

Biological Applications of Silicon Technology. H. M. McConnell (Stanford University) and a team of researchers have developed a silicon-based device called a cytosensor (microphysiometer) that can be used to detect and monitor the response of cells to a variety of chemical substances, particularly ligands for specific plasma membrane receptors. As pointed out by McConnell, "The microphysiometer measures the rate of proton excretion from 10^4 to 10^6 cells. The instruments serves two distinct functions. In terms of detecting specific molecules, selected biological cells in this instrument serve as detectors and amplifiers. The microphysiometer can also investigate cell function and biochemistry. A major application of this instrument may prove to be screening for new receptor ligands. In this respect, the instrument appears to offer significant advantages over other techniques." More detail is given in the McConnell reference listed.

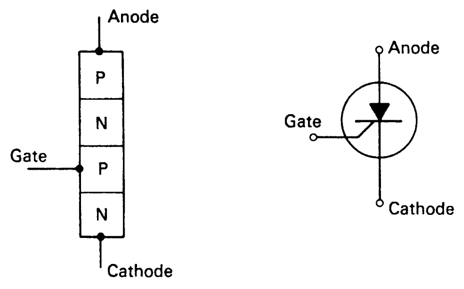
Additional Reading

Amato, I.: "Shine On, Holey Silicon," *Science*, 922 (May 17, 1991).
Aufderhaar, H. C.: "Silicon," in "Metals Handbook," 9th edition, Vol. 2, American Society for Metals, Metals Park, Ohio, 1979.
Binnig, G., and H. Rohrer: "The Scanning Tunneling Microscope," *Sci. Amer.*, **253**(2), 50–56 (August 1985).
Boland, J. J., and G. N. Parsons: "Bond Selectivity in Silicon Film Growth," *Science*, 1304 (May 29, 1992).

Bryzek, J., Mallon, J. R., Jr., and R. H. Grace: "Silicon's Synthesis: Sensors to Systems," *Instrumentation technology*, 40 (January 1989).
Carter, G. F., and D. E. Paul: "Materials Science and Engineering," ASM International, Materials Park, Ohio, 1991.
Connally, J. A., and S. B. Brown: "Slow Crack Growth in Single-Crystal Silicon," *Science*, 1537 (June 12, 1992).
Corcoran, E.: "Holey Silicon," *Sci. Amer.*, 102 (March 1992).
Dunn, W.: "Micromachined Sensors for Automotive Applications," *Sensors*, 54 (September 1991).
Golovchenko, J. A.: "The Tunneling Microscope: A New Look at the Atomic World," *Science*, **232**, 48–53 (1986).
Heath, J. R.: "A Liquid-Solution-Phase Synthesis of Crystalline Silicon," *Science*, 1131 (November 13, 1992).
Henkel, S.: "Silicon Microvalves Fabricated on Bimetallic Diaphragms," *Sensors*, 4 (December 1991).
Iyer, S. S., and Y-H Xie: "Light Emission from Silicon," *Science*, 40 (April 2, 1993).
Kasper, J. S., and R. H. Wentorf, Jr.: "Hexagonal (Wurtzite) Silicon," *Science*, **197**, 599 (1977).
Laine, R. M.: "Beach Sand: Material of the Future?" *Advanced Materials & Processes*, 6 (February 1992).
LeComber, P. G.: "Amorphous Silicon—Electronics Into the 21st Century," *University of Wales Review*, 31 (Spring 1988).
Link, B.: "Field-Qualified Silicon Accelerometers," *Sensors*, 28 (March 1993).
Maugh, T. H., II: "A New Route to Intermetallics (Metal Silicides)," *Science*, **225**, 403 (1984).
McConnell, et al.: "The Cytosensor Microphysiometer: Biological Applications of Silicon Technology," *Science*, 1906 (September 25, 1992).
Meyers, R. A.: "Handbook of Chemicals Production Processes," McGraw-Hill, New York, 1986.
Paquette, L. A.: "Silicon-Mediated Organic Synthesis," *Science*, **217**, 793–800 (1982).
Pugar, E. A., and P. E. D. Morgan: "Low Temperature Direct Reactions Between Elemental Silicon and Liquid Ammonia or Amines for Ceramics and Chemical Intermediates," in Report issued by Rockwell International Science Center, Thousand Oaks, California (September 1988).
Robinson, A. L.: "Consensus on Silicon Surface Structure Near," *Science*, **232**, 451–453 (1986).
Simpson, T. L., and B. E. Volcani, Eds.: "Silicon and Siliceous Structures in Biological Systems," Springer-Verlag, New York, 1981.
Staff: "Grace with Pressure (Silicon)," *Sci. Amer.*, **253**(2), 62–64 (August 1985).
Staff: "ASM Handbook—Properties and Selection: Nonferrous Alloys and Pure Metals," ASM International, Materials Park, Ohio, 1990.
Staff: "Silicon Atoms 'See the Light'," *Advanced Materials & Processes*, 6 (November 1990).
Staff: "Tough $MoSi_2$ Composites also Combat Oxidation," *Advanced Materials & Processes*, 26 (January 1991).
Staff: "Handbook of Chemistry and Physics," 73rd Edition, CRC Press, Boca Raton, Florida, 1992–1993.
Travis, J.: "Building a Silicon Surface, Atom by Atom," *Science*, 1354 (March 13, 1992).
Wentorf, R. H., Jr., and J. S. Kasper: *Science*, **139**, 338 (1963).
West, R.: "Isolable Compounds Containing a Silicon-Silicon Double Bond," *Science*, **225**, 1109–1114 (1984).
Yeganeh-Haeri, A., Weidner, D. J., and J. B. Parise: "Elasticity of Alpha-Cristobalite: A Silicon Dioxide with a Negative Poisson's Ratio," *Science*, 650 (July 31, 1992).
Yun, W., and R. T. Howe: "Recent Developments in Silicon Micro-accelerometers," *Sensors*, 31 (October 1992).
Yun, W., and R. T. Howe: "Sigma-Delta Modulator Interfacing with Silicon Microsensors," *Sensors*, 11 (May 1993).
Zdebick, M.: "A Revolutionary Actuator for Microstructures," *Sensors*, 26 (February 1993).

SILICON CONTROLLED RECTIFIER (or SCR). A semiconductor device consisting of four alternate layers of *n* and *p* type silicon, which functions as a current controlled switch. The schematic symbol and its correspondence to the four layers is shown in the figure. Two outstanding features are associated with the SCR, namely, the speed of switching and ratio of controlled to controlling currents. Load currents of tens to hundreds of amperes may be turned on in a few microseconds and turned off in about ten times as long a time. Typically, the controlling current is several orders of magnitude less than the controlled current.

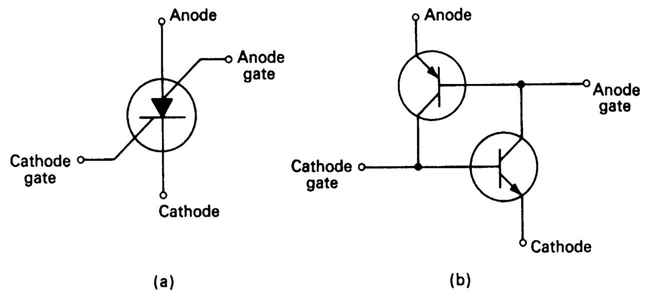
In operation, a comparatively small anode current flows when the gate current is zero and the anode-to-cathode voltage is held below a critical value known as the *forward breakover voltage*. A positive pulse of gate current of suitable magnitude triggers the device into a high



Diagrammatic representation of silicon-controlled rectifier and schematic symbol.

conduction mode at which time the anode current is determined almost entirely by the characteristics of the external circuit in series with the anode-cathode path, providing this current exceeds a small minimum value called the *holding current*. Once the high conduction mode is achieved, the gate current has no further control over the anode current. To turn off the anode current, the anode-to-cathode voltage must be reduced substantially (so that the anode current falls below the holding current). For turn-off in minimum time, it is necessary to reverse the polarity of the anode voltage.

SILICON CONTROLLED SWITCH (or SCS). An electronic device consisting of four layers of semiconductor material alternately doped with *p* and *n* type impurities. It differs from the silicon controlled rectifier in that external connections are made to all four layers rather than to only three, as in the case of the latter. The schematic symbol and an indication of the equivalence of the device to a combination of two junction transistors are shown in the figure. It is primarily used for switching and control applications.



Schematic symbol of silicon controlled switch (a) and equivalence to interconnection of two transistors (b).

SILK COTTON TREES. Of the family *Bombaceceae* (silk cotton family), there are several genera of these trees from which valuable commercial fibers and oils are obtained. The silk cotton tree (*Ceiba pentandra*), also called the ceiba or kapok tree is a tropical tree with a large trunk that may reach a height of about 100 feet (30.5 meters) and diameter of 10 feet (0.3 meter). The tree is found in Brazil, Java, India, Central America, and Ecuador. The trunk has large roots that stabilize the tree and taper from underground to buttress the natural trunk. The tree produces large pods which yield silk cotton called *kapok*. This material is used for filling mattresses and in padding cushions and furniture. Kapok is very light and resilient and also can be used as an insulating material. Most of the kapok used in the United States comes from Java. The Javanese tree was introduced from Brazil where the tree is referred to locally as the samaúma. The long, silky, and white fibers appear much like cotton. However, the fibers cannot be spun because of their brittleness. Kapok is also sometimes referred to as Illiani silk and Java cotton. Chemically, kapok is similar to jute, but with different proportions of the constituents, kapok being lower in alpha cellulose and higher in pentosans and lignin.

Other important genera of the silk cotton family include: *Calotropis gigantea*, a shrub grown in the East Indies and India and which produces a silky cotton known as *madar*, of an inferior quality to kapok.

Calotropis procera, found in the same region, produces *akund*. Sometimes these inferior fibers are mixed with kapok. The species *Bombax ceiba* and *B. malabaricum*, grown in India, yield a red silk cotton called *simal*. A fiber known as *shilo* is obtained from *B. ellipticum*. The Indian tree *Cochlospermum gossypium* yields a white silk cotton similar to kapok. Mexican kapok is produced from *Bombax palmeri* and from *Ceiba schottii* and *C. acuminata*. Ecuadoran kapok comes from the genus *Chorisia*. Balsa fiber from the genus *Ochroma*, produced in Colombia, Ecuador, and Central America, is of a darker color.

The seeds of the kapok tree yields a semidrying oil which finds use in the manufacture of margarine and soaps.

SILKWORM (*Insecta*, *Lepidoptera*). The caterpillar of the moth, *Bombyx mori*, which produces the silk of commerce. Caterpillars of the family *Saturniidae* also spin cocoons of silk, and are known as the giant silkworms. This family includes the common luna, cecropia, and polyphemus moths of North America, as well as many other species.

SILL. A tabular mass of igneous rock that has been intruded laterally between layers of sedimentary rock, beds of volcanic lava or ejectmenta, or even along the direction of the foliation in metamorphic rock. The term sill is synonymous with intrusive sheet.

SILLIMANITE. The mineral sillimanite is an aluminum silicate, the formula Al_2SiO_5 being like that of andalusite and kyanite. It is orthorhombic, usually in slender prisms, but may be fibrous or massive. Its hardness is 6.5–7.5; specific gravity, 3.23–3.27; luster, vitreous to silky; color, various shades of gray, grayish-green, and grayish-brown; transparent to translucent. It occurs in granites and gneisses as tiny prisms and aggregates, and is often associated with andalusite, cordierite and corundum. Sillimanite has been found in Bavaria, the Czech Republic and Slovakia, France, India, the Malagasy Republic, Myanmar and Ceylon, the latter two localities furnishing transparent sapphire-blue gem stones. In the United States sillimanite has been found in Connecticut, New York, Pennsylvania, Delaware, North Carolina, and in California, where in Inyo County is the largest deposit in the world. This mineral was named in honor of Benjamin Silliman for many years professor of chemistry and natural science at Yale University. Sillimanite is used in the manufacture of spark plug "porcelains" and laboratory ware.

SILTSTONE. A term signifying a clastic sedimentary rock in which the particles are of silt grade.

SILURIAN PERIOD. A major subdivision of the Paleozoic Era. Type locality, Wales and Shropshire, England. The formations of this system were first studied and described by R. I. Murchison in 1835. The Silurian Period began 380 million years ago and lasted for 50 million years. In Murchison's time, and for some time afterward, the Paleozoic Era, below the Devonian, was divided into the Cambrian and Silurian. This practice still holds in parts of Europe. The Silurian formations are well exposed in Eastern North America, especially in New York State, and the length of the Appalachian geosyncline where the sediments are principally red sandstones and shales of delta origin. There was little or no volcanic activity except in southeastern Maine. The Silurian is also well exposed in nearly all of Western Europe, Northern Siberia, Myanmar, Central Asia, Himalayas, Morocco, Australia, Peru and Bolivia. The system is characterized by all types of sediments, with evidence in certain areas of an arid climate, including thick deposits of salt and gypsum. The maximum thickness of sediments is 15,000 feet (4,500 meters) in Britain.

SILVER. Chemical element, symbol Ag (from Latin *argentum*), at. no. 47, at. wt. 107.868 ± 0.003, periodic table group 11, mp 961.93°C, bp approximately 2,212°C, density 10.50 g/cm^3 (20°C). Elemental silver has a face-centered cubic crystal structure.

Silver is a white metal, softer than copper and harder than gold. When molten, silver is luminescent and occludes oxygen, but the oxygen is released upon solidification. As a conductor of heat and electricity, silver is superior to all other metals. Silver is soluble in HNO_3 containing a trace of nitrate; soluble in hot 80% H_2SO_4; insoluble in HCl or acetic acid; tarnished by H_2S, soluble sulfides and many sulfur-containing organic substances (e.g., proteins); not affected by air or H_2O at ordinary temperatures, but at 200°C, a slight film of silver oxide is formed; not affected by alkalis, either in solution or fused. There are two stable, naturally occurring isotopes, ^{107}Ag and ^{109}Ag. In addition, there are reported to be 25 less stable isotopes, ranging in half-life from 5 seconds to 253 days.

In terms of cosmic abundance, the estimate of Harold C. Urey (1952), using silicon as a base with a figure of 10,000, silver was assigned an abundance figure of 0.023. In terms of abundance in sea water, silver is ranked number 43 among the elements, with an estimated content of 1.5 tons per cubic mile (0.324 metric ton per cubic kilometer) of sea water.

Electronic configuration is $1s^2 2s^2 2p^6 3s^2 3p^6 3d^{10} 4s^2 4p^6 4d^{10} 5s^1$. First ionization potential is 7.574 eV; second, 21.4 eV; third, 35.9 eV. Oxidation potentials: $Ag \rightarrow Ag^+ + e^-$, $E^0 = -0.7995$ V; $Ag^+ \rightarrow Ag^{2+} + e^-$, -1.98 V; $2Ag + OH^- \rightarrow Ag_2O + H_2O + 2e^-$, -0.344 V; $Ag_2O + 2OH^- \rightarrow 2AgO + H_2O + 2e^-$, -0.57 V; $2AgO + 2OH^- \rightarrow Ag_2O_3 + H_2O + 2e^-$, -0.74 V. Other important physical properties of silver are given under **Chemical Elements.**

Occurrence and Processing: Silver is widely distributed throughout the world. It rarely occurs in native form, but is found in ore bodies as silver chloride, or more frequently, as simple and complex sulfides. In former years, simple silver and gold-silver ores were processed by amalgamation or cyanidation processes. The availability of ores amenable to treatment by these means has declined. Most silver is now obtained as a by-product or co-product from base metal ores, particularly those of copper, lead, and zinc. Although these ores are different in mineral complexity and grade, processing is similar.

All the ores are concentrated in complex mills by selective froth flotation to produce individual copper, zinc, lead, and, infrequently, silver concentrates. The copper and lead concentrates are smelted to produce lead and copper bullions from which silver is recovered by electrolytic or fine refining. The silver bearing zinc concentrates are commonly processed by leaching and electrolytic methods. Silver is ultimately recovered as a by-product from zinc plant residues. Canada is a leading silver mining country. Other important sources of silver are Mexico, the United States, Peru, the former U.S.S.R., and Australia. See also **Mineralogy.**

A substantial portion of the total world silver supply is obtained from recycled scrap. Much of this scrap comes from photographic film, jewelry and the electrical field. The high value of the scrap dictates accurate sampling and careful feed preparation. Efficient and fast processing is required to minimize metal losses and a tie-up of high-value materials. The highly complex nature of plant feed, with respect to physical form, chemical composition, and grade, requires use of complex and highly flexible processing procedures.

Uses of Silver: Silver in the 20th century can be classified an industrial commodity. For most of the 19th century, silver was a monetary metal. Industrial consumption of silver is principally in photographic film, electrical contacts, batteries and brazing alloys. Sterling silver and silver plated copper alloys are used extensively for tableware and jewelry and other decorative art. Recently, the field of commemorative and collector arts has become a substantial market for silver alloys, particularly sterling silver.

The predominent place of silver salts as photographic receptors is not the result of any unusual primary sensitivity to illumination, but is due to the fact that they undergo an unusual secondary amplification process called "development." Silver salts, like the salts of many metals, when immersed in solutions of many reducing agents, are changed to metallic silver. The photographic system depends upon the fact that when certain mild reducing agents (called "photographic developers") are chosen, the rate of reduction is increased many fold if the silver salt crystals carry very small amounts of metallic silver at the developer-crystal interface. The effect produced by the original light exposure is amplified in the development process by a factor of 100 billion. Whereas new photographic or recording devices are being developed not involving silver, none yet approach the packing density of a fine-grained image possible using silver. Thus it appears that silver will be used in photographic recording for many years to come.

Among the electrical uses for silver are electrical contacts, printed circuits, and batteries. By far, the primary use is in electrical contacts where the high electrical and thermal conductivities, as well as corrosion and oxidation resistances, of silver are major reasons for its selection. Although silver has a strong tendency to weld under heavy currents, this is counteracted by alloying or by adding nonmetallic substances (such as cadmium oxide) to the silver matrix. The use of silver-cadmium oxide and silver-tungsten materials in electrical contact applications is widespread. The alloys used to improve the wear resistance and to reduce the sticking tendency of silver include silver-gold, silver-copper, silver-palladium, and silver-platinum. More complex alloys include silver-copper-nickel, silver-magnesium-nickel, silver-gold-cadmium-copper, and silver-cadmium-copper-nickel. Silver-cadmium oxide alloys are unique materials and are prepared either by combining silver and cadmium oxide by powder metal techniques or by the internal oxidation of a silver-cadmium alloy. Electrical alloys, which are impossible to combine by conventional melting, lend themselves to powder metal fabrication. Such composite structures as silver-graphite, silver-iron, and silver-tungsten are good examples of these types of materials.

In silver batteries, the silver oxide-zinc secondary battery has found its place in applications where energy delivered per unit of weight and space are of prime importance. The major disadvantages lie in their high cost and relatively short life. Consequently, a large part of the silver battery market is concerned with defense and space components. See also **Battery.**

Prior to World War II consumption of silver in silverware and jewelry was the largest industrial use of silver. Competition from stainless steel in flatware and holloware has contributed to a decline in overall use. Most consumption of silver in silverware and jewelry is in the form of sterling silver, an alloy of silver with approximately 7.5 weight percent copper. Silver plate, which is silver electroplated on a base metal, varies widely in specification. The thickness, expressed for example in pennyweights of pure silver per gross of teaspoons can range from a low of 1 to as high as 200.

In the 1920s and 1930s, low-temperature silver-copper brazing alloys were found to be useful on copper and its alloys and iron and its alloys (including stainless steel). Silver and copper form a simple eutectic system with limited solid solubility. This system can absorb elements such as zinc, cadmium, tin, and indium. These additions lower its melting temperature. It also can absorb higher melting elements such as nickel or palladium. These raise its melting temperature, but may improve its wetting characteristics, corrosion resistance, and strength at elevated temperatures. Silver solders or brazing alloys have the ability of making joints far stronger and more durable than common soft-solder (such as lead-tin) alloys. They are used in most refrigeration systems to join copper tubing. Also, extensive use is found in the assembly of automotive parts, military components, aircraft assemblies, and other hard goods manufacture. The nominal composition of a popular brazing alloy, ASTM Classification BAg-1 is silver 45%, copper-15%, zinc-16%, cadmium-24%.

One silver alloy containing about 70% silver, 26% tin, 3% copper, and 1% zinc is unique in that it is used extensively by dentists in combination with mercury to fill cavities in teeth. The "amalgam" manufacturers supply dentists with the alloy in the form of powder (filed, or more recently, atomized). This is mixed with mercury, using from 8 to 5 parts of mercury to 5 of alloy, and the cavity is packed. In the cavity, a metallurgical reaction takes place in which the silver-tin compound in the alloy becomes a durable silver-tin-mercury compound.

Silver, its oxides, halides and other salts play important roles in chemistry. Silver is an excellent catalyst in oxidizing reactions such as in the production of formaldehyde from methanol and oxygen, ethylene oxide from ethylene and oxygen, and glyoxal from ethylene glycol and oxygen. Silver has oligodynamic properties, that is, the ability of minute amounts of silver in solution to kill bacteria. Modern technology has made use of this property in various ways, mainly as a means of purifying water.

Small amounts of silver are used annually in such diverse applications as a backing for mirrors, and in control rods for pressurized water nuclear reactors. Miscellaneous uses like this account for only a small fraction of total silver consumption.

Chemistry of Silver. Silver(I) oxide, Ag_2O, is made by action of oxygen under pressure on silver at 300°C, or by precipitation of a silver salt with carbonate-free alkali metal hydroxide; it is covalent, each silver atom (in solid Ag_2O) having two collinear bonds and each oxygen atom four tetrahedral ones; two such interpenetrating lattices constitute the structure. Silver(I) oxide is the normal oxide of silver. Silver(II) oxide, AgO, is formed when ozone reacts with silver, and thus was once considered to be a peroxide. Silver(III) oxide, Ag_2O_3, has been obtained in impure state by anodic oxidation of silver.

All of the silver(I) halides of the four common halogens are well known. The fluoride may be prepared from the elements, the chloride by action of hydrogen chloride gas at 150°C, upon silver, and the bromide and iodide by ionic reactions in solution. The chloride, bromide, and iodide are essentially insoluble in H_2O, but the fluoride is soluble. There is also a subfluoride, Ag_2F, which may be prepared as a cathodic deposit by electrolysis of silver(I) fluoride AgF, or by evaporation of finely divided silver with silver(I) fluoride in dilute hydrofluoric acid. It is an anisotropically conducting solid and is considered to be made up in the solid state of two silver layers, metallic-bonded to each other, and ionic-covalent bonded to a single fluorine layer. It has reverse cadmium iodide structure. Silver subchloride, Ag_2Cl is made by reaction of Ag_2F and phosphorus trichloride. Silver(II) fluoride, AgF_2, made by action of fluorine upon a silver(I) halide, is a fluorinating agent or catalyst for fluorinations. The silver(I) halides vary markedly in ionicity, the values given by Pauling being AgF 70%, AgCl 30%, AgBr 23% and AgI 11%. This is reflected in their crystal structures and in their solubilities in water (or rather, their relative insolubility), the first three having sodium chloride structure, AgI having wurtzite structure; AgF having a molal solubility of 14, and the pK_{sp}'s of the others being 9.75, 12.27 and 16.08, respectively.

Silver differs markedly from copper in forming few oxy compounds. One of these is silver oxynitrate or silver(II, III) nitrate which has the empirical formula $AgO_{1.148}(NO_3)_{0.453}$, in which the average oxidation number of silver is 2.448. It is prepared by action of fluorine upon aqueous silver nitrate or is obtained as an anodic deposit by electrolysis of silver nitrate in dilute HNO_3.

Silver enters into complex formation with many ions and molecules. With halogens, the silver complexes are fewer than the copper ones. Silver chloride dissolves in HCl with the formation of such chloroargentate ions as $(AgCl_2)^-$, $(AgCl_3)^{2-}$, and possibly $(AgCl_4)^{3-}$. Complex ions with bromide, $(AgBr_2)^-$ and $(AgBr_3)^{2-}$ are more stable, as are those with iodide, than those with chloride. Complexes of the type Ag_2Cl^+, Ag_3Cl^{2+}, Ag_2Br^+, Ag_3Br^{2+}, Ag_4Br^{3+}, $Ag_2Br_6^{2-}$, Ag_2I^+, Ag_3I^{2+}, Ag_4I^{3+}, $Ag_2I_6^{4-}$, $Ag_2I_7^{5-}$ and $Ag_3I_8^{5-}$ are also known. With ammonia the ions $(Ag(NH_3)_2)^+$ and $(Ag(NH_3)_3)^+$ are definitely known and others may exist. Similar complexes are formed with amines and diamines. With cyanides, silver forms very stable complexes, the number of CN^- ions in the complex depending somewhat upon the excess of cyanide, so that $(Ag(CN)_2)^-$, $(Ag(CN)_3)^{2-}$, and $(Ag(CN)_4)^{3-}$ are definitely known. With thiosulfates, silver forms various complexes. In dilute solution, $(Ag_2(S_2O_3)_2)^{2-}$ exists, while in high concentration of $S_2O_3^{2-}$ ion, the complex $(Ag_2(S_2O_3)_6)^{10-}$ has been identified. In HNO_3 solution Ag^+ is easily oxidized to Ag^{2+} by peroxydisulfate. From this solution complex compounds of dipositive silver can be prepared, which are stable because coordination radically alters the oxidation potential of Ag(I) to Ag(II). They include pyridine complexes such as $(Ag(py)_4) \times (NO_3)_2$. 8-Hydroxyquinoline complexes containing the ions $(Ag(oxin)_2)^{2+}$, and o-phenanthroline complexes containing the ion $(Ag(o\text{-}phen)_2)^{2+}$. Silver(III) is known in the square, planar complex AgF_4^-, which has been prepared as $KAgF_4$ by direct fluorination of a mixture of potassium chloride and silver chloride. Silver(III), like Cu(III), also occurs in tellurate and periodate complexes.

Other silver compounds include: Silver chromate (Ag_2CrO_4), yellow to red to brown precipitate by reaction of silver nitrate solution and potassium chromate solution.

Silver dichromate ($Ag_2Cr_2O_7$), red precipitate by reaction of silver nitrate solution and potassium dichromate solution, changing to silver chromate upon boiling with H_2O.

Silver phosphate (Ag_3PO_4), yellow precipitate, by reaction of silver nitrate solution and disodium hydrogen phosphate solution, soluble in HNO_3 and in NH_4OH, turns dark on exposure to light.

Silver sulfate (Ag_2SO_4), white precipitate, by the action of silver nitrate solution and potassium sodium or ammonium sulfate solution or H_2SO_4, mp of silver sulfate 652°C.

Silver sulfide (Ag_2S), black precipitate, by the reaction of silver nitrate solution and hydrogen sulfide.

Silver forms several compounds or complexes with proteins by the action of silver oxide with gelatin in alkali solution, or with ablumin, or by suspension in casein solution and by other methods. Such silver-protein complexes containing from 19 to 23% of silver are known as "mild silver protein" and are used as antiseptic solutions. They are readily soluble in H_2O.

Donald A. Corrigan, Handy & Harman, Fairfield, Connecticut.

Additional Reading

Carapella, S. C., Jr., and D. A. Corrigan: "Properties of Pure Silver," *Metals Handbook*, 9th Edition, Vol. 2, American Society for Metals, Metals Park, Ohio, 1979.

Coxe, C. D., McDonald, A. S., and G. H. Sistare, Jr.: "Properties of Silver and Silver Alloys," *Metals Handbook*, 9th Edition, Vol. 2, 1979.

Coxe, C. D., McDonald, A. S., and G. H. Sistare, Jr.: "Silver-Base Brazing Alloys," *Metals Handbook*, 9th Edition, Vol. 2, American Society for Metals, Metals Park, Ohio, 1979.

Friend, W. Z.: "Corrosion Resistance of Precious Metals," *Metals Handbook*, 9th Edition, Vol. 2, American Society for Metals, Metals Park, Ohio, 1979.

Gale, N. H., and Z. Stos-Gale: "Lead and Silver in the Ancient Aegean," *Sci. Amer.*, 176–192 (June 1981).

Greener, E. H.: "Dental Materials," *Encyclopedia of Materials Science and Engineering*, MIT Press, Cambridge, Massachusetts, 1986.

Hamer, W. J.: "Silver" McGraw-Hill Encyclopedia of Chemistry, McGraw-Hill, New York, 1983.

Hammond, C. R.: "The Elements" in *Handbook of Chemistry and Physics*, 67th Edition, CRC Press, Boca Raton, Florida, 1986–1987.

Lechtman, H.: "Pre-Columbian Surface Metallurgy," *Sci. Amer.*, 56-53 (June 1984).

Sinfelt, J. H.: "Bimetallic Catalysts," *Sci. Amer.*, 90–98 (September 1985).

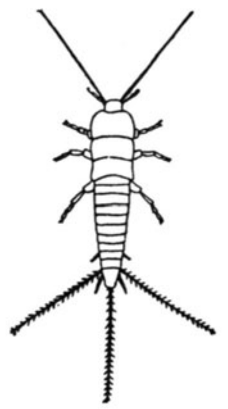
Staff: "Handbook of Chemistry and Physics," 73rd Edition, CRC Press, Boca Raton, Florida, 1992–1993.

Waterstrat, R. M., and G. Dickson: "Dental Amalgam (Hg, Ag, Sn, Cu, Zn)," *Metals Handbook*, 9th Edition, Vol. 2, American Society for Metals, Metals Park, Ohio, 1979.

Zysk, E. D.: "Precious Metals and Their Use," *Metals Handbook*, 9th Edition, Vol. 2, American Society for Metals, Metals Park, Ohio, 1979.

SILVER FIRS. See **Fir Trees.**

SILVERFISH (*Insecta, Thysanura. Lepisma*). One of the primitive wingless insects. It is about $\frac{1}{2}$ inch (12 millimeters) long, broad in front and tapering behind, with two long slender antennae and three similar processes at the caudal end of the body. It is covered with lustrous grayish scales, hence the common name. The insect eats starchy materials and sometimes defaces book bindings, wall paper, and laundry, but as a rule it is not sufficiently abundant to be a pest. It is often found in damp buildings. It is also called the fish moth.


Silverfish.

SILVERSIDES (*Osteichthyes*). Of the suborder *Mugiloidea*, family *Atherinidae*, the silversides are related to barracudas and mullet. They also are termed antherinid smelts, but should not be confused with the true smelts. See **Smelts.** There are numerous species, ranging up to a length somewhat in excess of 20 inches (51 centimeters). The *Antherinopsis californiensis* (jack smelt) is a silvery fish found in the waters off the California coast. Fresh and brackish-water silversides are found in the Great Lakes and in inland waters as far south as Florida and as far west as Texas. Mexican freshwaters support the *Chirostoma*, a rather abundant fish reaching a length of about 20 inches (51 centimeters) and well regarded as a food fish. Several species occur in the Mediterranean, only a few of which are found in the more northern waters off the British Isles.

Spawning grunions.

Occurring in the waters of Lower California and as far northward as southern California, the *Leuresthes tenuis* (grunion) attains a length of 5 to 7 inches (12.5 to 18 centimeters) and is well known for its spawning habits. The female grunion lays her eggs in a shallow indentation in the beach and always at night during a high tide. The male fertilizes the eggs immediately, whereupon the grunion "flop" back to sea, aided by the waves. Within just a few minutes after they are touched by the waters of the subsequent high tide, the eggs are hatched. Spawning normally occurs within one or two days after a full or new moon. The *Hubbsiella sardina* has habits similar to the grunion, but will spawn during daylight hours. Grunions mature at 1 year and live about 3 years. See accompanying diagram.

SILVERY POUT. See **Codfishes.**

SIMPLEX OPERATION (Communication System). A method of operation in which communication between two stations takes place in one direction at a time. This includes ordinary transmit-receive operation, press-to-talk operation, voice-operated carrier, and other forms of manual or automatic switching from transmit to receive.

SIMULATOR (Computer System). A device, data processing system, or computer program which represents a system or phenomenon and which mirrors or maps the effects of various changes in the original, enabling the original to be studied, analyzed, and comprehended by means of the behavior of the model. The term also is used to designate a routine which is executed by one computer, but which imitates the operations of another computer.

SINGLE-ENDED AMPLIFIER. A device designed to amplify signals between a single input terminal and the ground or signal reference point. Generally, the output signal is generated between a single output terminal and the same signal reference point.

A disadvantage of the design stems from installation difficulties caused by the common connection between the input and output circuits. Unfortunately, any potential that exists between the amplifier signal reference point (ground) and the reference point for the input signal appears in series with the input signal. Thus, this additional potential is amplified with the signal and cannot be distinguished from the true signal voltage. Since, by definition, a potential between two signal reference points is a common-mode voltage, the common-mode rejection of the single-ended amplifier is zero. See **Common-Mode Rejection.**

Where amplifiers of this kind are used in digital-data acquisition and instrumentation systems, careful attention must be given to the grounding circuits. Obviously, a major requirement is for the input signal reference point to be at precisely the same potential as the signal reference point for the amplifier input. Thus, a low-impedance connection between these two reference points at all frequencies of concern is needed. This is a difficult condition to achieve where the signal source and the amplifier are separated by a significant distance. Thus, amplifiers of this type seldom are used where such separation exists. For small subsystems where the ground structure is well controlled, single-ended amplifiers are used.

SINGULAR. A term for unusual or peculiar behavior of a mathematical entity. Thus, a singular matrix is one for which the determinant vanishes and a transformation involving that matrix is also singular. An integral equation with infinite limits of integration or an infinite kernel is singular. An algebraic or transcendental equation or function often becomes singular at one or more special values of the variables defining it. This is called a singular point or a singularity. If one studies these equations as the description of a curve, one is interested in points where tangents to the curve become indeterminate or infinite (see **Singular Point of a Curve**). On the other hand, the function itself might become infinite or show some peculiarity at a point. This is of special concern in the study of a differential equation (see **Singular Point of a Function**). The solution of a differential equation can also be singular. Such a solution cannot be obtained from the general solution by specifying the value of its parameters. See **Clairaut Equation.**

SINGULAR POINT OF A CURVE. If the tangent to a curve becomes an indeterminate form at one or more points, either finite or infinite, that point is called a singular point. The methods of differential calculus can, however, be used to find the tangents at such points and to investigate the behavior of the curve there. Many possibilities exist but only a few of them can be treated here. Conic sections do not have singular points; hence, the results apply only to higher plane curves.

When a curve crosses itself, there will be two tangents at that point, which is called a *double point.* If the two tangents are different, it is a *node* (L. *nodus*, knot) or *crunode* (L. *crux*, cross). If the two tangents coincide, the point is a *cusp* (L. *cuspis*, *point*), a *parabolic point*, or a *spinode* (L. *spina*, thorn). A third possibility is that of imaginary tangents and the point is then a *conjugate* or *isolated point*, also sometimes called an *acnode* (L. *acus*, needle).

When k branches of a curve cross, the point is a multiple point of order k. When $k > 2$, several new kinds of behavior appear, the most common of which are: *point of osculation* (L. *osculare*, kiss) or *tacnode* (L. *tactus*, touch), the curve recedes from the point of tangency in opposite directions; *cusp of the second kind*, also called *rhamphoid* (Gr. παμφος, beak), where the two branches of a curve lie on the same side of a common tangent. At a cusp of the first kind, also called *keratoid* (Gr. καπας, horn), the two branches of the curve are on opposite sides of the tangent. This is the only kind of cusp which can occur at a double point.

The preceding cases apply generally to algebraic curves. There are further possibilities for transcendental curves: *end point, terminal point*, or *point d'arrêt*, the curve terminates abruptly; *salient point* or *point saillant*, two branches of the curve end without a continuous derivative, thus without a common tangent.

Typical examples of each of these types, where the singular point occurs at the origin, are: node, lemniscate of Bernoulli; point of osculation, $a^4y^2 = x^4 (a^2 - x^2)$; cusp of first kind, semi-cubical parabola or cissoid; cusp of second kind, $(y^2 - x^2)^2 = x^5$; conjugate point, $y^2 = x^2 (x - 1)$; end point, $y = x \ln x$; salient point, $x = \cot y/x$.

See also **Cusp (Mathematics);** and **Maximum and Minimum (Mathematics).**

SINGULAR POINT OF A FUNCTION. A value of the complex variable z, for which $f(z)$ is an analytic function, is called an ordinary point. Any point which is not an ordinary point is a singular point. According to a theorem of Liouville, if $f(z)$ has no singular point for z finite or infinite, it must be a constant. Another Liouville theorem, more familiar to chemists and physicists, occurs in statistical mechanics. See **Liouville Equation.**

Singular points or singularities are classified as: (1) poles or unessential singularities; (2) essential singularities or poles of infinite order; (3) branch points caused by the fact that the function is not single-valued.

Let $w = u + iv$ be a single-valued function of $z = x + iy$, where u, v are real single-valued functions of x and y. Then $z = z_0$ is a pole of order k, provided that $(z - z_0)^k w(z)$ is analytic and not zero at $z = z_0$. The number k is an integer greater than unity and is the order of the pole. Singular points of this kind are non-essential because they are effectively removed if $w(z)$ is multiplied by $(z - z_0)^k$. They are called poles because a three-dimensional plot of w, x, y shows that w becomes infinite at the singular point and thus a pole of infinite height occurs there on the z-plane. Typical examples are $w = 1/z(z - a) \times (z - b)$, which has three simple poles at $z = 0, a, b$, respectively; $w = \csc z$, an infinite number of poles on the real axis.

A singularity which is not a pole or a branch point is called essential. It is actually a pole of infinite order. A simple example is $w(z) = \sin 1/z = 1/z - 1/3!z^3 + 1/5!z^5 - 1/7!z^7 \ldots$ and it is seen that no finite value of k in $z^k w(z)$ will remove the singular point of this function at $z = 0$.

The essential or non-essential character of a singularity can be investigated by expansion of the function in a Laurent series. In most cases, the result can be obtained more simply by inspection of the function. If the point $z = \infty$ is of interest, transfer to the new variable $z = \overline{1/\xi}$ and the function at $\xi = 0$.

When a function is multivalued its discontinuities are called branch points. Analytic continuation between two points will then give different values for $f(z)$ if the two paths include a branch point. Branch points always occur in pairs. The line joining them is a branch line and a contour crossing a branch line changes from one set of values of $f(z)$ to another. The two (or more) independent values of $f(z)$ are its branches. The values of the different branches of $f(z)$ are identical at a branch point. See also **Riemann Surface.**

Singular points are of considerable importance in the study of linear differential equations, especially when they are to be solved by expansions in series.

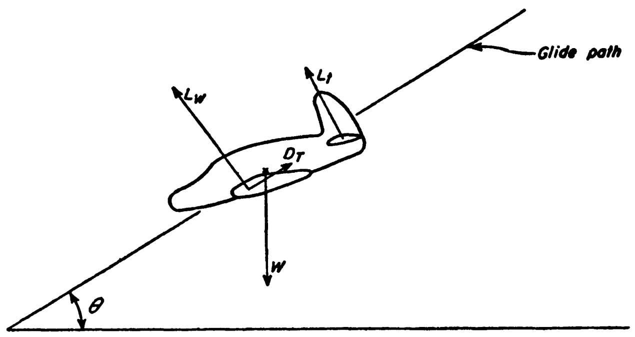
The point $z - z_0$ is a regular singular point of the second-order equation $(z - z_0)^2y'' + P(z)(z - z_0)y' + Q(z)y = 0$, provided: (1) z_0 is not an ordinary point; (2) both $P(z)$ and $Q(z)$ are analytic functions at $z = z_0$. If these conditions are not met, the singularity is irregular. This classification of singular points may be extended to linear differential equations of any order.

At a regular singular point the difference between the two exponents of the indicial equation is arbitrary, except that it cannot be zero or integral. If this difference equals $\frac{1}{2}$, the singularity is elementary.

An irregular singular point arises from the confluence of two or more regular singular points. In the case of a second-order differential equation, the indicial equation at an irregular singular point is of first degree or less and there is only one series solution there or none.

See also **Complex Variable.**

SINKING SPEED (Aircraft). The vertical component of the velocity of an airplane gliding without acceleration in power-off condition is the sinking speed. It is analogous to "rate of climb" of climbing flight. Methods exist for estimating sinking speed using other data of the airplane, to wit, air speed, altitude, weight, parasite area, wing span, and airplane efficiency factor. Gliding at minimum sinking speed should not be confused with the flattest glide.



Forces acting on a glider in a glide.

Referring to the figure, the following equations of statics may be obtained, using the lift-drag axes as positive axes:

$$L_w + L_t = W \cos \theta \quad (1)$$

$$D_T - W \sin \theta = 0 \quad (2)$$

where L_w = lift of the wing, L_t = lift of the horizontal tail surfaces small relative to L_w, and may therefore be neglected, D_T = total drag of the glider, W = weight of the glider.

With the assumption that the lift on the tail surfaces is negligible, the following relationships are obtainable from equations (1) and (2) and from the usual lift and drag equations:

$$L = \frac{\rho v_g^2 C_L S}{2} \quad \text{and} \quad D = \frac{\rho v_g^2 C_D S}{2}$$

where v is the glide path speed.

$$L_w = W \cos \theta$$
$$\cot \theta = \frac{L_w}{D_T} = \frac{C_L}{C_D}$$

Also, the sinking speed v_s, equals the gliding speed v_g multiplied by sin θ, and the horizontal component equals the gliding speed v_g multiplied by cos θ.

SINTERING. The heating of an aggregate of fine metal particles at a temperature below their melting point so as to cause them to weld together and agglomerate. See also **Agglomeration.**

SINUS. A pouch or cavity in any organ or tissue, or an abnormal cavity or passage formed by the destruction of tissue. The term is applied to a very large number of such structures in the human body, such as a dilated portion of a vein containing venous blood; a chronically infected tract such as a fistula; the air cavities within the cranial bones, especially those located near the nose and connecting with it. See illustration that accompanies entry on **Skeletal System.** They are called accessory sinuses of the nose or paranasal sinuses. They extend from the nasal passages into bones of the skull, and are named according to the bones in which they lie as the frontal, ethmoidal, sphenoidal, and maxillary sinuses. The maxillary sinuses are also called antra. The maxillary sinuses are found on either side of the nose. They are large in size, and lie between the floor of the eye socket above and the upper teeth below. The frontal sinuses lie in the forehead above the roof of the eye socket, one on either side of the midline of the forehead. The ethmoidal sinuses are three groups with numerous air cells, situated between the eyes in either side of the midline or septum of the nose. The sphenoidal sinuses, two in number, are situated above and behind the nose proper. They are all lined with a delicate mucous membrane continuous with the nasal mucous membrane. When the sinuses become infected, sinusitis is said to be present.

SINUSITIS. Infection of the mucous membrane lining of the paranasal sinuses. This is a common condition, but sometimes overdiagnosed. The disease may be acute or chronic. Rhinitis (viral or allergic) sometimes predisposes sinusitis. Also, the disease may follow blockage of the nasal septum by nasal polyps. Other less frequent causes include sudden changes in altitude, foreign bodies or tumors in the intranasal system, and certain processes of a systemic nature, such as cystic fibrosis. Any process which interferes with the drainage or ventilation of the sinuses may induce sinusitis. Common causative microorganisms include pneumococci, Group A and other streptococci, and *Haemophilus influenzae*. Less frequent causes are staphylococcus, *Mycobacterium tuberculosis*, *Aspergillus* fungi, *Mucor* fungi, viruses, and *Mycoplasma pneumoniae.* Involvement may be confined to one sinus, or more commonly to both sinuses in adults. Children are more inclined to suffer from *ethmoiditis.*

Symptoms of sinusitis vary with the site. See **Sinus.** Infection of the *frontal sinus (frontal sinusitis)* causes pain and tenderness over the lower forehead. There may be purulent (puslike) drainage. In *maxillary sinusitis*, there is pain and tenderness over the cheeks. The pain may be referred to the teeth; the hard palate may become inflamed. *Ethmoid sinusitis (ethmoiditis)*, in the acute phase, is characterized by a dull headache, which may be moderate or intense. Ethmoid pertains to the bones at the front part of the base of the skull, forming part of the septum and walls of the nasal cavity. These are perforated bones through which the olfactory nerves pass. There may be swelling between the inner corner of the eye and nose. The nasal discharge may be profuse or entirely absent, depending upon whether the drainage passages are partially or fully blocked. In subacute ethmoiditis, headache may be bothersome, but pain is usually not severe. Chronic inflammation of the ethmoid sinuses is manifested by headache, cough, a general feeling of fatigue, and often a slight fever. Acute *sphenoid sinusitis* occurs relatively frequently, with accompanying headache which has been described as excruciating and may be only partially relieved by drugs. Sphenoid pertains to the wedge-shaped compound bone at the base of the skull. Symptoms of sleeplessness and a fear of choking are sometimes present, caused by the thick discharge from the sinus. In the chronic form of the disease, the patient may experience pressure pains spreading in diverse directions, accompanied by a thick, sticky postnasal drip.

Therapy for acute sinusitis includes analgesics, topical heat, and decongestants. Pseudoephedrine is often used. Antihistamines may be used to aid in decongestion. There is no strong body of evidence to date that indicates the usefulness of antibiotics in the treatment of acute sinusitis per se, although these drugs are frequently administered as a preventive measure. When used, ampicillin, penicillin, cloxacilin, or erythromycin are frequently the drugs of choice. Antibiotics are always indicated in toxic patients and when treatment with decongestants is insufficient. In chronic sinusitis, irrigation and surgical drainage may be required.

Although relatively uncommon, sinusitis can predispose a number of conditions, such as osteomyelitis of the frontal bones (particularly in children), orbital infections, such as orbital cellulitis, septic cavernous sinus thrombophlebitis, among others. These are serious, sometimes life-threatening conditions, but uncommonly seen because of antibiotic therapy.

SIPHON (Zoology). A passage between the mantle folds of bivalve mollusks through which water enters or leaves the mantle cavity. In some species these passages are developed into long muscular tubes and in others they are no more than poorly marked openings. They are two in number, a dorsal or excurrent siphon and a ventral or incurrent siphon. Water is taken in through the latter, passes through the gills to the chambers above them, and flows out through the dorsal siphon.

A part of the mantle border in some of the marine gasteropods also forms a tube through which water can be drawn into the mantle cavity. This tube is known as the siphon.

The term oral siphon is applied to the canal leading to the mouth of ascidians and the opening from the atrial cavity is sometimes known as the atrial siphon.

Unlike these organs, all of them associated with respiration, the siphon of sea urchins (*Echinoidea*) is a slender tube associated with the alimentary tract.

SIRENIA. See **Sea Cows.**

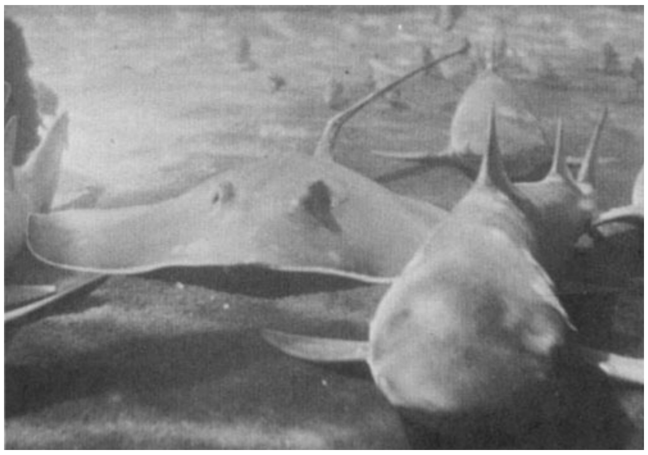
SIRIUS (α Canis Major). This is the brightest star in the sky and volumes have been written concerning its matchless brilliancy. Historically, it is undoubtedly the most interesting star in the heavens and references to it are found throughout all ancient literatures back to the earliest known writings. Aside from its surpassing brilliancy, the fact that it may be observed from every habitable portion of the earth has served to make it an object of veneration by all peoples. Sirius was worshiped in the valley of the Nile long before Rome was even heard of, and many ancient Egyptian temples were so arranged that the light from this star would penetrate to the inner altars.

Sirius has a true brightness value of 23 as compared with unity for the sun. Sirius is a white, spectral type A star and is relatively nearby (π = 0.″375), estimated distance from the earth 8.7 light years. The mass of Sirius is well known because it is a double star. The companion is a white dwarf. See also **Constellations;** and **Star.**

SISKIN. See **Finch.**

SKARN. A petrological term describing the process of contact metamorphism by which certain mineral silicates such as amphibole, pyroxene and garnet replace limestone and dolomite.

SKATES AND RAYS (*Chondrichthyes*). Like sharks, the skates and rays are cartilage fishes. Of the order *Batoidei*, skates and rays may be subdivided into: (1) Electric rays (*Torpedinidae*); (2) guitarfishes (*Rhinobatidae*); (3) sawfishes (*Pristidae*); (4) skates (*Rajidae*); (5) stingrays, whiprays, butterfly rays, and round rays (*Dasyatidae*); (6) eagle rays, bat rays, and cow-nosed rays (*Myliobatidae*); and (7) devil rays (*Mobulidae*). Certain features characterize the skates and rays as compared with other elasmobranches (cartilaginous skeleton, platelike scales, lack of air bladder). Some of the differences include—instead of to the head, the pectoral fins fasten to the sides of the head; the pectoral fins are much enlarged; gills are located on the undersides of the fins; method of respiration is altered to accommodate for bottom dwelling in muddy environs; there is no free upper eyelid as found in sharks.


Sting ray shown in the company of several lemon sharks. (*A. M. Winchester.*)

The electric ray carries a large electric organ under the wings and adjacent to the head. On the average, an electric ray can generate from 75 to 80 volts, but potentials as high as 200 volts have been recorded. Several days are needed by the ray to recharge the electric-producing apparatus. These fishes are of no food value and largely of interest because of their electrical characteristics. As with other fishes having electrical apparatus (see *Catfishes;* and *Gymnotid Eels*); the electrical ability of the fish is used to stun potential or overactive food and for defense. However, it also has been postulated that if the fish can control the discharge to a low, possibly microvoltage level, the apparatus may be used for detection and navigation along the lines of sonar. This would be reasonable in the case of the rays which live in muddy waters where visibility at best is poor. The electric ray holds a positive charge on its upper surface; negative on the undersurface. Approximately 40 species of electric rays are cataloged. The largest genus is *Torpedo*. The American Atlantic *Torpedo* attains a length of 5 feet (1.5 meters) or more and a weight of from 160 to 200 pounds (73 to 91 kilograms). The electric rays are ovoviviparous—eggs layed and hatched within the parent. Widely distributed in ocean waters of temperate or tropical regions, electric rays prefer small crustaceans for their diet and range from the surface down to depths of about 3,000 feet (900 meters).

In appearance, the guitarfish looks a bit like a shark and a bit like a ray, perhaps something like a flattened out shark. The guitarfishes are of nine genera with a total of over 150 species. The common Mediterranean guitarfish (*Rhinobatos rhinobatos*) frequents Mediterranean waters from Angola, Africa to Portugal; the *Rhinobatos horkelii* (Brazilian guitarfish) is found along the South American coast. The American Pacific guitarfish (*Rhinobatos productus*) is found in the waters of the Gulf of California and as far north as Monterey Bay. Guitarfishes are also ovoviviparous. They eat small crabs and fish. They survive quite well in captivity.

The sawfish looks like a guitarfish with a long, double-edged saw fastened to its nose. These fishes are classified as giant rays and six species are known. Records indicate that some species attain a length of up to 35 feet (10.5 meters) and a weight up to 5,000 pounds (2268 kilograms). Sawfishes prefer salt and brackish waters, but also migrate to fresh water. The landlocked Lake Nicaragua has a significant population of sawfishes. They also are found in the Indian River (Florida). The *Pristis pectinatus* is the common western Atlantic sawfish and possesses from 25 to 32 pair of rostral teeth. The *Pristis pristis* (eastern Atlantic sawfish) has fewer pairs of such teeth. The sawfish uses the saw for clubbing its prey and also for digging. In preparing for a meal, the sawfish sweeps through the water flailing the saw, thus wounding its victims, which then are consumed in due course. Even in captivity, the sawfish retains its habit of swinging the saw back and forth during feeding time.

Also an elasmobranch, the skate (*Rajidae*) has a very flat body and a pair of pectoral wings. Cartilage projected forward of the body creates a very elongated nose. Skates prefer tropical and temperate waters and are widely found with exception of the areas of Micronesia, Polynesia, the Hawaiian Islands and the waters off the northeastern coast of South America. They prefer sandy, muddy bottoms of shallow water, preferably less than 600 feet (180 meters) deep. The number of species exceeds 100. The hedgehog skate (*Raja erinacea*) is common and can be found in the western Atlantic from Nova Scotia south to South Carolina. Adults measure about 20 inches (51 centimeters) in length and weigh about a pound. The *Zearaja nasuta* that frequents New Zealand waters attains a length of over 6 feet and may weigh up to 70 pounds. Possibly the largest of these fishes is the *Raja binoculata*, occurring in the American Pacific, which attains a length of about 8 feet (2.4 meters). Some species of skates have electric organs, but the voltages are considered to be quite small. It is possible that the skate may have electrogenic capabilities. Skates found along European coasts are sold commercially, but they are not a major item.

Stingrays are characterized by venom spines. See accompanying diagram. The spines can be replaced if lost. Death can result if the venom gland is driven into a swimmer. It has been recorded that large stingrays can thrust the venom spine through the planking of a small boat. There are approximately 80 species of dasyatid stingrays, ranging from fishes of about $1\frac{1}{2}$ pounds (0.7 kilogram) and about 1 foot (0.3 meter) across the wings to a weight of 750 pounds (340 kilograms) and wingspan of 6 to 7 feet (1.8 to 2.1 meters) as found in the Australian stingray. Stingray young are born alive.

Although enjoying a very bad reputation, the devil rays (*Mobulidae*) under scientific investigation have proved to be quite docile. The *Mobula diabolis* (Australian devil ray) measures but 2 feet (0.6 meter) across the wingtips, whereas *Manta* may attain a wingspan of 22 feet and weigh up to 3,500 pounds (1588 kilograms). They usually travel singly or in pairs, seldom in schools. They are ovoviviparous.

SKELETAL SYSTEM. An aggregation of rigid or semirigid structures that provide mechanical support for the body and usually a lever system on which the muscles act.

The simplest type of skeletal system is found in the sponges (see **Porifera**) and some of the coelenterates. Sponges have scattered spicules of calcareous or siliceous matter among their loosely integrated tissues or are held together by a meshwork of fibers composed of the material spongin, so familiar in the sponges of commerce. The commercial sponge is, in fact, the skeleton freed of organic matter. Spicules are variously formed bodies, some straight rods, some with radiating axes from three to six in number, and some expanded at the ends. Approximately similar scattered bodies are found also in alcyonarians, and in some of these animals they are united to form a continuous mass in which the polyps are imbedded or a solid core surrounded by softer material. Rock corals lay down a calcareous deposit beneath the base to which each polyp of the colony adds.

In many invertebrates the body wall is the sole support of the animal. Even though it is soft, the incorporation of muscular layers in it provides for movement without rigid skeletal structures.

The echinoderms differ in having calcareous plates (ossicles) throughout the integument. In the sea urchins they are closely joined to form a shell and in the sea cucumbers they are small and scattered. Other groups show an intermediate condition with closely associated but movable ossicles. These structures in the sea urchin also illustrate lever action in the use of the spines for locomotion and in the peculiar chewing organ known as Aristotle's lantern.

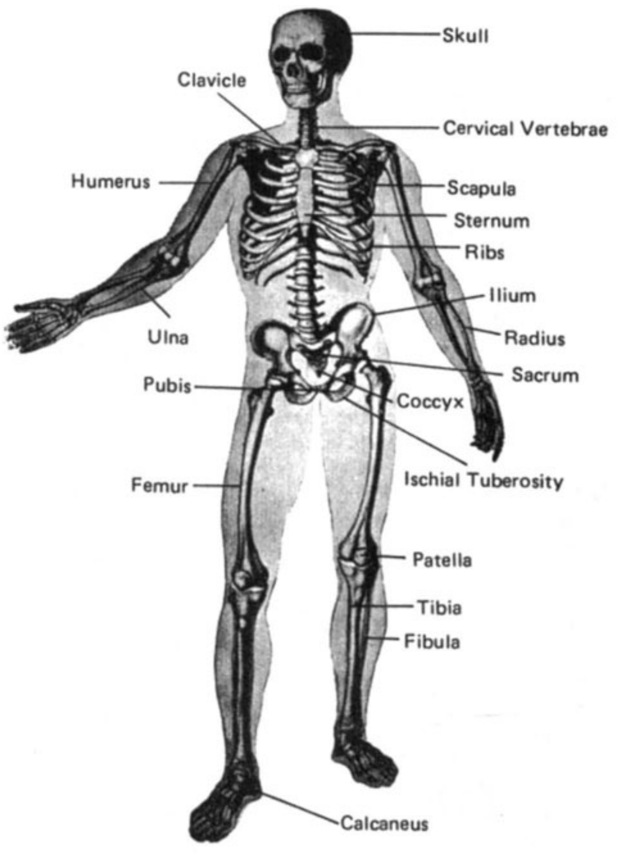


Fig. 1. Skeletal system of human.

The arthropods and vertebrates differ in the presence of a complex skeletal system which supports the body and forms the foundation of the jointed appendages. In no other phyla are leverlike appendages found.

The arthropods have an exoskeleton made up of hard plates (sclerites) of chitin and calcareous matter, formed in circumscribed areas of the cuticula. They are separated by flexible zones which provide for freedom of movement, and are moved by muscles attached to their inner surfaces. The appendages are formed of rigid segments attached to each other and to the body by the flexible tissue of the joints. The skeleton forms a sheath enclosing the muscles in such appendages.

Among the half-million species of arthropods wide variation is inevitable. In general, the exoskeleton of the head forms a compact capsule. The skeleton of the thorax also tends to become compact, although its segments are distinctly recognizable in many species. This portion of the exoskeleton also forms a shieldlike carapace in many species. The abdomen retains greater mobility, although its segments may be reduced in number and the exoskeleton of each may be consolidated to a ring surrounding the body. A moderately complex segment of the exoskeleton contains a dorsal, a ventral, and two lateral sclerites known respectively as the tergum or tergite, the sternum or sternite, and the pleura or pleurites. In many cases this plan is simplified by fusion or made more complex by subdivision of the sclerites.

An endoskeleton of limited extent in the arthropods is made up of internal projections from the exoskeleton known as apodemes. They serve as places of attachment for muscles.

One important result of exoskeletal support is the provision of wings in the insects as thin-walled sacs of the body wall. By the apposition of the upper and lower walls of these sacs they become thin membranous planes sufficiently strong and rigid to support the body in flight. Thus the jointed appendages are freed from participation in adaptation for flight.

Exoskeletal structures also appear in the vertebrates but here they are merely hard parts without the usual supporting functions of the skeleton. They are derived from either or both layers of the skin. The category includes teeth and beaks, horns, claws, hoofs and nails, scales, feathers, and hair. Bony plates associated with scales, as in the alligator and armadillo, are also exoskeletal.

The supporting skeleton of the vertebrates (Fig. 1) is an endoskeleton composed of bones and cartilage. It is made up of two chief divisions, an axial skeleton consisting of a vertebral column, ribs, and skull, and an appendicular skeleton including pectoral and pelvic girdles, each bearing a pair of appendages. The pharyngeal wall of fishes is supported by the visceral skeleton which persists to a limited extent in the more advanced classes. In the elasmobranch fishes now living, the skeleton is composed entirely of cartilage. In most vertebrates it is made up very largely of bone.

The skull in the elasmobranch fishes consists of a mass of cartilage below, behind, and partially enclosing the brain. This structure is called the chondrocranium. In the bony fishes it is replaced by bones in the same position and is supplemented by superficial bony plates enclosing the remainder of the brain. The jaws of the elasmobranch are also supported by cartilages and in the bony fishes these cartilages are supplemented by bones. Above the fishes a chondrocranium appears in the embryo and is replaced during development by the ethmoid bone and parts of the sphenoid, temporal, and occipital bones, taken in order from front to back. The remaining bones are not preformed in cartilage. They are more numerous in the lower form than in humans. In the human skull (Fig. 2) they are the pair of nasal bones in the bridge of the nose, the pair of lachrymals in the orbits, the vomer in the nasal septum, the large frontal bone of the forehead, the parietals in the top and sides of the skull, the upper part of the occipital forming the lower back wall of the skull, and the temporals above the ears. The upper jaw is based on the maxillary bone from which the palatine extends into the hard palate. A slender rod, Meckel's cartilage, supports the lower jaw of the human fetus but the permanent lower jaw is the mandible, a dermal bone formed independently.

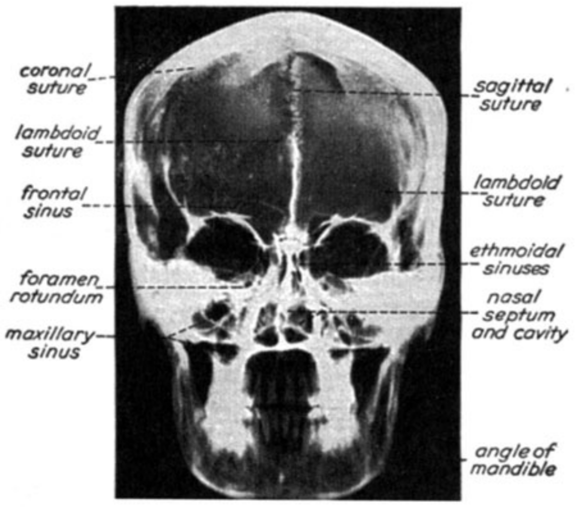


Fig. 2. Antero-posterior radiography of a skull. (*Cunningham.*)

The vertebrae that make up the spinal column consist typically of a centrum bearing a neural arch above and a haemal arch below. The neural arch surrounds the spinal cord and is surmounted by a neural spine. In the fishes the haemal arch encloses the dorsal aorta and in the region of the body cavity open haemal arches form ribs known as fish ribs. In other classes, and in a few fishes, these ribs are replaced by similar slender bones known as true ribs. They develop between the segments of the body lateral to the fish ribs and above them, and articulate with processes of the vertebrae. These ribs join a median ventral structure, the sternum or breast bone, composed of bones and cartilages.

The vertebrae differ in various regions of the body. The first, with which the skull articulates, is the atlas. The second, the axis, is noteworthy for its odontoid process, a solid anterior extension of the centrum. This process is the centrum of the atlas. These two and the remainder of the series in the neck are called cervical vertebrae. The following series with which the ribs articulate are the thoracic vertebrae. The lumbar vertebrae extend to the articulation of the pelvic girdle, where one vertebra or a fused series constitute the sacrum. The following caudal vertebrae lie in the tail or, in the apes and man, are fused to form a mass called the os coccyx.

The pectoral and pelvic girdles are composed of three pairs of bones, two passing downward and toward the median line of the body from the articulations of the appendages and one upward. Of the more constant bones in the pectoral girdle, also called the shoulder girdle, the upper bone is the scapula (shoulder blade), the anterior of the lower bones is the clavicle (collar bone), and the posterior is the coracoid. The two pairs attach to the sternum. They are supplemented in amphibia and reptiles by other bones. The bones of each half of the pelvic girdle are the dorsal ilium, the anterior pubis, and the posterior ischium. This girdle in many forms is firmly attached to the sacrum.

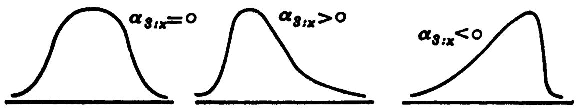
The girdles are modified in various ways. The principal tendency of the pectoral girdle is toward simplification by the loss of bones, so that in many species only the scapula and clavicle remain, as in man, the coracoid being reduced to a process on the scapula. In the hoofed mammals only the scapula persists. In the moles, however, the girdle is large and strong. The pelvic girdle, in correlation with the stresses that it bears in locomotion, becomes compact. The bones of each side tend to fuse and the pubic symphysis, at the ventral union of the two halves, is very firm. The name pelvis or pelvic bone is often applied to this composite structure.

The primitive appendages are the paired fins on the fishes. These structures are precursors of the pentadactyl appendage of terrestrial vertebrates. From this basic plan the structure of appendages in various groups has come by modification of the proportions and relations of the bones, by loss of digits, and to a limited extent by the fusion of separate bones. Both loss and fusion have occurred in the wings of birds, and maximum loss is found in the legs of the horse and allied species, where a single functional digit remains.

The appendages also contain sesamoid bones, small rounded bones developed in the tendons spanning movable joints. The most constant of these bones is the patella or knee cap.

The visceral skeleton in its primitive condition consists of a series of small bones and cartilages supporting the branchial arches between the gill slits. In the existing vertebrates the embryonic primordia of the hinged jaws are derived from these arches but the visceral skeleton is otherwise greatly reduced except in the fishes. It persists in part as the small bones (ossicles) of the middle ear, the hyoid bones supporting the tongue, and the cartilages of the larynx and upper part of the trachea.

SKEWNESS (Mathematics). That property of a frequency distribution involving its symmetry or asymmetry. Skewness is usually measured by the departure from zero of the third *moment* about the mean, standardized by division by σ^3 where σ is the standard deviation.



Demonstration of skewness.

SKIMMER. See **Shorebirds and Gulls.**

SKIN. The covering of the body of vertebrates. It consists of two parts, an outer epidermis and an inner dermis or corium. The former develops in the embryo from the outer germ layer, the ectoderm, and the latter from the middle germ layer, the mesoderm.

The epidermis is composed of many layers of cells in two principal strata, the stratum corneum and the stratum germinativum. The flattened cells next the surface are hardened by deposits of pareleidin, a substance related to keratin, and are said to be keratinized. They make up the stratum corneum. Below them are several layers of thicker cells whose active proliferation gives rise to the cells of the stratum corneum. These layers constitute the stratum germinativum. In the outer cells of this stratum granules of keratohyalin appear as forerunners of the pareleidin of the stratum corneum, forming the thin stratum granulosum. A thin clear zone just outside of the granular layer, known as the stratum lucidum, is regarded as the basal layer of the stratum corneum. In its cells the granules of keratohyalin become a diffuse intermediate substance, eleidin.

The corium is a dense connective tissue layer extending from the fatty subcutaneous tissue. It is obscurely divided into an inner stratum reticulare and an outer stratum papillare which rises in papillae beneath the epidermis. Epidermal derivatives including hair follicles, sweat glands, and sebaceous glands extend into the corium, and it contains nerve endings, tactile corpuscles, and blood vessels. Smooth muscle fibers attached to the hair follicles lie in it, and in some parts of the skin it contains other muscle fibers. The voluntary muscles of the face by which expression is controlled end in it.

Aging of the Skin. The deep, felt-like layer (dermis) gives the skin its suppleness and substance. The connective tissue fibers of the dermis form a woven pattern that, in infants, gives the skin a soft, velvety texture. As the skin ages, this woven pattern becomes tighter and less resilient, tougher, and wrinkled. These essentially undesirable changes are hastened by certain deleterious factors, including overexposure to ultraviolet (UV) rays of the sun, exceptionally large doses of vitamin C (ascorbic acid), exposure to heat, wind, drying agents (such as alcohol), and frigid temperatures. These changes are noted as "wrinkling" and often include the formation of "crow's feet" at the outer corner of the eyes, as well as some withering at the joining of the lips.

To date, these aging processes are not subject to reversal. Moisturizing agents (wool fat and lanolin) as well as extra doses of vitamins A and E and retinoic acid (Retin-A) have been tried, but documentation to date has been less than convincing. At best, these substances may slightly slow the natural skin aging processes.

Diseases of the skin are numerous and are described in the article on **Dermatitis and Dermatosis**. In the following article on **Skin Cancer,** the principle life-threatening skin diseases are described. The principal trauma to the skin includes burns, either from fire and heat or from chemical substances. See **Burn.**

Progress has been made in using "artificial" skin, particularly in the treatment of burn cases. As reported by Erickson, some success has occurred in the development of so-called skin "stand-ins," or dermal substitutes, which may partially replace the need for testing skin medications and cosmetics on laboratory animals. These substitutes are comprised of complex cultures of cells, look like real skin, and are available in sheets. The possible use of such new dermal substitutes (*TestSkin*) ultimately may revolutionize skin-product testing and research and obsolete most needs for laboratory animals. The concept also is being investigated for possible use in burn therapy. (See *Sci. Amer.*, 168, September 1990).

SKIN CANCER. Probably because the skin is the largest organ of the body and because it is so continuously exposed to the hazards of sunlight, ionizing radiation, abrasion, chemical attack, and other hostile conditions, it is more frequently the seat of cancer than any other site on the body. Although not responsible for all carcinomas of the skin, chronic exposure to sunlight, where the individual is not protected by intense melanin pigmentation, is the most important etiologic factor in the development of skin cancer and this itself may be further mediated genetically.

In the early 1990s, the number of newly diagnosed skin cancers per year ranged between 500,000 and 600,000 cases per year in the United States. These statistics are paralleled in other countries located in similar climate zones and with comparable life-styles. Part of the increase in the skin cancer treatment load may be attributed to the growing awareness of the general populace to public and institutional educational programs that have stressed the dangers of exposure to sunlight as well as a growing awareness of individuals to seek professional help more frequently when confronted with changes in their skin and the appearance of doubtful lesions. However, the bulk of the increase is due to new cancers brought about by failure to heed the warnings.

The death rate from skin cancer varies widely with the type of cancer and the timing of diagnosis and therapy. Deaths due to nonmelanoma cancers, especially basal-cell carcinomas, are comparatively low. In 1992, Weinstock et al. reported on a population-based study of mortality from nonmelanoma skin cancers that indicated that 59% of deaths were due to squamous-cell carcinoma and 20% were due to basal-cell carcinomas. Increased risk was associated with cancers on the ears and eyelids.

There were 32,000 deaths in the U.S. in 1991 attributed to cutaneous melanoma. This incidence has nearly tripled since the 1950s. Most increases were in young adult whites (ages 25 to 29 years). The annual incidence is higher than that of any other cancer. Sunlight is considered the most important environmental factor favoring the pathogenesis of melanoma. There appears to be some familial connection, with a risk factor of 2 to 8 times for individuals whose parents had melanoma. It was learned that dysplastic nevi are potential precuror lesions from which melanoma may develop.

The prognosis of malignant melanoma worsens with increasing thickness of the tumor. It also is related to site, age, and sex. Early diagnosis and removal of the lesion usually has been successful in warding off the usual incurable metastatic phase of the disease. In patients with tumors less than 0.75 mm thick, the 5-year survival rate is 96% dropping to 47% in cases where the tumor exceeds 4 mm. The 5-year survival rate falls to 36% in those persons with distant metastes.

In an excellent article by Phillips et al. (reference listed), the authors describe the latest surgical margins in the management of Stage I melanoma (i.e., those limited to the skin).

A number of benign skin lesions may develop into malignancies if left untreated. The most common of these is *actinic keratosis* (senile or solar keratosis) which will occur in areas frequently exposed to sunlight—the face, lower arms, neck, scalp (where not shielded by hair) and the dorsal surfaces of the hands. Varying in size up to one centimeter (diameter), the lesions may be pink to tan, smooth or scaly, and of almost any shape. *Leukoplakia* is much like actinic keratosis except that it will appear on mucous membranes (lips, oral mucosa, vulva, etc.) the lesions being somewhat elevated, irregularly shaped and sharply bordered white patches. Squamous cell carcinoma in situ (*Bowen's disease*) presents as single or multiple sharply defined plaques that are slightly thickened and brownish-red with varying amounts of scale. The lesions resemble eczema or psoriasis, but fail to respond to local therapy.

Although only about 5% of these lesions become invasive and less than 2% metastatize, more important is the recognition that the patient is at significant risk of developing carcinoma of the respiratory, gastrointestinal, and genitourinary systems. Intraepidermal metastases (*Paget's disease*) is an uncommon carcinoma presenting as a sharply defined, red, weeping, crusted or scaly lesion resembling atopic eczema which does not respond to topical corticosteroids. It may develop within the epithelium of the mammary ducts and extend upward into the epidermis or appear on the pubis, perineum, genitals and other sites related to underlying apocrine or eccrine sweat glands.

Basal Cell Carcinoma. This accounts for more than 75% of all skin cancers and these are among the most strongly antigenic of all human cancers. They arise from the epidermis, cytologically resemble the normal basal cells, and show little tendency to undergo the usual differentiation into squamous cells which produce keratin. Although these tumors rarely metastasize, they are locally invasive and may go deeply into nerves, bone, and brain. Like most cancers, they are remarkably painless in their course and this lack of symptoms leads to prolonged neglect of a lesion. The typical basal-cell carcinoma is an uninflamed smooth, waxy nodule which appears to be translucent and may show varying amounts of melanin pigment in the form of small dots. Such nodules often ulcerate and may reepithelialize leading the patient to believe that the nodule is resolving. A patient with one basal-cell carcinoma is likely to have others at the same time or shortly thereafter. Because the tumors are malignant, they should be excised or otherwise destroyed after diagnosis has been confirmed. A border of normal tissue, adjacent to the tumor, must be removed to prevent recurrence from invasive strands of tumor cells.

Squamous Cell Carcinoma. This also arises from the epidermis, but shows significant squamous differentiation and usually keratin production. The lesions have a variable tendency to metastasize depending upon their size, extent of invasions, and their mode of origin. The typical squamous-cell carcinoma is a painless, firm, red nodule or plaque with visible scales on the surface. Ulceration may occur, in contrast with the basal-cell variety and, when this is so, they have the lowest frequency of metastasis. On the other hand, where they arise from mucous membranes, burn scars, sinus tracts, or from apparently normal skin, they have a much higher tendency to metastasize.

Primary Malignant Melanoma of the Skin. This is the most serious disease in dermatology. Pigmented moles are among the most common growths on the skin of humans, yet cancer involving pigment cells is relatively uncommon, constituting about 1.5% of all cancers. The diseases is, however, virtually uncurable by chemotherapy or x-rays and so far has only responded to deep excision.

Even in its early stages, primary malignant melanoma is relatively easy to detect. The variegation of color and irregularities in surface pattern and configuration are characteristic even when the tumor has red, white, or blue surface color. Any of the several types of melanoma may exist for several years in the preinvasive stage.

These tumors are usually seen in the fourth and sixth decades of life. White people with light complexions are more prone to melanomas than persons with dark pigmented skin and there is a significant relationship between incidence and exposure to sun. Three forms are recognized: (1) *Lentigo malignant melanoma* with a median onset of 70 years. The lesions are various shades of brown and black and the margin is flat. The lesions are usually found on the head and neck, being less common on hands and legs. They usually commence as freckles with irregular outlines and may grow very slowly over 5 to 15 years. (2) *Superficial spreading melanoma* is the most common form, with a median onset of 55 to 60 years. The lesions may have several different colors from brown to gray and rose-pink. The margins are distinctly palpable and the lesions are found on all body surfaces, beginning as small irregular, brown pigmented areas with various shades of color. Growth into papules or nodules may require from 1 to 5 years. (3) *Nodular melanoma* has a median age of onset of about 50 years. The lesions are uniformly bluish-black with a depigmented halo and the margins are palpable. Usually commencing as a papule or nodule with a smooth, scaly, eroded surface, the lesions grow rapidly.

Survival rates are highest for the lentigo type and may be as high as 95%. In nodular melanoma where distant metastases frequently form, the prognosis is much poorer. Immunotherapy or chemotherapy when combined with surgery may improve the situation.

Sarcomas of the Skin. These may be primary or metastatic and single or multiple. They involve connective tissue, the vascular system, adipose tissue, muscle and/or the lympho-reticular system. Kaposi's idiopathic hermorrhagic sarcoma is manifested by firm, reddish-brown or bluish plaques or nodules on the hands and feet.

Mycosis Fungoides. This is the most common primary lymphoma (tumor of lymphoid tissue) of the skin and technically is a relatively indolent cutaneous T-cell lymphoma characterized by a neoplastic proliferation of mature helper T cells. Each year, about a thousand new cases are presented in the United States. However, the incidence is increasing. The lesion may present as a nonspecific erhythematous, scaly, eczematous eruption that may go on for years before a diagnosis is made. It gradually develops thicker, scaly, annular plaques reminiscent of psoriasis. In the accelerated phase the disease is characterized by skin tumors, lymph-node involvement, and a spreading to the viscera. These conditions are associated with a survival of 2.5 years. Patients who present only superficial skin involvement have a better prognosis (some 12-plus years). Sézary syndrome, where there are plaques, tumors, erythrodermic, and lymph-node or blood involvement, but no visceral involvement, has a median survival of 5 years.

Protection from the Sun

Life-styles have changed a lot during the last few decades, and, unfortunately, the incidence of skin cancer has increased markedly. Cases of skin cancer in earlier years were almost always associated with outdoor manual laborers, and thus a tanned skin was shunned by many of the elite. In recent years, particularly with the younger generation, a tanned skin has become a status symbol by people who regard a nicely tanned body as a symbol of health and strength. Young males and females alike have taken to sun bathing, using tanning parlors, and taking tanning pills even though during the last few years numerous public and private institutions have warned against excessive exposure to solar rays and its relationship with skin cancer.

While tanning pills generally have been considered hepatotoxic (adverse effects on the liver), a number of satisfactory and safe UV blocking and screening lotions have become available. The labels on these products should be read carefully to make certain that sufficient protection will be provided over what periods of exposure time. But, even with such safeguards, sunbathers must always act in moderation.

Professional dermatologists suggest:

1. Always wear sunglasses that absorb from 90 to 100% of UV light. Glasses with gray lenses are preferred to avoid undue alterations in color perception. Glasses which wrap around the side provide protection against reflected-light trauma. Sunglasses protect against *solar keratitis,* which is a sun-induced inflammation of the cornea of the eye. The glasses also provide long-term protection against later cataract development.
2. Apply lip balm to lips, zinc oxide paste to the nose, and wear a broad-brimmed hat. A weekly lanolin facial treatment is excellent for maintaining a good skin condition.
3. Use a sun block (screen) over all areas of the body that will be exposed to the UV light. Light-skinned people require a highly protective blocking agent (rated 12 to 20). Dark-skinned people can use a lighter blocking agent.
4. Under no circumstances should babies or older family members be permitted to bask in the sun. Exposure for just a few minutes can be damaging.
5. Do not relax precautions on cloudy days. Such days are are a special hazard because clouds are radiolucent, offering no barrier to UV rays.
6. The foregoing also apply when in the water. A water film provides no protection and, in fact, may partially remove the blocking agent (usually PABA [para-amino benzoic acid]).

R. C. Vickery, M.D., D.Sc., Ph.D., Blanton-Dade City, Florida.

Additional Reading

Koh, H. K.: "Cutaneous Melanoma," *N. Eng. J. Med.*, 171–182 (July 18, 1991).
Phillips, T. J., et al.: "Recent Advances in Dermatology," *N. Eng. J. Med.*, 167 (January 16, 1992).
Sun, M.: "Shieseido Grant (Skin Cancer Research)," *Science*, 810 (August 21, 1989).
Weinstock, M. A., et al.: "Non-Melanoma Skin Cancer Mortality," *Arch. Dermatol*, 1194 (July 1991).

SKIN DISEASES. See **Dermatitis and Dermatosis.**

SKIN EFFECT. The alternating current flowing in a conductor is not uniformly distributed over the cross section of the conductor but becomes denser the farther the cross section being considered is from the center of the conductor. This effect, known as skin effect, is caused by the varying inductance of the different parts of the conductor. If a circular conductor is considered to be specific, and imagined to be composed of many elemental filaments in parallel it can be seen readily why the current flow is more dense as the distance from the center becomes greater. The current in each filament sets up a flux around it which links both it and any other filaments which may be within the lines of flux. It will be realized at once that the filaments in the center can be linked by many more lines than those on the outside, since the outer ones are beyond much of the flux of the inner ones. Thus the inductance of the inner filaments is greater than that of the outer ones and so the reactance of the inner ones is greater. This greater impedance of the inside causes more of the current to flow in the outer-layers and gives rise to the skin effect. As the effective cross-section of the conductor is decreased by this, the resistance to ac is greater than the resistance to dc. In certain cases, such as the induction motor, where the conductor is largely surrounded by iron, the skin effect is very much greater than in a conductor in free space. In this motor the effect is utilized to obtain high starting torque without serious impairment of the running characteristics. In the layout of ac buses, various arrangements such as hollow squares, channels, etc., are used to save material since the inner part would be of very little use. At radio frequencies the effect is very pronounced so only a thin outer layer of the conductor is effective, thus greatly increasing the resistance. It is convenient to speak of a "skin depth" which is defined as the depth below the surface of a conductor at which the current density has been reduced to $1/e$ times its value at the surface of the conductor. The skin depth δ is given by the expression

$$\delta = \sqrt{\pi f \mu T}$$

where δ is depth in meters, f is frequency in cycles/second (Hz), μ is the permeability in henries/meter and T is the conductivity in mhos/meter.

SKIN FRICTION. The drag force on a body arises from tangential stresses at its surface, usually viscous, from the normal pressure distribution over it, and, for supersonic flow, from wave-drag. The component calculated from the tangential viscous stress is the skin friction and, for a bluff body, is the component most sensitive to change of Reynolds number. Roughly, the skin friction coefficient, the stress per unit area divided by the dynamic head of the free stream is proportional to the one-fifth-power of the Reynolds number.

SKINKS. Of the class *Reptilia,* subclass *Lepidosauria*, order *Squamata* (scaly reptiles), suborder *Sauria* (lizards), infraorder *Scincomorphs* (skinks and allies), these animals are found in three subfamilies: *Tiliquinae,* which includes the Giant, Cape Verde, Stumped-tail, Blue-tongued, and Spiny-tailed skinks; Scincinae, which includes the common skink and the Eastern, Hemprich's, Arabian, Sand, and Algerian skinks, among many others; and Lygosominae, which has many genera, including the East Indian Brown-sided, Emerald, Spotted, Casque-headed, Dart, and Blind skinks, among others. Closely related and part of *Scincomorpha* are two additional families, the tailed lizards and night lizards, which are described under **Lizards**.

Skinks are distinguished by the head, which is covered on top with large, symmetrically arranged, ossified plates, and by the round scales on the back and belly that overlap like roof shingles. The tongue is free and moderately long. It is slightly notched at the end and bears imbricating scalelike papillae. The body is usually cylindrical, and the head ends in a sharp snout. The feet in most species bear five toes; the tail is pointed at the tip. The skinks demonstrate a gradual transition from lizards that have four strong legs to the legless "snake-type" lizard. In this transition, it is hypothesized that, initially, the trunk became longer, the limbs became smaller and more delicate, and the number of fingers and toes was reduced until finally becoming stumpy vestiges, with the forelegs disappearing entirely and, lastly, the hindlegs vanishing as well. Thus, the manner of locomotion changed from walking to snakelike slithering. These changes reflected a move from the original mode of life above ground to the specialized life of the subterranean burrower. These changes are reflected in varying degree among the numerous genera of skinks. The form of the tail also was involved in the slowly evolving transition. Some genera of skinks, such as the Australian stumped-tailed skink (*Tiliqua rugosa*), has only a stump tail, whereas others have a spinelike tail. Ear openings also changed, wherein the external ear opening is found reduced and has become covered with scales. This may have occured as the result of the animal's transition to a subterranean form of life—to protect or even close the openings of sense organs. In many skinks, according to I. E. Fuhn, the lower eyelid, customarily movable and scaly in lizards, evolved into a transparent window that allows the animal to see through when it is closed. If the lids fuse, a rigid, transparent "spectacle" is formed, much as in snakes. Where skinks colonized rocky regions or trees, it became advantageous to develop adhesive mechanisms on the fingers and toes. There are exceptions, as in the case of the giant skink, which evolved a prehensile tail.

Skinks are *thermophilous* (heat-loving) and are found mainly in the tropical and subtropical regions of the earth, such as southern Asia, Africa, and regions of Australia and Polynesia. Some, however, are found in the Americas and Europe. Because of their highly terrestrial preference, skinks are not usually found near bodies of water.

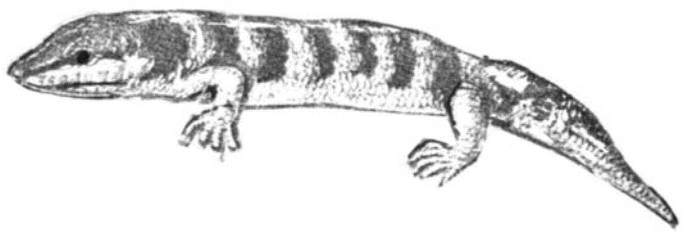
Formal classification of the skinks is based largely upon the structure of the skull, thus accounting for the three families previously mentioned.

Fuhn suggests that the *Tiliquinae* are relicts of the Tertiary Period. The giant skink (*Corucia zebrata*) is, on the average, about 65 centimeters long. It is presently found in the tropical forests on San Cristobal in the Solomon Islands. The largest of all modern skinks, this animal lives in trees, a characteristic that is rare among skinks. Especially adapted to arboreal life, its characteristics of long prehensile tail, strong limbs armed with claws, large ear and eye openings, broad head, and short, blunt nose are distinguishing features. The upper body is greenish-white, marked with irregular brown bands; the head is sometimes a reddish-brown. This animal has not been studied extensively in its habitat.

Also of *Tiliquinae* is the Cape Verde skink, which at one time was a favorite animal for terrariums. It is relatively easy to sustain in captivity, but, because of extensive hunting for its flesh by fishermen in its natural habitat, the species is threatened. A close relative of this creature became extinct as early as 1708.

Among the *Tiliqua,* the blue-tongued skink (*Tiliqua scincoides*) is well known and named because of its cobalt-blue tongue. Also studied is the stump-tailed skink (*T. rugosa*), which achieves a length of about 36 centimeters. The latter is found in Western Australia, South Australia, and New South Wales. Notably, this skink is quite short and wide, has a flat tail that is widened at the end and, because of its round end, that appears to have been chopped short (sometimes referred to as "bob-tail"). The threat display of the animal has been described by Robert Mertens: "It rolls itself into an arc and, hissing, opens its mouth to show the wide, long, cobalt-blue tongue in sharp contrast to the red mouth. Though its mouth, armed with strong teeth, is wide open, the animal does not bite, but rather tries to scare its enemy away simply by its threat display." The "bobtail" often lives on sandy dunes and can burrow into upper layers of sand. It moves slowly, but in an emergency can run quite fast. The animal's diet consists of vegetation, fruit, and small animals. The blue-tongued skink is ovoviviparous, the female bearing two or three young at a time. The animal has been featured in terrariums.

Of the subfamily *Scincinae,* the common skink is the best representative of this group. See accompanying illustration. The species (*Scincus scincus*) generally is adapted to live in the desert or other sandy regions. The characteristic ability of these lizards to move about rapidly in the sand was noted in early Biblical times. At one period, in the late 1700s, it was believed that the lizard had aphrodisiac and medicinal properties. All members of this species have sturdy, cylindrical bodies. The tip of the snout has an elongated wedgelike shape, and the tail, conically shaped, is short. The legs each bear five toes with serrated fingers. The coat of scales is remarkably smooth and firm.


Common skink (*Scincus scincus*).

Scincinae also includes the genus *Chalcides,* with its well-known cylindrical skinks, which are distributed in the Mediterranean, from eastern Africa and Arabia into Iran, Pakistan, and India. The species *Chalcides chalcides* is a snakelike animal, with three toes on the limbs. This animal may reach a length of about 42 centimeters. During rapid locomotion, the legs are pressed against the body and not used. The upper side is olive to bronze in coloration, with up to 11 dark, longitudinal stripes. The tail is moderately long, and the head is not set off from the trunk. The limbs are short, with five toes. The smooth and shiny scales are brown. These animals are ovoviviparous, bearing about three young at a time. A popular terrarium specimen is *C. ocellatus*. Its natural habitat is Morocco and environs. It prefers valleys in which date palms grow or else groves of tamarisk on clay soil, where it uses the tracks formed during dry periods as hiding places. This skink is active all day until sunset, during which time it feeds on crickets. Because the animal has a snakelike appearance, there are numerous superstitions about it. For terrariums, Franz Werner observes: "It loves the sun and needs a warm, sunny container with a sandy floor and stones under which it can withdraw at twilight. One can feed it with small mealworms and flies. The skink can become quite tame and will come out to take food handed to it."

SKIP DISTANCE. As the frequency of a radio wave is increased, the minimum angle of incidence at which the wave will be reflected from the ionosphere rather than pass on through becomes greater. This means that the higher the frequency the farther from the transmitter the reflected sky wave strikes the earth. This distance between the transmitter and the point closest to it at which the sky wave can be received is the skip distance. The ground wave is attenuated more rapidly the higher the frequency so that at high frequencies there may be a region in which the ground wave has become too weak for use and in which the sky wave cannot be received because of its skip. Above about 4 MHz this effect becomes very noticeable, the dead region for higher frequencies running to a distance of a few hundred miles from the transmitter.

SKIPJACK TUNA. See **Tunas.**

SKIPPER (*Insecta, Lepidoptera*). An insect much like the butterflies but in many ways intermediate between the butterflies and moths. Most species are distinguished by their knobbed and hooked antennae. The few that lack the hook are less easily distinguished from the butterflies except by the veins of the wings or by other details of structure. Most of the skippers of the temperate zones are of moderate size and modest colors but many tropical species are brilliant. They constitute the superfamily *Hesperioidea*.

The name skipper refers to the vigorous and often erratic flight of these insects. Their small wings and powerful muscles accompany rapid flight. Many species perch readily and make short darting flights from place to place.

SKUA. See **Shorebirds and Gulls.**

SKUNK. See **Mustelines.**

SKUTTERUDITE. This mineral includes an isomorphous series with *smaltite-chloanthite*, essentially cobalt/nickel arsenides, (Co, Ni) As_{2_3}, crystallizing in the isometric system. The usual habit is cubic, octahedral, or cubo-octahedral. The mineral also occurs in massive and granular forms. Skutterudite has a metallic luster; hardness of 5.5 to 6.0, a specific gravity of 6.5. The mineral is opaque with tin-white to silver-gray color. The nickel-rich material alters surficially to annabergite (green color); the cobalt-rich material to erythyrite (rose color). The streak is black. The mineral is an essential ore of cobalt and nickel.

Skutterudite is found in moderate-temperature veins, commonly associated with other cobalt/nickel minerals, e.g., cobaltite and nickeline. The mineral was named for its occurrence at Skutterud. Norway. Important ore sources are Norway, Bohemia, Saxony, Spain, France, and New South Wales, Australia. Notable occurrences are in Ontario, Canada, mainly Sudbury, South Lorrain, and Gowganda.

SKY COVER. In surface weather observations, a term used to denote one or more of the following: (a) the amount of sky covered but not necessarily concealed by clouds or by obscuring phenomena, i.e., any atmospheric phenomenon that obscures a portion of the sky from the point of observation; (b) the amount of sky concealed by obscuring phenomena that reach the ground; or (c) the amount of sky covered or concealed by a combination of (a) and (b).

"Opaque" sky cover is the amount of sky cover that completely hides everything above it; *transparent sky cover* is that portion of sky cover through which higher clouds, blue sky, etc., may be observed; and *total sky cover* is the two taken together. Sky cover, for any level aloft, is described as "thin" if the ratio of transparent to total sky cover at and below that level is one-half or more.

Sky cover is reported in *eighths* under international observing and reporting practice. The "0" indicates $\frac{0}{8}$, or clear; and "8" indicates $\frac{8}{8}$, or total cover.

Sky cover is also reported in *tenths*, so that 0.0 indicates a clear sky; 1.0 ($\frac{10}{10}$) indicates a completely covered sky. According to the *summation principle* used in weather observing practice, the sky cover at any level is equal to the summation of the sky cover of the lowest layer plus the additional sky cover provided at all successively higher layers up to and including the layer in question. Thus, no layer can be assigned a sky cover that is less than a lower layer, and no sky cover can be greater than 1.0. If, at any level, the ratio of transparent sky cover to total sky cover is one-half or more (excluding transparent portions of surface-based obscuring phenomena), the layer at that level is classified as thin.

The term *ceiling*, in aviation weather observing practice, refers to the ascribed height of the lowest layer of clouds or obscurring phenomena when that layer is reported as "broken," "overcast," or "obscuration," and not classified "thin" or "partial." When none of these conditions is satisfied, the ceiling is termed "unlimited." Whenever the height of a cirriform cloud layer is unknown, a slant "/" is reported in lieu of a height value; at all other items, the ceiling is expressed in feet above the surface.

The following classifications of sky cover are used in aviation weather observations:

a. Clear: sky cover less than 0.1 (no ceiling).
b. Scattered: sky cover 0.1 to 0.5 (clouds or obscuring phenomena aloft only; no ceiling).
c. Broken: sky cover 0.6 to 0.9 (minimum requirement for a ceiling; must be some clouds or obscuring phenomena aloft).
d. Overcast: sky cover 1.0 (must be some clouds or obscuring phenomena aloft).
e. Partial obscuration: sky cover 0.1 to 0.9 (surface-based obscuring phenomena only).
f. Obscuration: total sky cover (surface-based obscuring phenomena only).

See also **Clouds and Cloud Formation;** and **Weather Technology.**

Peter E. Kraght, Certified Consulting Meteorologist, Mabank, Texas.

SKYLARK. See **Lark.**

SLAG. Slag is a fused product occurring in connection with metallurgical and combustion processes. It is composed of the oxidized impurities in a metal, and of a fluxing substance, and of ash. In the steel industry, slag is the neutralized product of anhydrous compounds entering into the process. Slag is of great importance to the operator of a steel furnace or a cupola, in that, through the slag, impurities are separated and removed from the metal. By floating as a molten covering on the pool of metal, slag protects it from oxidation and serves to keep it clean. By controlling the character of slag, and continuous observation, the metallurgist insures that the metal is of the quality desired.

Molten ash is one of the products of combustion of coal in certain high-capacity boiler furnaces. It is also called slag. In some plants, the ash is removed from the furnace in this fluid form. Such furnaces are known as slag tap furnaces. Slag has some commercial value as ballast, coarse aggregate for concrete, road metal, etc.

SLATE. A fine-grained homogeneous sedimentary rock composed of clay or volcanic ash which has been metamorphosed (foliated) so as to develop a high degree of fissility or slaty cleavage which is usually at a high angle to the planes of stratification. This high degree of fissility makes the better grades of slates an extremely useful roofing material which, however, has been somewhat replaced in recent years by synthetic and manufactured substitutes. The finest slates in the world come from Wales, Britain.

SLEEP. Recurrent physiological loss of consciousness at regular intervals, predominantly in humans once in 24 hours and lasting for a more or less constant time of an average of about 8 hours for adults. The newborn infant sleeps 18 to 20 hours a day. The periods of sleep last from 2 to 3 hours, and usually are interrupted by hunger. As the baby's stomach grows and more food is consumed, the periods of hunger are farther apart and the periods of sleep are longer. When the child is 6 months old, it sleeps less—16 to 18 hours a day. At the age of about one year, the child sleeps about 12 hours at night, with usually a morning and an afternoon nap. Although it may be more difficult for older people to get the necessary sleep, there is little if any truth that older people do not require as much sleep as younger persons. Some of the problems of the aged can be derived because of a lack of sustained sleep. The amount of sleep for any person varies with each individual. The prime requisite is that the individual should sleep long enough to awaken rested and refreshed.

The human body cannot work around the clock. It must have a regular opportunity to catch up on its repair activities and to dispose of wastes that have accumulated during the day faster than they can be discarded. During sleep, the entire body slows down. The liver stores starches needed for the next day. The kidneys clean the blood. The rate of metabolism is at its lowest, being just sufficient to keep the vital parts of the body in operation. The blood pressure drops; the pulse rate is slower; and breathing is irregular and slowed down. The body is less sensitive to stimulation by pain, light, or sound. Even the temperature may drop by as much as a degree. Since the brain does not have muscles, it does not tire in the same way the body does. But there is evidence that the brain requires sleep for other reasons, particularly as sleep is related to memory.

The autonomic nervous system and the involuntary muscles, such as those of the heart and the respiratory mechanism, continue to work in sleep, at a slightly reduced rate, so that the whole body does not share equally in the recuperative process. A number of bodily changes are constantly observed in sleep; these include deep muscular relaxation, loss of some reflex activity, slowing of the heart rate and reduction in volume of many glandular secretions. Deep sleep brings no dreams; those which accompany light sleep are generally the result of external stimuli received at the time, and which, in the absence of the critical faculty of the cerebral cortex, are greatly misinterpreted.

Although research continues, there is no fully satisfactory explanation of the phenomenon of sleep. In particular, the biochemical aspects of sleep have received comparatively little detailed attention. There are many ancillary problems which have been investigated, but the central one—what is accomplished biochemically each day by long hours of sleep—has been attacked only a relatively few times and with no marked success. The symptomatic treatment of insomnia by the use of sleep-inducing drugs is largely on an empirical basis. Knowledge of the fundamental reasons behind a person's vital need to sleep can be obtained only by further biochemical investigation.

The amount of sleep enters into an assessment of the efficiency of the human body. Just how much work is actually obtained for the fuel that has been fed the body? When gasoline is burned in an engine, only about one-eighth of the energy released can be captured to drive the motor, and the remainder is lost in the form of the heat that is given off. Such an engine is said to have an efficiency of 12.5% because it wastes 87.5% of the fuel through heat loss. Some diesel engines have a 32% efficiency. If one were to calculate the work produced by a person performing strenuous work, it would be found that the efficiency of the human body is about 20%. Much of the fuel burned by the body is converted to heat to maintain the body temperature and is lost by the body to the surroundings without doing any real work. Since the human body cannot work regularly the full 24 hours of a day, its efficiency is in reality considerably lower—perhaps about 10 to 12%. Nevertheless, the human body is built to last a great many years as contrasted with most other energy-consuming mechanisms. Some of the basic accomplishments of the human body during a 24-hour period include:

—Kidneys filter over 150 quarts (142 liters) of fluids.
—Heart beats over 100,000 times.
—1,500 gallons (5678 liters) of blood are pumped through the body.
—Over 10 million red cells are destroyed every second, all of which must be replaced at an equally rapid pace.

—Nerves carry millions of messages.
—Eyelids blink thousands of times to cleanse and lubricate the eyes.

Electroencephalography (EEG) reveals a pattern of large slow waves which appear when a subject falls asleep. As shown in Fig. 1, this pattern can be made to appear experimentally in cats by partially cutting through the midbrain at a certain level. It is thought possible that such an incision cuts off the vital centers in the midbrain from certain postulated driving centers elsewhere in the brain, and that with the stream of awakening impulses removed, sleep occurs. But the means by which this process occurs almost unfailingly every 24 hours in the human being remains obscure. EEG and eye-movement studies, however, are shedding some light on the process.

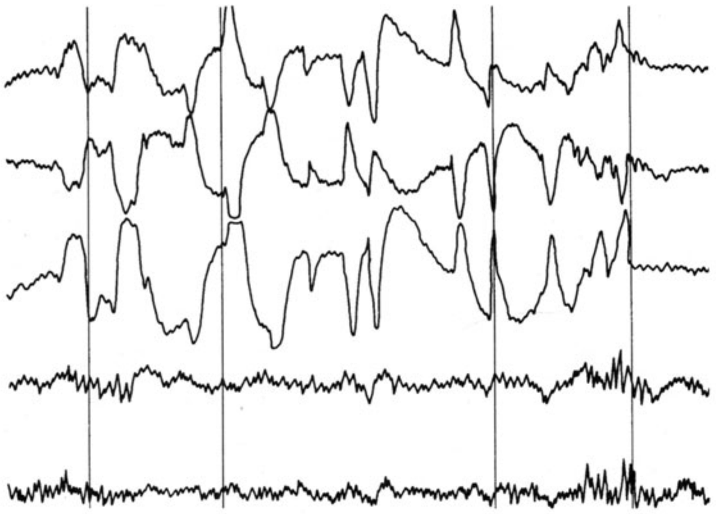


Fig. 1. Alpha. The subject drowses, on the edges of sleep. Top three EEG channels show rolling movements of right eye, left eye, and both. Two bottom channels trace brain's electrical changes. The even pattern of alpha rhythm indicates relaxation.

The alpha waves grow smaller as the subject passes through the gates of the unconscious. The eyes are very slowly rolling. For a moment, the subject may wake up during the early part of this descent, alerted by a sudden spasm that causes the body to jerk. This, like the brain waves, is a sign of neural changes within. It is known as the *myoclonic jerk*, resulting from a tiny burst of activity in the brain, somewhat related to that of an epileptic seizure. The myoclonic jerk, however, is normal in all human sleep. It is gone in a fraction of a second and descent continues. The subject does not note the peculiar transformation as true sleep is approached. The lower chart records of Fig. 1 indicate Stage 1 of sleep. The pattern of the sleeper's brain waves is small and pinched, low, irregular and rapidly changing. Occasionally the regular waves of the alpha rhythm break through. The sleeper may be enjoying a floating sensation or drifting with idle thoughts and dreams. The muscles are relaxing, the heart rate is slowing. The subject could be awakened easily and might insist that actually sleep had not occurred. But, after a few minutes at this port of entry, if not disturbed, the subject descends again to another level, another step removed from the real world.

As the sleeper passes into Stage II, the brain waves change again. Now they trace out quick bursts—a rapid crescendo and decrescendo, resembling a wire spindle, and unmistakable on the EEG chart. The eyes roll slowly from side to side, but if the experimenter gently opens a lid the sleeper will not see. If awakened, which can be accomplished easily with a modest sound, the subject may still think that actual sleep has not occurred. At this point, however, the subject has been soundly asleep for perhaps 10 minutes. Whatever happens now in the subject's imagination will be totally beyond conscious grasp.

Still the sleeper descends to Stage III. This is characterized by large slow waves that occur about one a second. Sometimes they are about five times the amplitude of the waking alpha rhythm. It will take a louder noise to awaken the subject, perhaps some repetition of the person's name. The muscles are very relaxed and the breathing is even. The heart rate has become slower; body temperature declines; blood pressure drops.

Some 20 to 30 minutes after the subject fell asleep, the sleeper reaches Stage IV, the deepest level. This is marked by large, slow brain waves, called delta waves (Fig. 2) that trace a pattern resembling jagged buttes. Stage IV is a relatively dreamless oblivion. The breathing is even, heart rate, blood pressure, and body temperature slowly falling. But, the sleeper does not remain in Stage IV. After 20 minutes or so in this depth, the subject begins to drift back up through the lighter levels toward the surface. By the time the subject has been asleep for about 90 minutes, brain waves of the lightest sleep, even resembling those of waking, will be indicated. Still the sleeper is not easy to awaken, lying limply, the eyes moving jerkily under closed lids as if watching something. This is a special variety of Stage I, known as REM (for rapid eye movement) sleep. See Fig. 3. If awakened during this period, the subject would almost certainly remember dreaming and most often in vivid detail. After perhaps 10 minutes in the REM state, the sleeper will probably turn over in bed and begin shifting down the levels of sleep again to the depths, only to return in another hour or so for a longer REM dream. Each night, the entire cycle may be repeated 4 to 5 times.

In an excellent paper on sleep and memory (see references), Fowler, Sullivan, and Ekstrand describe experiments conducted to determine the relationship of sleep with ability to remember. The interference the-

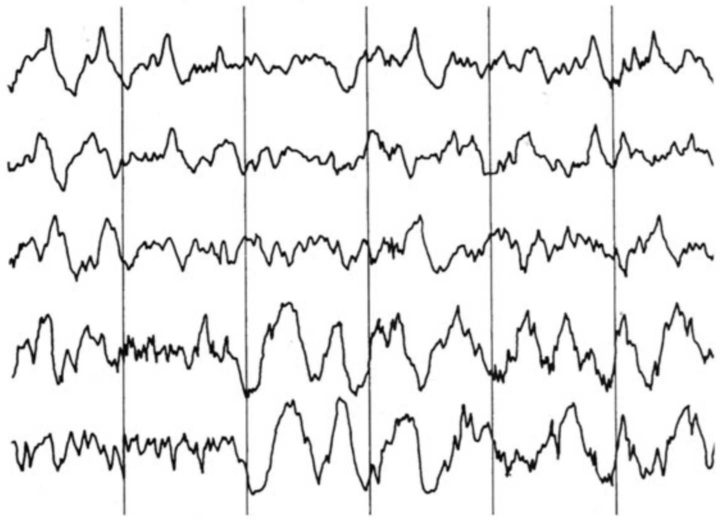


Fig. 2. Delta. Marked by large, slow brain waves (bottom channels). This is the deepest stage of sleep, the first to be made up for after total deprivation.

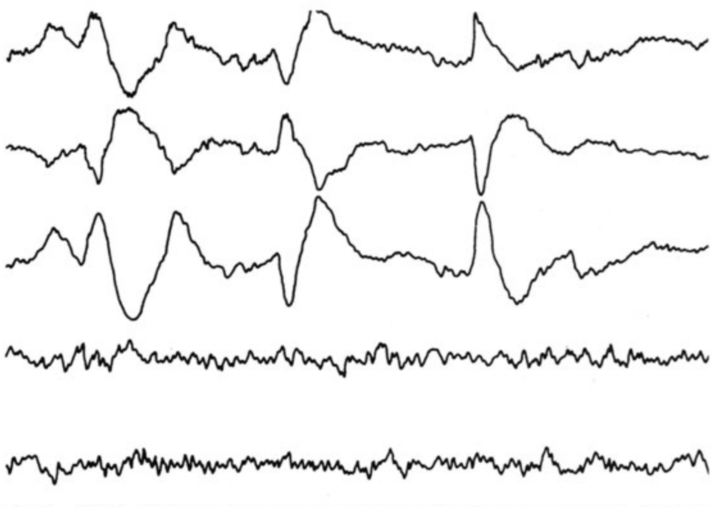


Fig. 3. REM. Rapid eye movements, spaced in the top three channels, indicate dreaming. This is the most dramatic of the night's phases. Although brain waves are low, a psychological storm is raging in breathing, pulse, brain temperature, and hormone circulation.

ory of forgetting essentially states that forgetting is due to interference from learning taking place during the retention interval. It then would possibly follow that by putting subjects to sleep, recall could be facilitated. The experiments described in the reference paper demonstrated that memory over an interval with relatively high amounts of rapid eye movement (REM) sleep was inferior to memory over an interval with relatively high amounts of Stage IV sleep. The findings to date suggest that, at least for humans, REM sleep does not facilitate memory consolidation and that Stage IV sleep may be beneficial to memory.

In studying patients complaining of insomnia, Guilleminault, Eldridge, and Dement uncovered a new clinical syndrome, sleep apnea (partial suspension of breathing) sometimes associated with insomnia. These findings came out of combining respiratory studies with sleep research. Quoting briefly from the aforementioned reference. "Sleep apneas have been reported in the cardiopulmonary syndrome of obesity (Pickwickian) and in other syndromes involving hypersomnia, such as narcolepsy. The apneas in hypersomniacs seem to be temporarily associated with sleep. Several distinct types of sleep apnea have been defined in these conditions. They include a "central" type, characterized first by cessation of breathing and then, after the apnea, by a simultaneous resumption of diaphragmatic movements and oral airflow; an "obstructive" or "peripheral" type characterized by the interruption of airflow secondary to upper airway obstruction, but with continuance of diaphragmatic and thoracic muscle contraction; and a "mixed" type, characterized by an initial central apnea followed by temporary upper airway obstruction at the subsequent resumption of diaphragmatic movements." It is concluded from these early studies that some unknown percentage of patients complaining of chronic insomnia may have profound disorders of respiratory control mechanisms, with resulting disruptive sleep. The investigators conclude with the suggestion that respiratory function during sleep should be evaluated in patients who complain of chronic insomnia characterized by several conscious arousals throughout the night and early morning and who also have a short latency before onset of sleep and a history of heavy snoring.

Somnambulism (sleepwalking) and talking while asleep are symptoms most commonly observed in persons with conflicts. Some investigators ascribe the sleepwalking (and other actions performed by a person who is apparently otherwise asleep) to underlying unconscious distress. The phenomenon indicates that certain brain areas, more than others, are inhibited during sleep.

SLEEPING SICKNESS. See **African Trypanosomiasis.**

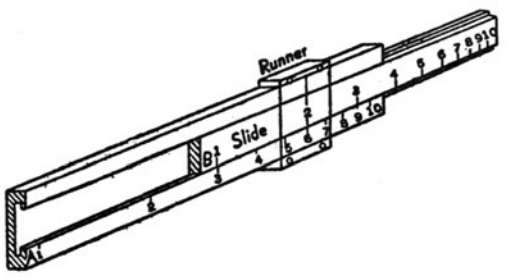
SLEET. See **Precipitation and Hydrometeors.**

SLICKHEAD FISHES (*Osteichthyes*). Of the order *Isospondyli*, family *Alepocephalidae*, these are very deep sea fishes, rarely seen in collections. They have slender bodies and are of a dark color and considered quite a small fish. The *Dolichopteryx* is equipped with "telescopic eyes."

SLIDER (*Reptilia*, *Chelonia*). Turtles related to the painted turtles and land tortoises. The several species are brownish or greenish, marked with yellow and in some cases also with red. They constitute the genus *Pseudemys*. Most of the species are confined to the southern states but one, known as the red-bellied terrapin, is found near coastal rivers from Florida to Cape Cod and another lives in the Mississippi valley as far north as Iowa. Both of these species are edible. Also called cooters.

SLIDE RULE. A hand-held, manually-operated analog computer for multiplying, dividing, extracting roots, and obtaining powers of numbers mechanically by logarithmic means. In construction, the simple slide rule consists of two adjacent logarithmic scales which may be so set that a reading on one is added to a reading on another. This represents the addition of logarithms when multiplying numbers. The slide rule for general use is found in several different forms, and there are many special slide rules such as stadia rules, electrical rules, hydraulic computing rules, etc.

The accompanying illustrates the principle of multiplication by the slide rule. Two logarithmic scales, A and B, are arranged on a slide rule so that they may be mechanically added. In the illustration given, the process of multiplying three by two is shown. The end of the B scale is aligned over the 3 on the A scale, so that to 3 on the A scale may be added 2 on the B scale. Now if these scales were uniformly divided, the result, of course, would be five units on the A scale, but since they are logarithmically divided, the result on the A scale is the product rather than the sum. Thus, under the 2 on the B scale, one reads the product of $3 \times 2 = 6$ on the A scale. A glass runner is provided so that the alignment of the numbers of the two scales may be facilitated. By halving the divisions, two complete scales could be placed on the rule. Using a full scale, such as A of the figure, and a double scale, one can extract square roots. A triple scale can be used for cube roots. Other scales contained on the ordinary slide rule are a uniform division scale for logarithms, and scales for reading the value of two trigonometric functions of an angle, the sine and tangent.



Simple slide rule.

With the advent of fast, low-cost, solid-state electronic pocket and portable calculators, the conventional engineer's slide rule has largely been replaced, but not fully because the availability of plastic materials and manufacturing techniques has greatly lowered the cost of the slide rule. Customized slide rules directed to the solution of specific problems remain quite popular, particularly among manufacturers of equipment who frequently provide slide rules to users of their products as an aid to application, maintenance, and procurement decisions. The principal features of a simple slide rule are shown in the accompanying diagram.

SLIP RINGS. These are conducting rings attached to a rotating part of an electrical machine to make connection through brushes with the stationary part of the circuit. They are used where it is not necessary to commutate the current being conducted.

SLIT. The long narrow opening by which radiation enters or leaves certain diffraction instruments. Slits are often used as line sources of radiation or of particles, and combination of two or more slits are employed as a collimator.

SLOPE. In rectangular coordinates, the ratio of the change of the ordinate to the corresponding change of the abscissa of a point moving along a line. If a straight line is determined by the points (x_1, y_1) and (x_2, y_2) its slope $m = (y_2 - y_1)/(x_2 - x_1)$. If the line is not straight its slope is given by

$$\lim_{x_2 \to x_1} m = (dy/dx)_x = {}_{x_1}$$

which is the tangent or derivative at the point.

See also **Coordinate System.**

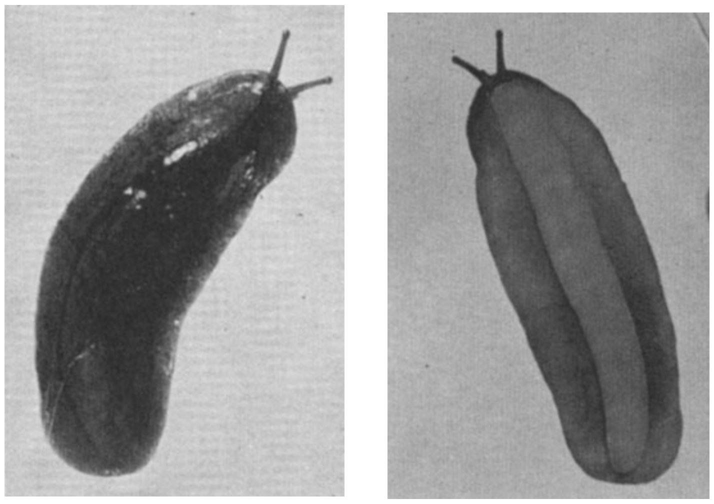
SLOTH-BEAR. See **Bears.**

SLOTHS. See **Edentata.**

SLUDGE. When fresh sewage is admitted to settling tanks a certain amount of the solid matter in suspension will settle out, 50% more or less for sedimentation periods of an hour and a half or so. This collection of solids is known as fresh sludge. Such sludge will become actively putrescent in a short time and in modern treatment plants must be passed on from the sedimentation tank before this stage is reached. This may be done in two common ways. The fresh sludge may be passed through the slot in an Imhoff tank to the lower story or digestion chamber. Here, decomposition by anaerobic bacteria takes place with considerable liquefaction and reduction in volume. After the decomposition process has run its course (in 6–9 months) the resulting sludge is called "digested" sludge and is relatively inoffensive in character. It may be disposed of by drying on sludge drying beds and spreading on the land. It has little, if any, fertilizing value, being in the nature of humus. The sludge digestion chamber is operated on a periodic schedule of sludge withdrawals.

Alternatively, plain sedimentation basins with mechanical equipment for continuous collection of the fresh sludge may be used. The fresh sludge, so collected, is discharged into separate sludge digestion tanks which operate on the principle of the lower story of the Imhoff tank except that by means of higher and better temperature control the digestion cycle is much more rapid and efficient than for the Imhoff tank.

SLUG (*Mollusca, Gasteropoda*). Not an insect, but a mollusk, this pest is damaging to certain food crops and is a serious pest in glasshouse operations, notably on certain ornamental plants, such as coleus, geranium, marigold, and snapdragon. The slug is also a severe pest on lawns, particularly on dichondra. The pest eats very large holes, often entire leaves, always at night. The creatures are fleshy, wormlike, gray-to-brown, slimy. They are without body segmentation and are legless. Slugs range up to 3 inches (7.5 centimeters) in length. During the day, the pests hide under leaves, debris, etc., but in a heavy infestation, they emerge at night in the scores. Some of the species include: the *spotted garden slug* (*Limax maximum*, Linne); the *gray garden slug* (*Deroceras reticulatum*, Müller); the *gray field slug* (*D. laeve*, Muller); and the *greenhouse slug* (*Milax gagates*, Linne).


Slugs. (*A. M. Winchester*.)

The control measure most frequently used is poisoned bait, placed in suspected locales at night. The quantity required varies with the strength of the mixture used. Ingredients frequently used include bran as the base, plus calcium arsenate, metaldehyde, blackstrap molasses, and water. The molasses is used as an attractant.

The *pear slug* that damages cherry, pear, and plum trees is not a true slug, but rather the larva of a species of sawfly. See **Sawfly.**

SLUG TUNING. A means for varying the frequency of a resonant circuit by introducing a slug of material into either the electric or magnetic fields or both.

SLUICE. A channel through which water issues is, in some cases, called a sluice. A sluice may be a pressure conduit, or it may be an open flume. For instance, the word may be applied to an artificial channel to lead water from one point to another, especially a temporary wooden channel. Sluices are often incorporated in dam structures. A sluice through a dam is a conduit cast in the concrete and equipped with controls called sluice gates. The purpose of the sluice is to empty the reservoir if necessary, to control the water level, and to aid in passing floods.

SLURRY EXPLOSIVES. See **Explosive.**

SMALLPOX. This disease caused by a large, enveloped poxvirus featuring a complex helical nucleocapsid structure containing double-stranded DNA and several enzymes, was once described as the most devasting pestilence in human history. Worldwide eradication culminating in the last reported case in late 1977 (patient in Merka, Somalia recovered) indeed marked a major hallmark in the professions of public health and preventive medicine. In 1967, when the World Health Organization (WHO) instituted a campaign to eradicate smallpox from the last reservoirs on earth, the disease was still endemic in 33 countries and 11 other countries reported cases. The eradication process started many years ago, of course, with mandatory vaccination policies in a number of countries and with particularly tight control over travelers who could carry the disease from one country to the next. The thrust of the last decade by WHO concentrated on mass vaccinations of the populace of countries where the disease still persisted, as in Brazil, Indonesia, Bangladesh, India, Afghanistan, and Africa south of the Sahara. In the last few years of the WHO program, the emphasis also included strict isolation of smallpox patients. At a time when it appeared that the disease had finally been overcome, an outbreak occurred in 1977 among nomads in eastern Africa and in Mogadishu (capital of Somalia) during which time nearly 1400 cases were reported. One of the final steps was a last major mass vaccination program in Somalia.

Smallpox (also called variola), an acute contagious disease characterized by high fever, chills, headache, generalized pain, and vomiting and with an accompanying skin eruption that resulted in severe scarring, has been known since antiquity and lesions have been seen on the skin of an Egyptian mummy of the twentieth dynasty. The classical scarring which labeled a person for life as a victim of smallpox was often accompanied by development of blindness. The disease originally spread from India and Central Asia to Europe and became especially prevalent at the time of the Crusades in the eleventh century. It was first introduced in America by a Negro slave of Cortex in 1520, producing an epidemic that killed several million people. One century later it appeared in New England. It was first called smallpox in Europe, to differentiate it from grand-pox (syphilis of the skin). The discovery of vaccination by Jenner in 1798 is one of the landmarks in the development of medical science.

Black or hemorrhagic smallpox, a virulent fulminating form of the disease, commonly caused death within two to six days. Fatalities from smallpox often arose from complications, including abscesses, pneumonia, septicemia, nephritis, and ulcers on the cornea of the eye. In recent decades, these complications were treated with chemotherapy. However, no specific drug to assure the cure of smallpox had been found by the time of its eradication.

R. C. Vickery, M.D.; D.Sc.; Ph.D., Blanton/Dade City, Florida.

Additional Reading

Basu, R. N., Jezek, Z., and N. A. Ward: "The Eradication of Smallpox from India," World Health Organization, Publications Centre, Albany, New York, 1980.

Henderson, D. A.: "The Eradication of Smallpox," *Sci. Amer.* (October 1976).

Razzell, P.: "Edward Jenner's Cowpox Vaccine," Caliban Books, Firle, Lewes, Sussex, England, 1977.

Razzell, P.: "The Conquest of Smallpox," Caliban Books, Firle, Lewes, Sussex, England, 1977.

SMECTIC LIQUID CRYSTALS. See **Liquid Crystals.**

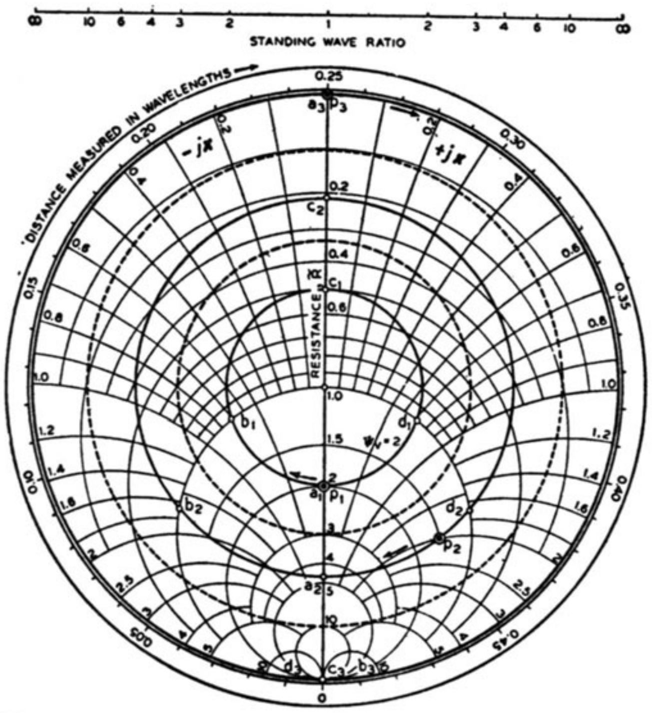
SMELTING. The process of heating ores to a high temperature in the presence of a reducing agent, such as carbon (coke), and of a fluxing agent to remove the accompanying rock gangue is termed smelting. Iron ore is the most abundantly smelted ore. It contains about 20% gangue (clay and sand). The ore is heated in an air blast furnace with coke and limestone (fluxing agent) at a temperature above the melting point of iron and slag (fusion mixture of impurities and flux). The molten iron (the more dense material) and molten slag (the less dense material) are removed separately from the furnace. See also **Arsenic; Cadmium; Cobalt; Copper; Indium; Iron Metals, Alloys, and Steels; Lead; Silver;** and **Tin.**

SMELTS (*Osteichthyes*). Of the order *Isospondyli*, family *Osmeridae*, smelts (more properly termed true smelts) seldom exceed 14 inches in length, and thus classified as a small fish. They prefer the cold or temperate in-shore waters of the northern hemisphere. The osmerid smelts are frequently confused with the atherinid smelt (silverside family). See also **Silversides (Osteichthyes).** The true smelts have a small adipose fin on the dorsal fin, which is absent in the silversides.

Most true smelts, numbering some 13 species, are found in the Pacific, with no occurrence in the Indian Ocean or the southern hemisphere. They can be extremely abundant. From time to time, for example, the *Spirinchus thaleichthys* (Sacramento smelt) has occurred in extremely large numbers in San Francisco Bay. Smelts are, of course, logical prey for larger fishes. Fisheries for adult smelts have long prospered.

The *Thaleichthys pacificus* (eulachon) is found in the waters off the American Pacific northwest. Nicknamed "candlefish," the fish, when dried and because of its excessive oil content, can be burned as a candle or torch, a practice sometimes used by Indians. The *Hypomesus olidus*, which attains a length up to 12 inches (30 centimeters), is found in the Sacramento River delta (fresh water), whereas its Japanese counterpart is a marine form, frequenting fresh waters only for spawning. The *Osmerus mordax* (American Atlantic smelt) occurs from Virginia northward to the Gulf of St. Lawrence. In the early 1900s, this fish was introduced to the Great Lakes where it has done very well. On the Pacific coast, ranging from southern California northward to Alaska, *Hypomesus pretiosus* (surf smelt) prefers the surf. It is a popular bait. The *Osmerus eperlanus* (European smelt) is well regarded commercially and occurs in numerous European rivers, including the Seine.

SMITH DIAGRAM (or Smith Chart). A diagram with polar coordinates, developed to aid in the solution of transmission line and waveguide problems. It is composed of the following sets of lines: (1) constant resistance circles; (2) constant reactance circles; (3) circles of constant standing-wave-ratio; (4) radius lines representing constant line-angle loci. The chart employs normalized quantities for maximum flexibility.



Smith diagram showing circles of constant standing wave ratio, each corresponding to a particular terminal impedance as follows: (a) The terminal impedance (p_1) is $\tilde{Z}(1) = 2 + j0$. (b) The terminal impedance (p_2) is $\tilde{Z}(1) = 1.5 + j2$. (c) The terminal impedance (p_3) is $\tilde{Z}(1) = 0 + j0$.

SMITHSONITE. Smithsonite is zinc carbonate, $ZnCO_3$, a hexagonal mineral with a rhombohedral cleavage. It is a brittle mineral; hardness, 4–4.5; specific gravity, 4.3–4.5; luster, vitreous to dull; color, usually white, but may be colored yellowish or brownish or perhaps blue or green due to impurities. It is translucent to opaque. Smithsonite is a secondary mineral after sphalerite or may replace limestone or dolomite. It is sometimes called calamine (but true calamine is a zinc silicate) and often associated with it. Smithsonite occurs in Siberia, Greece, Rumania, Austria, Sardinia, Cumberland and Derbyshire, England; New South Wales, Australia; South West Africa, and Mexico. In the United States it is found in Pennsylvania, Wisconsin, Missouri, Arkansas, and Utah. This mineral was named in honor of James Smithson, whose legacy founded the Smithsonian Institution at Washington, D.C.

SMOG AND SMOKE POLLUTION. See **Pollution (Air).**

SMOKE TREE. See **Cashew and Sumac Trees.**

SMOOTHING (Of a Curve). The replacement of a curve, or of a sequence of points by another that is in some sense more regular, and yet whose ordinates, for any abscissa, are changed as little as possible. The irregularities in a sequence of points may be due to errors in measurement. If theory requires the theoretically correct points to lie on a given curve, one may apply some method of curve fitting, possibly least squares. If not, one may select arbitrarily a simple function, possibly a polynomial, and fit it by least squares. If the purpose is merely to obtain a smooth graph, this may be drawn visually. A somewhat more sophisticated method is to take, say, 5 consecutive points, fit a parabola, and replace the middle point by the one on the parabola. The next parabola requires four of these points and one new one.

See also **Neighborhood of a Point.**

SMUT FUNGI. See **Fungus.**

SNAIL (*Mollusca, Gasteropoda*). A name properly applied to most members of this class, but usually given only to the small fresh-water and terrestrial species with coiled shells. The term does not correspond with any part of the scientific classification of mollusks beyond this general application. See accompanying illustration.

Snail.

Snails are regarded as a food delicacy in many countries of the world. Closely related to the slug, the snail is a garden pest with tastes and habits much the same as those of the slug. See also **Mollusks;** and **Slug.**

SNAIL DARTER. See **Perches and Darters.**

SNAKE FLY (*Insecta, Neuroptera*). A peculiar insect found only in the western United States. It has four membranous wings and the body is prolonged at the anterior end. These insects are insect eaters and are found commonly on the bark and foliage of trees.

SNAKES. According to one view, primitive snakes were specialized burrowers, whereas according to another view, they were surface-dwelling or aquatic boalike snakes from which specialized burrowers evolved in one line and fast-moving terrestrial and arboreal types in another. Snakes are in the class *Reptilia* (reptiles), the order *Squamata* (scaly reptiles), and the suborder *Serpentes* (snakes). Although there are variations to this classification, this is the classification followed by Grzimek (1971).

About 3000 species and subspecies of snakes range from southern Canada, northern Sweden and the Kamchatka Peninsula southward throughout Africa and most of the Americas, Australia, and temperate and tropical islands except Polynesia (only one tiny species, for example, occurs in Hawaii). There are no snakes in Ireland or New Zealand. Islands have been populated by snakes from adjacent mainland or by subsequent chance. Ireland lacks them because it was separated from continental contact before the last glacier departed, and no introduction has since occurred.

As shown by the accompanying table, there are 3 infraorders, 12 families, and nearly 20 subfamilies, but this number has been variously estimated in accordance with the degree of significance attached to specializations of diverse groups. Even with these complications, a few generalizations can be made.

The most primitive families of snakes possess internal vestiges of the skeletal girdle of the hind limbs, and in some a vestige of the thigh (femur) remains, often visible externally as a small spur on either side of the anal opening. Snakes do not possess movable eyelids. The skin over the eye is in the form of a large scale flush with the rest of the cranial surface in the most primitive snakes. In more advanced forms, the scale fits the eye perfectly and is designated a *brille* or tertiary spectacle. Snakes sleep, but little outwardly visible evidence reveals when they are asleep.

The ears cannot directly receive airborne sounds, since the external ear and middle ear cavity are missing. The middle ear ossicle or columella is present, but is embedded in muscle except where it fits into the *fenestra ovalis* of the capsule of the inner ear. Snakes can detect some ground-borne vibrations, since the inner ear is largely intact. They are deaf to music of snake charmers and to their own noises. The tongue in all snakes is long, slender and forked at the tip. It functions primarily in picking up airborne particles that go into solution on its moist surface. By placing the tips of the tongue into the paired vomeronasal organs at the anterior tip of the roof of the mouth, the snake can detect odors.

Snakes periodically molt the outer epidermal layers of the entire skin, including eye covering and tips of tongue. Frequency varies with species and health, but is not known to be less than once per year. Molting commonly occurs three times per year, but more frequently if the snake has been injured. The molt usually proceeds posteriorly, starting at the snout and lower jaw where the skin is rubbed backward over the head, turning it wrongside out and leaving it usually in one piece. In preparation for the molt, the basal layer of epidermis rapidly proliferates, producing a completely new outer epidermis, with cuboidal and squamous cells. Breakdown of the basal cuboidal cells of the older layer forms a milky fluid, giving a dull cast to the snake as a whole and a milky appearance to the eye; at this time snakes can see but dimly if at all and tend to remain inactive and hidden from view. They are more likely during this period to be defensive if disturbed. The fluid disappears by evaporation or absorption after about three days, and about two days thereafter the slough is shed.

Many snakes lay eggs. Others bear their young live. Parental care is generally absent. The eggs may be abandoned after deposition in damp sand, soil, sawdust, ground litter, or rotting logs. Eggs hatch in about three days to three months. Some viviparous types possess a placenta. As pointed out later, some species, such as cobras, pay more attention to their eggs than other species.

Like other reptiles, snakes are ectothermic, their body temperatures being controlled largely by behavior and to only a slight degree by intrinsic mechanisms.

Males court by repeatedly rubbing chin and body along length of the female's body. In rare circumstances the two animals intertwine the bodies and rise straight upward until only tail and rear part of the body remain on the ground. Ordinarily the latter behavior is a "combat dance" between males. Snakes, like lizards, possess two copulatory organs (hemipenes). Copulation occurs with either and with both structures, but not with both at once, and persists for long periods, from one-half to several hours.

Snakes have numerous well-developed salivary glands, the secretions of which have been shown experimentally to be powerfully toxic in almost all species. Yet the bites of most snakes produce no signs or symptoms of poisoning because of the absence of venom-conducting fangs. So-called nonpoisonous snakes have only "solid" teeth, lacking a groove down one side. All "poisonous" species possess one to four pairs of fangs. Fangs are teeth specialized for injection of venom, having either a tubelike structure acting as a needle, or a groove down one side which conducts venom like a capillary tube when it is imbedded in a victim's flesh.

Snakes have teeth, usually all conical, slender, curved, of about the same size and shape. They are found on five paired jaw bones: the maxilla, palatine, pterygoid, dentary, and rarely the premaxilla. They vary in total number from about 20 to 200. Three sorts of fang modification have evolved. Snakes possess venoms containing numerous protein enzymes capable of promoting various types of tissue degeneration. Snakes feed upon animals, never plant material. They often scavenge. Snakes usually swallow the food head first. The trachea can be extended forward below the food far enough to enable the snake to breathe while swallowing.

Four modes of locomotion are concertina, sidewinding, rectilinear, and horizontal undulation.

Blind Snakes, Primitive Snakes, and Wart Snakes

The infraorder of blind snakes (*Scolecophidia*) contains two snake families whose members burrow in the ground: (1) Blind snakes (*Typhlopidae*), and (2) slender blind snakes (*Leptotyphlopidae*). These two families bear such a remarkable resemblance that it may appear as if they are very closely related and belong to the same family. However, differences in skull structure and skeletal elements cause them to be divided into distinct families. These are very primitive snakes; they have small, poorly differentiated scales that remain small on the belly and do not form the large scutes or ridges seen on more advanced snake species, and they have vestiges of upper hind leg bones and the pelvic girdle. Some zoologists believe that blind snakes are actually closer to some lizards than to other snakes.

Both blind snake families are characterized by a subterranean mode of life. These snakes have a smooth, round body that makes them look like big worms, and a blunt head that merges smoothly with the rest of the body. The tail is short and often has a spiny structure whose function has not been clarified. Perhaps the spine helps anchor the snake in the ground so the body can move more easily. Some species use the spine as a weapon. The eyes are tiny and are covered by large head scales, but they gleam through these scales. The eyes are probably used to distinguish light and dark. The skull is used a great deal as blind snakes burrow beneath the ground, and it is shortened and forms a hard capsule.

The *common blind snake* (*Typhlops braminus*) has a habit of finding its way into flower pots that are left in the open for long periods of time. This species crawls about among the roots and feeds on the insect larvae and other creatures that may be dwelling there. This snake occurs in Madagascar, India, Sri Lanka, southeast Asia, and the Indo-Australian islands, but is also found in southern Mexico. Other blind snake genera occur only in tropical South America.

The *Texas blind snake* (*Leptotyphlops dulcis*) and the western blind snake (*L. humilis*) occur as far north as the sandy, dry prairies of the southern United States, where they burrow in the ground, beneath rocks, or inside fallen trees. They often crawl about on the ground in the early evening hours, but they disappear again after dark. Their diet consists of insect larvae, ants, and termites. They usually lay a few elongated eggs.

The primitive snakes (*Henophidia*) comprise five primitive, nonpoisonous families, three of which have just a few species. Pipe snakes (*Aniliidae*) are three genera from South America and southern and southeastern Asia. They have vestiges of the pelvic girdle and the hind limbs. The hind limb remnants look like small claws on each side of the anal opening. In contrast to blind snakes, the head scales on pipe snakes have not fused much. The eyes are small, but do not lie beneath scales. The scales on the belly are clearly larger than those on the back, although they are not the size of those found in advanced snakes.

The *false coral snake* (*Anilius scytale*) has a length of some 75–85 centimeters (25–34 inches) and lives in South America. It is brilliant red and has black rings on its back, giving it an appearance not unlike poisonous coral snakes. It lacks the yellow bands. This snake bears live young. Diet consists of lizards and snakes.

CLASSIFICATION OF SNAKES

Class REPTILIA (Reptiles)
Order SQUAMATA (Scaly Reptiles)
Suborder SERPENTES (Snakes)

Infraorder: BLIND SNAKES (*Scolecophidia*)
Family: Blind Snakes (*Typhlopidae*)
Examples: Peter's blind snake (*Typhlops dinga*); Common blind snake (*T. braminus*)
Family: Slender Blind Snakes (*Leptoryphlopidae*)
Examples: Texas blind snake (*L. dulcis*); Western blind snake (*L. humilis*)

Infraorder: PRIMITIVE SNAKES (*Henophidia*)
Family: PIPE SNAKES (*Aniliidae*)
Examples: False coral snake (*Anilus scytale*); Pipe snake (*Cylindrophis rufus*)
Family: SHIELD-TAILED SNAKES (*Uropeltidae*)
Family: SUNBEAM SNAKES (*Xenopeltidae*)
Example: Sunbeam snake (*Xenopeltis unicolor*)
Family: ACROCHORDIDS (*Acrochordidae*)
Example: Javan wart snakes (*Acrochordus javanicus*); Indian wart snake (*Chersydrus granulatus*)
Family: PYTHONS AND BOAS (*Boidae*)
Subfamily: *Loxoceminae*
Example: Mexican python (*Loxocemus bicolor*)
Subfamily: *Pythoninae*
Examples: Reticulate python (*Python reticulatus*); Indian python (*P. molurus*); Rock python (*P. sebae*); Blood python (*P. curtus*); Timor python (*P. timorensis*); Ball or royal python (*P. regius*); Angola python (*P. anchietae*); Carpet python (*Morelia argus*); Brown python (*Liasis fuscus*); Boa (*Bothrochilus boa*); Black-headed python (*Aspidites* and *Chondropython*); Burrowing python (*Calabaria reinhardtii*)
Subfamily: Boas (*Boinae*)
Examples: Wood snakes (*Tropidophis*); Cuban boa (*Epicrates angulifer*); Rainbow boa (*Epicrates cenchris*); Rubber boa (*Charina bottae*); California boa (*Lichanura roseofusca*); Madagascar boa constrictor (*Acrantophis madagascariensis*); Sand boas (*Eryx*); Brown sand boa (*Sanzinia johnii*); Emerald tree boa (*Corallus caninus*); Boa constrictor (*Boa constrictor*); Anaconda (*Bunectes murinus*); Yellow Anaconda (*Eunectes notaeus*)
Subfamily: *Bolyeriinae* (Round Island Boas)

Infraorder: ADVANCED SNAKES (*Xenophidia* or *Caenophidia*)
Family: COLUBRID SNAKES (*Colubridae*)
Subfamily: *Xenoderminae* (Xenodermin Snakes)
Subfamily: *Sibynophinae*
Subfamily: *Xenodontinae*
Examples: Eastern hognosed snake (*Heterodon platyrhinos*); Western hognosed snake (*H. nasicus*)
Subfamily: Natricinae
Examples: Ringed snake (*Natrix natrix*); Diced snake (*N. tessellata*); Common water snake (*N. sipedon*); Diamond-back water snake (*N. rhombifera*); Brown water snake (*N. taxispilota*); Glossy water snake (*Regina rigida*); Queen snake (*R. septemvittata*); Kirtland's water snake (*Clonophis kirtlandii*); Common garter snake (*Thamnophis sirtalis*); Mexican garter snake (*T. eques*)
Subfamily: *Colubrinae*
Examples: King snakes (*Lampropeltis* spp.); Milk snake (*L. triangulum*); Common king snake (*L. getulus*); Green snakes (*Opheodrys* spp.); Racer snakes (*Coluber* spp.); Eastern coast coachwhip (*Masticophis flagellum*); Speckled racers (*Drymobius*, spp.); Patch-nosed snakes (*Salvadora* spp.); Leaf-nosed snakes (*Phyllorhynchus*, spp.); Rat snakes (*Elaphe* spp.); Rat snake (*E. obsoleta*); Fox snake (*E. vulpina*); False water cobra (*Hydrodynastes gigas*); Indian rat snake (*Ptyas korros*); Keeled rat snake (*P. carinatus*)
Subfamily: *Calamarinae* (Dwarf Snakes)
Example: Linne's dwarf snake (*Calamaria linnaei*)
Subfamily: *Lycodontinae* (Lycodontine Snakes)
Examples: Indian wolf snake (*Lycodon aulicus*); African wolf snake (*L. capense*); Cape file snake (*Mehelya capensis*); Brown house snake (*Boaedon fuliginosus*); Rainbow snake (*Farancia erythrogrammus*); Mud snake (*F. abacura*)
Subfamily: *Dipsadinae*
Subfamily: *Pareinae*
Subfamily: *Dasypeltinae* (Egg-eating Snakes)
Examples: African egg-eating snake (*Dasypeltis scabra*); Indian egg-eating snake (*Elachistodon westermanni*)
Subfamily: *Homalopsinae* (Homalopsine Colubrid Snakes)
Examples: White-bellied water snake (*Fordonia leucobalia*); Tentacled snake (*Erpeton tentaculatum*)
Subfamily: *Boiginae* (Boigine Snakes or Boigine Vipers)
Examples: Tree snakes (*Boiga*, spp.); Mangrove snake (*B. dendrophila*); Twig snake (*Thelotornis kirtlandii*); Green vine snake (*Oxybelis fulgidus*); Flying snakes (*Chrysopelea* spp.)
Family: COBRAS (*Elapidae*)
Examples: King cobra (*Ophiophagus hannah*); True cobras (*Naja* spp.); Asian or Indian cobra (*N. naja*); Egyptian cobra (*N. haje*); Spitting cobra (*N. nigricollis*); Black-lipped cobra (*N. melanoleuca*); Cape cobra (*N. nivea*); Gold's tree cobra (*Pseudohaje goldii*); Desert black snakes (*Walterinnesia* spp.); Shield-nose snakes (*Aspidelaps* spp.); Water cobras or water snakes (*Boulengerina* spp.); Black mamba (*Dendroaspis polylepis*); Mamba (*D. angusticeps*); Kraits (*Bungarus* spp.); Oriental coral snakes (*Calliophis* spp.); Long-glanded coral snakes (*Maticora* spp.); American coral snakes (*Micurus* spp.); Southern coral snake (*M. frontalis*); Eastern coral snake or Harlequin snake (*M. fulvius*); Western coral snake (*Micruroides* spp.); Arizona coral snake (*M. euryxanthus*); Death adder (*Acanthophis antarcticus*); Australian tiger snakes (*Notechis* spp.); Australian copperhead (*Denisonia* spp.)
Family: SEA SNAKES (*Hydrophiidae*)
Subfamily: *Laticaudinae*
Examples: Black-banded sea krait (*Laticauda laticauda*); Olive-brown sea snake (*Aipysurus laevis*)
Subfamily: *Hydrophiinae*
Examples: Yellow sea snake (*Hydrophis spiralis*); Beaked sea snake (*Enhydrina schistosa*)
Family: VIPERS (*Viperidae*)
Examples: True vipers (*Viptera* spp.): Adder or common European viper (*V. berus*); Asp viper or European asp (*V. aspis*); Sand viper (*V. ammodytes*); Mountain viper (*V. sznthia*); Saw-scaled viper (*Echis carinatus*); Horned viper (*Cerastes cerastes*); Puff adders (*Bitis* spp.); Rhinoceros viper (*B. nasicornis*); Dwarf puff adder (*B. peringueyi*); Bush vipers or tree vipers (*Atractaspis* spp.); Common burrowing viper (*A. irregularis*); Night adders (*Causus* spp.); Common night adder (*C. rhombeatus*)

(continued)

CLASSIFICATION OF SNAKES *(continued)*

FAMILY: PIT VIPERS (*Crotalidae*)

Examples: Rattlesnakes (*Crotalus* spp.); Santa Catalina rattlesnake (*C. catalinensis*); Eastern diamond-back rattlesnake (*C. adamanteus*); Western diamond-back rattlesnake (*C. atrox*); Prairie rattlesnake (*C. viridis*); Timber rattlesnake (*C. horridus*); Green rattlesnake (*C. lepidus*); Sidewinder (*C. cerates*); Tropical rattlesnake (*C. durissus*); Pygmy rattlesnakes (*Sistrurus* spp.); Lance-head snakes (*Bothrops* spp.); Asian lance-head snakes (*Trimeresurus* spp.); Bushmaster (*Lachesis mutus*); Haly's viper (*Agkistrodon halys*); Himalayan viper (*A. himalayanus*); Chinese copperhead (*A. acutus*); Malaysian moccasin (*A. rhodostoma*); Cottonmouth or water moccasin (*A. piscivorus*); Copperhead (*A. contortrix*); Tropical moccasin (*A. bilineatus*)

NOTE: Some of the better known as well as other species selected at random are included in the various examples given. The examples represent only some of the species of snakes.

Shield-tailed snakes (*Uropeltidae*) are called that because of the greatly enlarged, modified scale found near the tip of the tail. The function of this structure has not been clarified, but it possibly serves to anchor the animal to the ground. Shield-tailed snakes are thought by some to burrow by pushing the tail tip forward and turning it from side to side. They lack externally visible hind limb vestiges. Shield-tailed snakes are live-bearers that give birth to 3 to 8 young at a time. The diet consists of earthworms and insect larvae. Most species do not exceed 30 centimeters (12 inches) in length, but the *Rhinophis oxyrhynchus* achieve a length of about 60 centimeters (24 inches).

Sunbeam snakes (*Xenopeltidae*) are distributed across India and southeastern Asia. The toothed pre-maxillary bone, rudimentary pelvic girdle, and hind limb vestiges are all primitive characteristics. The sunbeam snake (*Xenopeltis unicolor*) achieves a length of about 1 meter (3.2 feet). The snake has a round body, and its dorsal scales shimmer with a rainbow effect. When a sunbeam snake is excited, it twitches its tail vigorously, with much the same action as in rattlesnakes. Sunbeam snakes lack rattles. Their prey consists of small snakes, frogs, small rodents, and birds.

Wart snakes (*Arochordidae*) are distributed from the coastal regions of India and Sri Lanka across the Indo-Australian islands as far as the Solomons. Unlike most other snakes, the scales of wart snakes do not overlap, but are set adjacent to each other (an arrangement often found in lizards) and bear a sharp ridge. The scale characteristics give the skin a rough, warty appearance. Wart snakes live in the brackish zone of rivers, but sometimes swim short distances in the sea. They are sluggish, slow-swimming snakes that can often stay underwater on the bottom for long periods without breathing. The nasal openings are on the upper side of the snout, enabling them to breathe by simply protruding this part out of the water. Wart snakes only rarely come onto land, where they crawl about clumsily.

Boids (*Boidea*)

These snakes arose toward the end of the Cretaceous period as species closely related to monitorlike lizards. Boids comprise not only the largest snakes alive today, but also small and medium-size species. The top of the head has small scales or large, symmetrical plates, and the back has small scales. The belly is covered with larger plates. The pupils are vertical. There are paired lungs. All three parts of the pelvic vestiges are present, and the skull bones are joined by flexible linkages. Prey is killed by constriction.

There are only four species in the entire boid family that pose a danger for people—the rock python, the Indian python, the reticulate python, and the anaconda. See Fig. 1. Boids are the only *large* animals on earth that are mute. At most, they can utter a hiss. They are also deaf and cannot preceive air vibrations. They can sharply perceive movements in the ground or on the substrate upon which they rest. In recent years, it has been shown that some snakes can perceive loud noises with the sensitive tips of the tongue.

As in all snakes, the eyes of boids lack lids and are covered by the epidermis. The skin is shed along with the rest at molting time, and a new epidermal covering develops. The membrane protects the eye. The eyes of snakes are less mobile and less elastic than avian and mammalian eyes, and therefore their accommodation capability is less. Boids have fairly keen vision, however. They can distinguish outlines well at short distances and can recognize immobile objects in the vicinity and climb about on them. Prey and conspecifics are distinguished primarily by the way they move and from their odor.

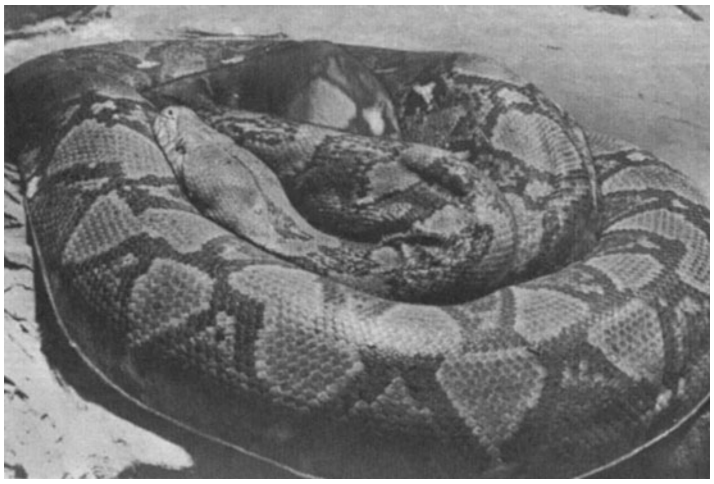


Fig. 1. Regal python. (*A. M. Winchester.*)

During the day, the boid pupils constrict and form vertical slits, while at dusk and at night they widen and look like the eyes of a cat. The firm covering offers effective protection for the eyes when the snake crawls through thickets and holes, and underwater, where they enable the snake to see. The protective membrane gives the impression that the snake is constantly staring. The eyes are only useful for perceiving prey at close range and for the perception of movement; boids have another organ that helps them detect prey nonvisually. Small rectangular openings in the scales on the lower and upper lips (the openings being analogous to the pit organ of pit vipers) enable the snake to perceive even faint warmth radiations. They can detect a human hand at a few dozen centimeters with this heat organ. The organ enables boids to find warm-blooded prey hiding in concealed sites. The finest of all senses, however, is olfaction (smelling), which arises from the Jacobson's organ in the roof of the mouth. The tongue leads small particles in the air to this organ. The heat organ and Jacobson's organ make snakes independent of light conditions, for regardless of bright or dark, they can pursue their prey and often overtake it.

Boids move not only by means of swinging movements in the horizontal plane, but also by stretching the front of the body forward and pulling the rest behind (earthworm movement). A boid on a smooth surface can leave a straight track rather than a winding one. Boids move even better in the water than on land. They swim with surprising agility, diving and surfacing. When swimming, they make use of their ability to pump air inside the body, preventing it from sinking. If the snake wants to dive it releases some of the air. Some boids (not just the purely aquatic South American anaconda) cover considerable distances in water, even undertaking extensive trips into the open sea, sometimes floating along with drifting tree stumps.

One boa constrictor is reported to have covered the distance from the South American mainland to Saint Vincent Island, a distance of 320 kilometers. When the volcano Krakatoa erupted in 1888, all life was destroyed on that Indonesian island. Biologists then recorded the return to the island of various plants and animals in the following decades. Reticulate pythons were one of the first reptiles to reappear there, arriving in 1908.

A great deal has been fancifully written about the strike of a python (for example, Kipling's descriptions in the *Jungle Book*). Snake breeders know that pythons never intentionally strike an object with their head. Nevertheless, they are very strong animals. The entire body, from head to tip of tail, is a single muscle package, which enables boids to gain a powerful hold on their prey. A strong man can probably defend himself against a 4-meter (13-foot) long python or anaconda, but he could not deal alone with one of these weighing 50 kilograms (110 pounds) or more. The boid seizes its prey or an enemy by darting quickly toward the animal and seizing it with its long, sharp teeth. Then the snake wraps around the prey two or more times, at the same time constricting its muscles. Boids do not inflict killing wounds with the teeth. They strangle their prey or kill it by causing vital blood vessels to burst.

Boids must first gain a hold with their teeth if they are to deal effectively with prey, so their teeth are of utmost importance. A reticulate python has about 100 very sharp teeth that point backward. If the snake grabs even a finger, the finger can not be pulled out. One would have to force open the snake's jaw and push the finger further in before it could be released from the teeth. Boids kept in zoos become fairly tame quickly. Even pythons caught in the wild defend themselves by biting and almost never by attempting to wrap themselves around the persons capturing them. In handling newly-arrived boids at zoos, one man must be assigned for each meter (3.2 feet) of length of the snake. Each man must be able to grab the snake firmly and not let go.

Although boids become tame in a short time, snake charmers often literally cool their animal down before doing a performance. The dulled snake lets the charmer do about anything when it is cool and it does not regain its normal activity level until after the display.

It was formerly believed that boids cover their prey with saliva prior to swallowing them, but it is now known that this does not occur until the prey is in the jaws and inside the snake's body. A snake can regurgitate its prey if the snake is startled, but this is uncommon. However, a goat or antelope that is swallowed can cause so much discomfort with its horns that the snake may rid itself of the animal.

Although large snakes can swallow huge amounts at a single time, they actually eat modest amounts and require just a little more than their own weight in food per year. After swallowing large prey, a boid may not eat at all for many weeks and may fast for several months without losing any significant body weight.

Gases forming in the snake's stomach cause the body to become even more inflated several days after eating prey, and the gases sometimes cause the snake considerable discomfort. The swallowing capacity of boids has its limits, and reports about them swallowing horses, cows, and other large animals are considered fantasy.

Large boids reach sexual maturity in three years (in captivity). Males recognize females by the anal secretions of the female. Copulation may last for several hours. While all pythons lay eggs, all boa species are live-bearers. Boa ovoviviparity is to some extent believed to be the result of a more advanced stage of adaptation. Female pythons, which coil around their eggs and may incubate them two or more months, are in greater potential danger than female boas, which can move away right after they give birth. The young boas, which are able to crawl into all sorts of recesses immediately after birth, are also less vulnerable than eggs. The eggs which female boas develop within their body are covered by a membrane instead of a shell.

Tree Boas (*Corallus*). The *emerald tree boa* (*C. caninus*), which attains a length up to 2 meters (6.4 feet) is from northern South America and Brazil and is one of the most beautiful boas of all. Its brilliant green coloration with whitish or yellowish bands offers superb camouflage when it crawls about in trees or shrubs as it hunts birds or lizards, or when simply resting in a tree. The body is compressed laterally, enabling the snake to press close to the tree branches. On the ground it is not a particularly agile snake. In the resting position, the prehensile tail often grasps a branch as the body is wound around the tail, with the coils of the body equally divided on both sides of the point where the tail is grasping. The long, powerful teeth enable tree boas to catch birds, which they snap at and then hold onto with these teeth.

The *garden tree boa* (*Corallus enhydris*) attains a length of up to 2.5 meters (8.2 feet) and is distributed in the Central and South American tropics, as well as the Antilles. The basic coloration may be brown, ochre, or gray, and sometimes it has a striking pattern. The garden tree boa uses an accordionlike climbing technique. The snake wraps the front of its body around a thin part of the tree trunk and then pulls itself higher, only to reanchor the body again and climb a bit more.

Boa Constrictor (*Boa constrictor*). This snake usually will reach a maximum length of 4 to 4.5 meters (13.1 to 14.8 feet) and will weigh up to 60 kilograms (132 pounds). This snake is probably the most familiar of the boids, although by no means the largest. It is distributed from Mexico to northern Argentina. Three closely related species occur on the West Indies islands and the plains of western Argentina. Boa constrictors, whose pattern varies from one individual to the next, are among the most beautiful of snakes. They are ground dwellers and are particularly prevalent in mountain forests. These snakes live chiefly on small mammals, birds, tegus, and iguanas.

Anacondas (*Eunectes*). These boas are more aquatic than those just described. The *anaconda* (*E. murinus*) can attain a length of 9 to 9.6 meters (29.5 to 31.5 feet) and the large ones are the longest of reptiles in the world today. The *yellow anaconda* (*E. notaeus*) is considerably smaller.

Contrary to popular preceptions, most boids (excepting sand boas and a few others) cannot tolerate extremely severe heat. They prefer temperatures of 20–30°C. Many species hibernate or estivate and can survive unfavorable seasons in a dormant state.

Colubrid Snakes

The colubrid family (*Colubridae*) contains more species than any other snake family. The relationship between the presence or absence of grooves and the individual species is evident when one inspects the venom gland, a structure that is necessarily found in every venomous snake. Venom glands probably developed through a process of adaptation from upper lip and salivary glands. The venom glands consist of the upper lip gland (which only produces mucus) and a poison-transporting venom gland. As early as 1894, investigators demonstrated that the secretions of a single venom gland from a ringed snake is potent enough to kill a guinea pig. The mucous gland lubricates the prey; the venom gland helps prepare it for digestion.

The number of species of colubrid snakes is extensive, as is obvious from the table, and only a few of the more interesting species can be highlighted here.

Ringed Snake (*Natrix natrix*). These snakes reach a length of about 1 meter (3.2 feet) on the average, but some species are longer. Ringed snakes are distributed across a large area, including most of Europe and a bit of northern Africa. The snake is named for the two fairly distinct orange, yellow, or whitish moonshaped spots on the rear of the head, which look something like a ring. It is difficult today to find a ringed snake more than 2 meters (6.5 feet) long. The longer ones are usually more cautious and therefore more difficult to catch. Due to their large size, they are forced to traverse a greater area and to visit many ponds in order to find sufficient nutrition. Females wander about seeking favorable egg-laying sites (often in dung heaps). Their enemies include cats, hedgehogs, raptorial birds, and humans. Many are killed on modern highways.

When ringed snakes are threatened, they usually respond with a series of stereotyped defensive behavior patterns—flicking the tongue and hissing, releasing the content of their stink gland; defecating or regurgitating food. Only rarely do they actually bite, and a ringed snake bite has no harmful effects for humans.

Diced Snake (*Natrix tessellata*). These snakes are found in central Europe and the Near East. They can achieve a length of between 75 and 100 centimeters (29.5 and 39 inches), but eastern species may be as long as 150 centimeters (59 inches). The diced snake is rather slender with a narrow head. The eyes are medium in size, are situated slightly high, and have round pupils. The top of the neck often has a dark V-shaped spot. The underside is whitish or yellowish, or it may have reddish and black spots. Some authorities believe that the diced snake is a relic from the warmer postglacial period. It is even more dependent on water than the ringed snake. It consumes a larger amount of fishes and often lies for hours or days underwater near the shore on low rocks, under which it flees when people approach. The eyes of the diced snake are well adapted to a highly aquatic life. While ringed snakes can be found considerable distances from water, the diced snake occurs in the immediate vicinity of water.

Viperine Snake (*Natrix maura*). This species achieves a length of over 80 centimeters (31.5 inches). Distribution is similar to that of the ringed and diced snakes. The viperine name arises from the zigzag band extending along its back, giving it a similar appearance to the common viper. The snake is harmless. Viperine snakes can be distinguished from vipers by the round pupils of the viperines. The stripe on the back of this snake sometimes appears as two rows of spots, but usually forms the aforementioned zigzag pattern. Dark spots or rings run along both sides of the central stripe, and they enclose a white or yellowish spot. Viperine snakes feed on fishes and amphibians.

Water Snakes (*Natrix spp.*). There are many water snakes in North America, mainly in the United States. The *diamond-back water snake* (*N. rhombifera*) is about 1.4 meters (4.6 feet) long when fully grown. The back is brown or olive-brown, with a chain of rhomboid figures. The belly is yellow and has halfmoon-shaped brown spots. Distribution is from the midwestern United States southward into Mexico. The *plain-bellied water snake* (*N. erythrogaster*) also reaches a length of about 1.4 meters (4.6 feet). The upper side is uniform black to redbrown and the underside is reddish. This snake occurs in the southeastern United States. The *queen snake* (*Regina septemvittata*) attains a length of about 50 centimeters (19.6 inches) and is similar to the plain-bellied water snake except it is more slender. This species occurs from Pennsylvania to Wisconsin. The *brown water snake* (*Natrix taxispilota*) on the average has a length up to 1.3 to 1.5 meters (4.3 to 4.9 feet) and is the largest North American water snake. It is heavy and plump and has a clearly distinguishable head. It is brown to rust-brown, with large rectangular spots. This species occurs in the southern United States. The *Brazos water snake* (*N. harteri*), reaching a length of about 90 centimeters (35.4 inches) and is known only in the Palo Pinto region on the upper Brazos River in Texas. *Kirtland's water snake* (*Clonophis kirtlandii*) reaches a length of about 45 centimeters (17.7 inches) and is a light-brown to gray snake with roundish black spots. Distribution is from the midwestern United States to Pennsylvania and New Jersey. *Graham's water snake* (*Regina grahami*) attains a length up to 60 centimeters (23.6 inches). It is dark brown with a broad yellow band on the first three rows of scales, bordered below by a narrow dark stripe. A pale stripe lined on both sides with black extends across the back, and the belly is yellowish with dark splotches. This snake occurs from Illinois to Louisiana and eastern Texas.

Water snakes generally live on fishes, anurans, and urodele amphibitans, crustaceans, and insects. Mating occurs in April and the young are born live between August and early October. Newborn common water snakes are about 22 centimeters (8.7 inches) long. With their plump body, dependence on water, and vigorous defensive behavior, the common water snake may be confused with water moccasins.

Garter Snake (*Thamnophis*). These snakes achieve a length of 50 to 150 centimeters (19.7 to 59 inches) and are the most prevalent colubrid snakes in North America. They are small, slender, and often considered "cute." There are several species. These ground-dwelling colubrid snakes are found throughout the 48 contiguous United States, as well as in southern Canada and northern Mexico. They are sometimes observed in urban areas. They differ from water snakes in having an undivided anal scute, and many species have three light stripes. All garter snakes are live-bearers. Some feed chiefly on earthworms, while others living near water concentrate on frogs.

Smooth Snakes (*Coronella*). These snakes have a small head which is distinct from the neck. The medium-size eyes have round pupils. The scales are smooth and are arranged in rows. The group includes small to medium-size species of ground-dwelling, nonvenomous colubrid snakes distributed in Europe and Asia. Smooth snakes occur mostly in woods, inhabiting clearings, paths, forests, and particularly piles of wood and rocks and shrub-covered ground, where they can sun themselves. That same habitat supports sand lizards, common lizards, and blind snakes and mice on which they feed. Like king snakes, smooth snakes also feed on other snakes.

Common King Snake (*Lampropeltis getulus*). These snakes attain a length up to 2 meters (6.6 feet). They prey upon other snakes, even poisonous ones. Their diet also includes small mammals, lizards, and fishes. When a king snake seizes another snake, it wraps its body firmly around the other animal and chokes it. Interestingly, rattlesnakes do not assume their typical position when they are brought face to face with a king snake. Instead, they press the front of the body against the ground and try to hit the king snake with another part of the body. The king snake has a light chain pattern on the back. Distribution is in the southern United States and northern Mexico.

Dwarf Snakes (*subfamily Calamarinae*). These small snakes are only from 25 to 30 centimeters (9.9 to 11.8 inches) long. The head is small and passes into the round, firm trunk without any noticeable neck section. There are some seven genera with approximately eight species, all from the southeast Asian tropics, chiefly Indonesia and the Philippines. They feed primarily on earthworms and insects. They spend all their lives on the ground, concealing themselves in the soft earth beneath fallen tree stumps and rocks. Defenseless and slow, they are often eaten by other snakes.

Flying Snakes (*Chrysopela*). These snakes do not fly, but they drop from one branch to a lower one so quickly that it appears as though they are gliding. They attain a length up to 1.5 meters (4.9 feet). They have a longitudinal ridge on each side of the belly. The *C. ornata* has brilliant coloration with a complex pattern. These snakes are found in Indonesia and southeastern Asia.

Cobras and Sea Snakes

The cobra family contains some of the most greatly feared snakes, about which so much has been written: cobras, which perform before snake charmers; the legendary black mamba; and the colorful, red-yellow-and-black ringed American coral snakes. Two members of the family are usually thought of as being among the most venomous snakes in the world—the southeast Asian king cobra and the Australian taipan. Cobra species occur in every continent except Europe. The family is particularly diverse in Australia, with about 75% of all Australian snakes being in the cobra family.

There are four dangerous venomous snake families known: cobras, sea snakes, vipers, and pit vipers. These four families form two major groups. The fangs are always in front of the mouth, but in cobras and sea snakes, they have a groove, which indicates how the originally open venom canal developed over many centuries of adaptation.

The *Elapidae*, like other venomous snakes, arose from nonvenomous snakes. Their anatomy is much more like that of the colubrids than is that of the plump, short-tailed vipers. Also, in cobras the fangs are firm, relatively short structures, while in vipers the short, vertical upper jawbone and fangs are rotated back against the palate when the mouth is closed. Thirdly, cobra fangs have a closed groove, but it has a deep indentation. Finally, cobras as a rule produce neurotoxins in their venom, while viper venom generally has a hematoxic (blood poisoning) action.

Elapids are generally slender, highly agile snakes with a colubridlike head that is not very distinct from the neck and which bears large, colubridlike scutes. The fangs are short, so the mouth does not have to open very wide for them to be useful. There are no other upper teeth in the highly developed species (mambas and coral snakes), while others have one or more small, ungrooved teeth behind the fangs in the upper jaw. Additional teeth are found on the palate, the pterygoid bone, and the lower jaw.

The length of these snakes varies from 30 centimeters (11.8 inches) in the bandy-bandy to over 5 meters (16.4 feet) in the king cobra. The body often has stripes that may be very colorful. Many cobras can flatten their body when they are excited, and *Naja* cobras can spread their neck ribs, forming the familiar hood. Cobras and sea snakes share a similar fang anatomy, but they differ in that cobras have a round tail. Their chief prey are small vertebrates. Several species prefer to feed on other snakes. They are ovoviviparous (live bearing, particularly in the Australian species) or oviparous (egglayers). Elapids do not constrict their prey.

The *Elapidae* contains more venomous snakes than does any other snake family. Most of them are ground-dwellers, often secreting themselves in rodent dens or burrowing in loose earth (coral snakes). A few are arboreal (mambas, cobras, and Gold's cobra), and one genus (winged water cobra) inhabits standing and slowly flowing water. Most elapid snakes are active at dusk and night and they avoid bright sunlight. There are 41 genera with approximately 180 species and 300 subspecies.

King Cobra (*Ophiophagus hannah*). This is the largest poisonous snake in the world. Specimens as long as 4 meters (13.1 feet) are not unusual, and the current record is 5.58 meters (18.3 feet). Whether the species is also the most dangerous poisonous snake is still being debated. The very large venom glands do produce a huge quantity of venom, and fatalities from king cobra bites have been recorded in Burma, India, and China. However, one should regard phrases, such as "the most dangerous venomous snake" with some skepticism. The effect of any individual bite on a human will vary with the conditions under which the incident occurs. The metabolic reactions of the person involved (death from shock has occurred, but not in every case) and the degree to which the snake is excited (usually exaggerated in reports) both play a role in the effect of a bite. It is not correct to measure potency of a snake's venom by using the number of people it has bitten. Actually, the most toxic snakes bite the least often (H. G. Petzold, 1972). Experimental studies have revealed that coral snake venom is much stronger than pit viper venom, but many more deaths are caused by pit vipers than by coral snakes because of the life habits and the prevalence of the species. See Fig. 2.

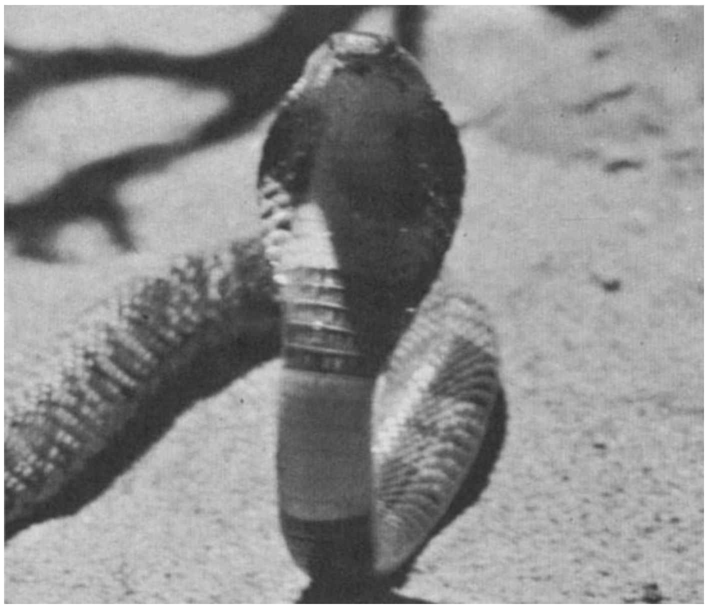


Fig. 2. Cobra. (*A. M. Winchester.*)

King cobras are much less aggressive than many other, smaller snakes, although one might not guess this when seeing a king cobra's hooded threat display. This display usually has a powerful effect on humans. The king cobra inhabits India, but not the nearby island of Sri Lanka. Distribution extends into southern China as far as Shanghai, Malaysia, the Sunda Islands to Bali, the Andaman Islands, and the Philippines. The head of the king cobra can grow to be near the size of a human head and coloration varies with habitat from olive-brown or gray to a deep, shining black. Forty to fifty narrow, irregular light stripes extend across the back. The eyes have a bronze hue. The two fangs are never more than 10 centimeters (4 inches) long, and there are usually three smaller upper teeth behind each of them.

The snake generally leads a secretive life. However, it feeds by day as well as by night. It prefers dense highland jungle, often near water, and in the Nilgiri Mountains of India it occurs as high as 2000 meters (6562 feet). King cobras often flee into water when they are pursued. Unlike most other venomous snakes, the king cobra is very agile and excitable. Its threat posture is almost universally known; the front of the body is raised and the loose neck skin is stretched wide as the movable neck ribs spread out. The king cobra's hood is not the disklike rounded form of the smaller Asian cobra (also known as the Indian cobra), and the king cobra lacks the eyeglass markings. An upright king cobra can sway its forebody for several minutes, but it never does so in the swanlike manner of *Naja* cobras. However, it is the only cobra species that can move forward while it is in the threat posture. Hardly anyone would easily dismiss an encounter with a threatening king cobra, which not only assumes that characteristic posture, but also emits a high, penetrating hiss through the nasal openings.

King cobras are especially feared during the mating season, and in India, paths and streets may be closed off by officials if a cobra nest is discovered nearby. This leads to a fascinating aspect of king cobras. The eggs are incubated. The characteristic "two-story" nest, made of foliage, was first described early in the 19th century. The first "story" contains the 18 to 40 white eggs; the female coils her body around them, while the male is usually stationed near the nest. Until recent years, it was believed that the snakes randomly gathered twigs that were blown about by the wind. Definitive observations were made at the Bronx Zoo (New York) of a breeding pair of king cobras. The 4-meter (13.1-foot) female grabbed bamboo shoots, branches, and foliage by slinging her forebody around them, and in two days, she had constructed a nest nearly 1 meter (3.2 feet) wide. In their Indian habitat, newborn king cobras are about 0.5 meter (1.5 feet) long and are deep black with yellowish-white stripes. Later the coloration becomes lighter and the contrast between the various hues diminishes.

King cobra venom is extremely potent. A human can die within 15 minutes after being bitten unless treated with the proper antivenin. The symptoms for king cobra poisoning are typical for the neurotoxic action of elapid venom. Victims suffer from vision disabilities and severe dizziness. This is followed by a slowing down of the reflexes and respiratory difficulties. Work elephants are sometimes killed by king cobras in India and Thailand if the snake manages to bite a tender spot, such as the tip of the trunk or the area where the toenail meets the skin. An elephant bitten by a king cobra usually will die within 3 to 4 hours.

Not every king cobra strikes at humans without provocation. In terms of the "art" of snake charming, few king cobras are used in such performances, and those that are used are generally limited to religious performances, as in Burma. The Egyptian cobra and the Asian or Indian cobra are the more typical performing snakes. Snake charming is an ancient art; the snake that Aaron charmed before the pharaoh was an Egyptian cobra. Although many charmers work with artificially cooled-down snakes or fangless cobras, masters of the art use normal cobras. Snakes are deaf and do not dance in their baskets to the tune of their charmer's flute. With their eyes and forebody, they follow the movements of the tip of the flute and the swaying body of their master, who merely needs to move with the rhythm of his music to make it appear that the snake is dancing. The charmer knows the striking distance of the cobra as well as any professional herpetologist and adjusts himself to the snake accordingly.

True Cobras (*Naja*). These snakes are less surrounded by myths, but are more prevalent. The *Asian* or *Indian cobra* (*Naja naja*) is the most prevalent and occurs over a large region from central Asia through India and southern China to the Sunda Islands and the Philippines. Coloration varies more within this species than in any other; there are light-brown, brown, olive-gray, and completely black common cobras, which may have lighter stripes. Asian cobras generally reach a length of 1.5 meters (4.9 feet), with the largest ones being found on Sri Lanka.

They inhabit a number of places—from the thickest jungle to open rice paddies, city parks and gardens, sheds, bazaars, and village streets. The proportion of the body that is raised when a cobra threatens depends upon the degree that the animal is excited and varies from one-fifth to one-third of the total body length. Raising of the hood is a sign of threat and defense and, together with raising the body, shows that the snake is prepared to attack. A threatening cobra does not bite upward; its strike is not directed higher than the mouth is at the time the snake is in the upright position.

Cobras are visually oriented animals; they show little response to mechanical manipulation.

The first attempts to develop immunology and serum therapy for snake bite occurred as early as 1891. Calmette was the first to collect cobra venom for injection into small animals. The first snake serum was available in 1895. When cobras are "milked," an expert holds the snake at the edge of a Petri dish and lightly massages its venom glands with thumb and middle finger. This releases about half the venom stored in the glands; only about one-fifth of the contents are released when a cobra bites. Cobra venom is extremely potent. As little as 0.00002 gram of dehydrated cobra venom can kill a guinea pig. Theoretically, one gram of the dehydrated cobra venom is sufficiently potent to kill 140 dogs, 167,000 mice, or 165 humans. The mongoose is reputed to be immune to snake bite, but this is not true. However, a mongoose can tolerate a dose of cobra venom that is eight times the quantity that would kill a rabbit.

Mambas (*Dendroaspis*). This snake is more feared and ill-famed in Africa than the resident cobras. The most familiar is the *black mamba* (*D. polylepsis*), which attains a length of about 4 meters (13.1 feet) and is the largest venomous snake in Africa. It is characterized by lightning-quick, elegant movements in branches and the vigor of its reactions. These large, slender snakes have a narrow head, smooth scales arranged in diagonal rows, very large eyes with round pupils, and enlarged lower teeth, which together with the poison fangs of the upper jaw grab a firm hold onto prey, usually birds, arboreal lizards, and tree frogs.

Contrary to their reputation, mambas do not make wild attacks on harmless travelers. The snakes are shy and will flee into the foliage when people approach. Some mambas are highly territorial and will stay in the same, small area for months.

Banded Krait (*Bungarus fasciatus*). This is a lacquer-black snake with shiny yellow bands on a strangely ridged back, a coiled snake that lies in the shade. The slender head is hardly distinctly different from the rest of the body. The krait achieves a length of 1.5 to 2 meters (4.9 to 6.6 feet). Kraits are nocturnal and sensitive to light. They are specialists and eat almost nothing but other snakes, and they readily attack the large, powerful rat snakes, cobras, and even smaller conspecifics. The poison fangs are only 2 to 3 millimeters long and the amount of venom released (30–40 milligrams, based on measurements of dehydrated venom taken at East Berlin's zoo) is not great. However, the venom contains extremely potent neurotoxic material that paralyzes the respiratory center and can kill a human being in 30 minutes. However, it is extremely unusual for a human being to be bitten by a krait. Kraits are egg-layers.

Coral Snakes (*Micrurus, Micruroides, and Leptomicrurus*). These snakes usually range between 60 and 80 centimeters (23.6 and 31.5 inches) in length, seldom up to 1 meter (3.3 feet). They are characterized by their bright coloration and pattern. In the majority of American coral snakes (*Micrurus*), the only teeth in the upper jaw are the two small poison fangs, an indication that this genus is a phylogenetically recent one. Little venom is released with each bite; the maximum for the Brazilian species (*Micrurus corallinus*) is 50–200 milligrams of liquid (i.e., 60 milligrams dehydrated). Most coral snakes are in Mexico, which has about 30 species. Colombia has 28 and Central America, excluding Mexico, contains 26 species. The number of species is reduced as one moves further north or south; there are just three coral snake species in the United States; Argentina contains only four.

Coral snakes feed chiefly on lizards and small snakes and, less often, on young birds, frogs, and insects. When they bite, they "chew" several times in succession, a behavioral pattern that is rare in venomous snakes. Repeated biting increases the amount of venon injected. The swallowing of prey is a slow process, because the jaws are not very flexible. Coral snakes do not have a strong tendency to bite, and cases of bitten humans are unusual, but when they occur they are serious injuries.

The *eastern coral snake* (*Micrurus fulvius*) achieves a length up to 60 centimeters (23.6 inches) and occurs from southeastern United States to Mexico and is one of the most prevalent species. It often inhabits rodent dens, where its bright coloration makes it stand out. The snout is black. This species assumes a curious defensive position when it is irritated. The head is hidden; the short tail is raised several centimeters and moved back and forth.

The *Arizona coral snakes* (*Micruroides euryxanthus*) also attain a length of about 60 centimeters (23.6 inches). These snakes inhabit arid terrain. Arizona coral snakes have a small, smooth tooth behind each poison fang. This species has red, yellow, and black bands. It is interesting to note that nonpoisonous colubrid snakes in North, Central, and South America bear the same red, yellow, and black ring markings as the poisonous coral snakes. However, the distinction between the harmless and dangerous snakes is rather easy to make in the case of North American snakes. In the coral (dangerous) snakes, a yellow (or white) ring is always bordered by a red one, while in the harmless colubrid snakes, such as the brilliantly colored milk snake (*Lampropeltis doliata syspila*), red and yellow are separated by a black ring. The problem of identification is more difficult in South America because of the presence of "true" and "false" coral snakes which have nearly identical markings.

In Australia, where elapids are particularly prevalent and diverse, there are numerous small, primitive, slightly potent species, as well as some that are among the most dangerous members in the entire family, whose bite can be fatal to humans. They include the *taipan* (*Oxyuranus scutulatus*), which attains a length of 3 to 4 meters (9.8 to 13.1 feet). This is the most poisonous of Australian snakes. It inhabits remote coastal stretches of northern Australia and New Guinea, as well as the islands of the Torres Strait. The snake is brown to brown-black, and melanistic individuals have been found in New Guinea. About 80% of taipan bites are fatal unless antivenin is administered in time. Even with serum injections, many people die. It has been reported that a riding horse in northern Queensland died within five minutes after being bitten by a taipan. The species has long poison fangs. The venom not only paralyzes the central nervous system, but also destroys red blood corpuscles.

When preying (chiefly on rats), the taipan makes several bites in rapid succession. The snake often remains in rat holes and is active early in the morning and late in the evening.

The gray-brown, dark-banded *death adder* (*Acanthopis antarcticus*) is a large Australian elapid, which attains a length of about 80 centimeters (31 inches). It is one of the more prevalent elapids in Australia and New Guinea. The appearance is much like that of a viper. It has keeled scales and, unlike nearly every other cobra species, has a broad, clearly distinct, nearly triangular head (usually a characteristic exclusive to vipers). When excited, the death adder can inflate itself into a sausage shape. Its venom (up to 236 milligrams of liquid per milking) is more potent than cobra venom. Although the bite of a death adder leaves practically no mark, about half of all untreated bites in humans result in death due to respiratory arrest. Death adders are chiefly active at dusk. Australian farmers intensely dislike the species because their sheep are in areas where the snake is most prevalent (central and western Australia) and a sheep that steps on a death adder will probably die because the snake reacts by biting the animal. Female death adders bear from 10 to 12 live young.

Sea Snakes (*Hydrophiidae*). These snakes have the same basic characteristics as the *Elapidae*. However, they have some unique features that reflect their marine existence; the tail is laterally compressed and forms a rudderlike structure. The trunk, particularly the rear portion, also is often laterally compressed. The scale covering is nearly uniform. These highly specialized marine inhabitants have lost the broad ventral scutes seen in terrestrial snakes, so the scales are nearly the same all over the body. The external nasal openings are on the top of the head and can be closed by a valve. Sea snakes have salt glands in the head for eliminating excess salt.

Most sea snakes belong to the subfamily *Hydrophiinae*, which contains 13 genera. Members of this subfamily share the paddle-shaped tail with all other sea snakes. These snakes have lost all dependence upon terrestrial life, and never go on land. Body musculature is degenerate. The strong current in the water takes over much of the work that the muscles would otherwise have to provide. Heavy specimens are rather helpless on land and can suffocate since their vestigal supporting muscles cannot even fill the lungs with air. One of the biological puzzles is the striking similarity in body shape, coloration, and patterning between several moraylike fishes and various sea snakes of the Indian and Pacific Oceans.

Vipers

Two snake families, whose venom apparatus has reached an impressive degree of effectiveness, are at the peak of phylogenetic development of the large suborder of snakes. Their poison fangs have no sign of grooves; they are solenoglyph teeth, which means that they actually have enclosed canals within the fangs that transmit the venom out of the body, very much like a hypodermic needle. Furthermore, members of these two families, excepting just a few species, have lost the large head scutes found in colubrid snakes and cobras and instead have many small scales on the head. These families are the vipers or adders (*Viperidae*) and the pit vipers (*Crotalidae*). Pit vipers are considered to be a distinct family because of their unique pit organs.

Vipers are found only in Europe, Asia, and Africa. Australia lacks both vipers and pit vipers, an indication that both families are phylogenetically recent, developed after the Australian continent became a separate land mass. Whether vipers developed from cobralike ancestors or directly from nonvenomous colubrid snakes remains obscure. It has been well established that vipers arose somewhere in western Asia.

The most important distinguishing characteristic of the venom apparatus in vipers is that the two upper jaws, which bear the fangs, are greatly shortened. Each upper jaw has a special joint that permits the jaw, along with the fang anchored firmly within it, to rotate 90 degrees. If the viper closes its mouth, the fangs lie back, tips inward, and are covered by a fold in the mucous membrane. When the mouth is opened, a lifting mechanism is activated, putting the fangs into a vertical position. The fangs (or more precisely, the upper jaw bones) are laid back with the same action as when a pocket knife is snapped together. The adaptation of folding back the fangs permits them to be extremely long, far exceeding the length of those in cobras. The fangs of the giant king cobra are not much longer than those of the rather small adder. The long fangs enable vipers to bite deeply into the tissues and cause the victim to suffer severe necrosis. The fangs snap back into the mouth after they are withdrawn from the victim.

Viper venom contains mainly hematoxic materials (substances that injure the blood and associated vessels). Thus, a viper bite has a very different effect from a cobra or mamba bite (their venoms being neurotoxic). Viper bites are accompanied by prominent local irritation and symptoms of severe blood poisoning, with burning pain, inflammation, pronounced discoloration, sudden drop in blood pressure, internal bleeding, degeneration of the tissues, and the formation of an abscess. Death ensues because the heart stops, not as the result of respiratory arrest as in the case of cobra bites.

Vipers or adders are generally compact, sturdy snakes. The head is triangular and is distinct from the rest of the body. Their length ranges from 30 centimeters (11.8 inches), as in the case of the dwarf puff adder, to about 180 centimeters (5.8 feet), as in the Gaboon viper. The pupils are usually vertical and elliptical. All face bones are movable. Each of the two shortened, retractile upper jaw bones bears only the tubular venom fang (which can only be activated for a short period) and often one to several significantly smaller reserve teeth of various sizes, none being a firmly positioned poison fang. The tail is short, and coloration is usually drab. In the genus *Vipera*, it is often of a dark zigzag pattern or a rhomboid band along the back. Desert species are sand-yellow, while jungle vipers often have a colorful carpet marking. There are 10 genera with 60 species and 110 subspecies.

Some vipers inflate their bodies into sausage shapes when excited. Almost all vipers assume a plate-shaped coiled position as a threat gesture, in which they lift up the neck and hold it in an S-shape. Other threat behaviors include loud hissing and rapid, forward jerks of the head. Some sand dwellers, such as the saw-scaled viper, create a particularly impressive sound by rubbing their scales together. Vipers feed chiefly on small vertebrates, particularly rats, mice, and lizards and less often on frogs and birds, paralyzing or killing their prey by biting it. Some of the smallest vipers prefer locusts. Many vipers are useful for controlling rodent pests.

The true vipers (*Vipera*) are distributed in Europe and Asia, with 8 species and, in Africa, with 4 species. The *adder* (*Vipera berus*) attains a length of 50 to 60 centimeters (19.7 to 23.6 inches), seldom over 80 centimeters (31.5 inches). The adder can tolerate the coldest climates of any viper species. Its distribution essentially follows that of deciduous forests, mixed forests, northern pine forests, and high-altitude forests. In Europe, the species occurs as far north as the Arctic Circle and also inhabits the Carpathian, Balkan, and Caucasus Mountains. The *asp viper* (*V. aspis*), about the same size as the adder, occurs chiefly in the Mediterranean region, extending northward as far as the southernmost part of the Black Forest of Germany and moving from there into central France and western and southern Switzerland.

Adders do not have the compact appearance of many other vipers. The head is only slightly broadened toward the rear and it is covered with rather large scales. Basic coloration in males is gray; in females, brown. Adders hibernate underground at a site which is moist, but protected from flooding. Live-bearing (ovoviviparity) permits the adder to penetrate far north and high up on mountains. The number of young varies between 6 and 20. Two vipers are seldom found together outside the mating season, although their home territories overlap and they may live within several meters of each other. Vipers hunt by lying in wait and grabbing appropriate prey that passes within striking distance. The chief enemies of European vipers are serpent eagles and the hedgehog.

Attempts to find snake antivenins were being made in antiquity, but truly effective venom antidotes were not developed until the end of the 19th century with the work of Calmette, as previously mentioned in connection with cobra antivenom, and the work of Phisalix and Bertrand, who produced asp viper antivenin at about the same time. Asp viper antivenin became available in 1896. The sand viper (*Vipera ammodytes*), which attains a length up to 90 centimeters (35.4 inches), is found in southeastern Europe and is the most dangerous (venomous) snake in Europe. Sand viper bites cause fatalities in remote parts of the Balkans. However, its venom does not compare with the potency of that in cobras or rattlesnakes.

The Gaboon viper, from the jungles of western, central, and eastern Africa is the largest of all vipers. It is the most prevalent of venomous snake species in the mountains of Cameroon. Its patterning is like that seen on ornate Oriental carpets; the deep purple-brown background has geometrical patterns in yellow, light brown, and blue, usually in an hourglass arrangement along the sides with long, rectangular, deep-purple bordered rings. The pale-brown head with its dark medial line has a chocolate-colored wedge-shaped spot whose tip extends to the eyes. With most exceptions, Gaboon vipers are docile and do not readily bite.

Pit Vipers

Pit vipers and vipers share a common extinct ancestor. Pit vipers probably arose once the egg-laying vipers had evolved and were developing along that line. Pit vipers survived because their new organ, the pit organ, was biologically successful, and this recent snake group is still evolving. More than two-thirds of all pit vipers inhabit the Americas, but their origin is probably in tropical Asia, where vipers and pit vipers occur together. There are no pit vipers in Africa or Australia, but the species has reached the southeastern edge of Europe, on the Caspian Sea.

In pit vipers, the pupils are vertical and elliptical. The fangs are long and solenoglyph (tubular). They often remain inside the prey, and may be excreted in the pit viper's feces. Fangs are replaced two to four times annually by reserve teeth. Contractions of the muscles surrounding the venom glands control the amount of venom released and direct the venom via a mucous-membrane covering to the teeth. Their length varies from 40 centimeters (15.7 inches) to 3.75 meters (12.3 feet). Coloration is typically gray, brown, or olive with a distinctive light-dark diamond pattern or dark spots. There is often a dark band extending from each eye to the corner of the mouth. Arboreal pit vipers (with prehensile tails) are green or yellowish. Behaviorally, pit vipers resemble vipers, but the threat posture consists of slightly flattening the body and rolling it together into a spiral, and then lifting the forebody from the ground and bending it into an S-shape. The end of the tail is also raised, and tail shaking occurs even in those pit vipers lacking rattles. The rattling is a warning reaction. The diet is the same as that of the vipers.

Venom in most pit vipers contains hematoxins like that of vipers. The tropical rattlesnake is an exception. Since pit vipers play a more significant role in North and South America than vipers do in Europe, organized medical aid for bitten people is more extensive in the Americas. The Instituto Butantan, near São Paulo, Brazil, was the pioneer establishment for dealing with the treatment of venomous snake bites. In 1901, Mineiro, who had gained fame by combating typhoid and pestilence epidemics in Brazilian ports, was assigned the task of establishing a serum therapeutic institute for the State of São Paulo.

All pit vipers have a deep pit between the nostril and the eye; its opening is larger than the nostril but smaller than the eye. The pit, which has been depicted in old Indian drawings, lies in an indentation of the upper jaw bone. Scientists puzzled for decades over the meaning of these pits. As early as 1824, it was known that tiny nerves lead to them and that they were therefore sensory organs, but there were arguments of just what sense modality they served. It was not until the late 1930s that most of the mysteries were cleared up. It is now established that the pit organs are heat-detecting sensors. They do not function like a thermometer, as once thought, but respond to temperature differences between the inside of the snake and the outside. Sensitivity has been suggested as being a small fraction of one degree Celsius. With this organ, the pit viper can sense the presence of warm-blooded prey at a distance up to 50 centimeters (19.6 inches). The pit organ is particularly helpful at night when the difference between the temperature of the prey and the ambient temperature is the greatest.

Rattlesnakes (*Crotalus*). These snakes are among the best-researched and most impressive snakes known. All rattlesnakes have the rattle at the tip of the tail. The biological function of the rattle was discussed for many years. The hard, dry, chainlike, loosely-jointed rattle elements at the tail tip are the remains of previous moltings. See Fig. 3. In newborn rattlesnakes, the tail ends in a spherical scale. While all other snakes cast off this terminal scale at every molt, rattlesnakes do so only at their first molt, which occurs 7 to 10 days after birth. This last scale hardens into a pistonlike hollow structure before the onset of the second molting; it has a ringlike construction. During subsequent molts, this last scale loosens like all the others, but it is not cast off. In the meantime, a new terminal scale has developed from inside this first one, and the new one has a greater diameter and holds the first one. Another terminal scale develops at the next molt, and so it goes, adding more and more rattles to the chain. Each of the hollow, podlike scales in the chain of rattles was, at one time, the closest to the snake's body. Thus, contrary to widespread popular concepts, the number of rattles does not indicate the rattlesnake's age in years. They show how many times the snake has molted and rattlesnakes molt at an average of three times each year. Rattles tend to break off and with an irregular molting schedule, there is no direct connection between number of rattles and a snake's age. Rattlers can shake their tails much faster than a human eye can perceive—up to forty or 60 cycles per second. The sound is more like a hissing buzz or whir than a rattle. The rattlesnake does not hear its own rattling because it is deaf. The most likely meaning of rattling is that it is a warning sound used to intimidate enemies—a tool for upsetting the enemy. One exception to the foregoing is the Santa Catalina rattlesnake (*Crotalus catalinensis*), which does not have rattles.

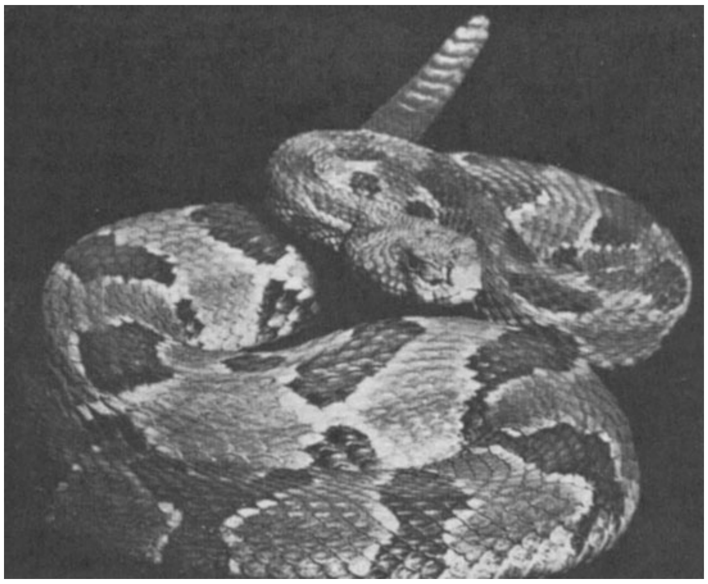


Fig. 3. Rattlesnake. (*A. M. Winchester.*)

Most species of rattlesnakes prefer dry, rocky, shrub-covered terrain, where they can conceal themselves inside crevices in the rocks or in mouse holes. Forest dwellers are more active during bright daylight hours than are species which inhabit more open country. Northern species often hibernate in large groups. A single mass winter quarter can support rattlesnakes from a large region and usually the same site will be used year after year. Northern populations of prairie rattlesnakes have been known to form groups numbering 200 to 300 on the average, but sometimes up to a thousand snakes. Rattlesnakes often share their quarters with nonvenomous snakes. Symbiotic relationships also occur among rattlesnakes, burrowing owls, and prairie dogs.

The diet of American rattlesnakes consists of mammals, primarily white-footed mice, pocket mice, voles, wood rats, and chipmunks. Large rattlers also may feed on ground squirrels and wild rabbits. Birds, such as domestic chicks and quail, are eaten less frequently.

Mating occurs in the spring in the southern United States. All rattlesnakes bear live young, born between August and October. Further north, especially in Canada, litters are born every other year in June and July. Most species bear 8 to 15 young per clutch.

Eastern Diamond-Back Rattlesnake (*Crotalus adamanteus*). This is the largest of all rattlesnakes and achieves a length up to 2.5 meters (8.2 feet) and a weight of 10 kilograms (22 pounds). This massive snake with its light-edged diamond figures is one of the most dangerous venomous snakes, chiefly because of the large amount of venom it produces. In a single 'milking,' this snake can yield 1050 milligrams of liquid venom (equal to 300–500 milligrams of dehydrated venom).

Western Diamond-Back Rattlesnake (*Crotalus atrox*). This snake attains a length of 2.2 meters (7.2 feet), but usually is somewhat shorter. The snake is aggressive and easily excitable and causes the majority of fatalities from snake bite in the United States. The species is useful in controlling rodents. The species is highly fertile, bearing 10 to 20 young at a time.

Prairie Rattlesnake (*Crotalus viridis*). This is a greenish-olive snake with longitudinal rows of round, brown spots. It is one of the most widely distributed of the rattlesnakes and exists in many subspecies. It is found in the Sierra Nevada mountains in California up to 4,000 meters (13,124 feet) and extends as far north as Canada. These snakes live primarily on mice.

Sidewinder (*Crotalus cerastes*). This small rattlesnake achieves a length of 60 to 70 centimeters (23.6 to 27.6 inches). Its sidewinding movement is an adaptation to desert life. Sidewinders can move at speeds of 3–4 kilometers (1.8–2.5 miles) per hour, making it one of the speediest of all rattlesnakes. It is almost purely nocturnal. Sidewinders are not particularly venomous; their prey includes other reptiles, especially smooth-throated lizards and whiptail lizards.

Tropical Rattlesnake (*Crotalus durissus*). Also called the cascaval or South American rattlesnake, this is probably the most dangerous of the rattlesnake species. Its venom contains substantial amounts of neurotoxic material. It is estimated that a cascaval bite is about as potent as a puff adder and cobra bite combined. A very special antivenin was developed in order to lower the death rate from bites in South America, particularly in Brazil. Approximately 75% of all untreated adults bitten die from a cascaval bite. A peculiar aspect of a cascaval bite is paralysis of the neck muscles of the victim. This causes the head to suddenly drop to one side. Severe neural disturbances later appear, including auditory and visual impediments that lead to blindness and unconsciousness.

The Agkistrodon Snakes. Much more is known about American *Agkistrodon* snakes than about the Asian species. They are as familiar to Americans living in the southern and eastern United States as are rattlesnakes. The most important of the American *Agkistrodons* include the cottonmouth or water moccasin, and the copperheads.

Cottonmouth or Water Moccasin (*Agkistrodon piscivorus*). This is a compact, powerful snake and is treated with great respect when it is encountered in rice fields in the southern United States. A cottonmouth bite is seldom fatal, but it has extremely unpleasant consequences. The popular name is due to the white inside of the jaw, which is visible when the snake is irritated, the open mouth being a threat gesture. Large cottonmouths are a deep blue-black, quite unlike the juveniles, which have jagged, red-brown, light-edged bands on a flesh-colored background. Five to 15 young are born after a long gestation period. Females reproduce every other year.

Cottonmouths often stay in swamps and on low branches overhanging the water. They swim well and flee into the water when alarmed. The species is not a true aquatic snake, since it does not catch most of its prey (frogs and fishes) while swimming, but rather by lying in wait on the shore. Cottonmouths apparently form pairs frequently, as shown by studies on several islands off Florida. Like rattlesnakes, they often winter in groups.

Copperhead (*Agkistrodon contortrix*). Although the fatalities from snake bites in the United States are mainly attributed to the western diamond-back rattlesnake, the number of incidents is greater from copperhead bites. The copperhead attains a length up to 1 meter (3.2 feet) and is distributed in four subspecies across many eastern and southern states. Copperheads are strikingly colored, beautiful snakes whose bands, arranged in an irregular zigzag, may be narrow or broad and red-brown or cinnamon on a gray or pinkish background. The top of the head is a light reddish-brown. This may be an adaptation to their favorite habitat, which is foliage-covered forest floors. They adapt their coloration to their background. Copperheads are extremely adaptable snakes. In some heavily populated areas where rattlesnakes disappeared many years ago, such as along the Hudson River in New York State, copperheads are still prevalent. Copperheads winter in groups, frequently with rattlesnakes. A female bears 8 to 12 live young, which have sulfur-yellow tail tips.

Additional Reading

Beardsley, T. M.: "Snakes in the Grass," *Sci. Amer.*, 22 (February 1990).

Crews, D., and W. R. Garstka: "The Ecological Physiology of a Garter Snake," *Sci. Amer.*, **247**(5), 158–168 (November 1982).

Goldstein, E. J. C., et al.: Bacteriology of Rattlesnake Venom and Implications for Therapy," *J. Infect. Dis.*, **140**, 818 (1979).

Greene, H. W., and R. W. McDiarmid: "Coral Snake Mimicry," *Science*, **213**, 1207–1212 (1981).

Gruber, U.: "Blind Snakes, Primitive Snakes, and Wart Snakes," in "Grzimek's Animal Life Encyclopedia," Vol. 6, pp. 359–380, Van Nostrand Reinhold, New York, 1972.

Hediger, H.: "Snakes," in "Grzimek's Animal Life Encyclopedia," Vol. 6, pp. 345–358, Van Nostrand Reinhold, New York, 1972.

Lillywhite, H. B.: "Snakes, Blood Circulation, and Gravity," *Sci. Amer.*, 92 (December 1988).

Newman, E. A., and P. H. Hartline: "The Infrared 'Vision' of Snakes," *Sci. Amer.*, **246**, 3, 116–126 (1982).

Petzold, H. G.: "Cobras and Sea Snakes," pp. 415–438; and "Vipers and Pit Vipers," pp. 439–485 in "Grzimek's Animal Life Encyclopedia," Vol. 6, Van Nostrand Reinhold, New York, 1972.

Reinhard, W., and Z. Vogel: "Colubrid Snakes," in "Grzimek's Animal Life Encyclopedia," Vol. 6, pp. 381–414, Van Nostrand Reinhold, New York, 1972.

Savitzky, A. H.: "Hinged Teeth in Snakes," *Science*, **212**, 346–349 (1981).

Staff: "Treatment of Snakebite in the USA," *Med. Lett. Drugs Ther.*, **24**, 87 (1982).

Sutherland, S. K., et al: "Rationalisation of First Aid Measures for Elapid Snakebite," *Lancet*, **1**, 183 (1979).

Watt, C. H., Jr.: "Poisonous Snakebite Treatment in the United States, *JAMA*, **240**, 654 (1978).

Weiss, R.: "Snakebite Succor: Researchers Foresee Antivenom Improvements," *Science News*, 360 (December 8, 1990).

SNAPPERS (*Osteichthyes*). Of the order *Percomorphi*, family *Lutianidae*, snappers prefer tropical shallow inshore waters and usually occur in large schools. They possess bright, iridescent coloration, usually of blue, red, and yellow shades. They are known to migrate over great distances in search of food. They are carnivorous. They are an abundant fish and particularly in the Indo-Pacific region. The *Ocyurus chrysurus* (yellowtail snapper) is found in the tropical Atlantic and attains a length of about 2 feet (0.6 meter). The *L. synagris* (spot snapper) and *L. analis* (muttonfish) are found in American Atlantic waters. The established reputation of snappers as good food fishes has been tarnished from time to time by outbreaks of tropical fish poisoning, largely attributed to certain species of snappers. The poisonous qualities, however, appear to vary with particular region and time of year. Snappers considered in the poisonous category include: *Lutjanus bohar* (onespot snapper); *L. gibbus* and *L. vaigiensis* (red snappers); and *Aprion virescens* (blue-gray snapper). See also **Foodborne Diseases.**

SNIPE. See **Shorebirds and Gulls.**

SNIPE FLY (*Insecta, Diptera*). A two-winged fly with long legs and a conical abdomen. It belongs to the family *Rhagionidae*, sometimes called *Leptidae*. Some of the species suck blood and in the western part of the United States are very annoying to human beings. This insect receives the name deer fly in the west, a term applied to a small horsefly in the east.

SNOW. See **Precipitation and Hydrometeors.**

SNOW LEOPARD. See **Cats.**

SOAP. See **Colloid System.**

SOAPS. Chemically, a soap is defined as any salt of a fatty acid containing 8 or more carbon atoms. Structurally a soap consists of a hydrophilic (water compatible) carboxylic acid which is attached to a hydrophobic (water repellent) hydrocarbon. Soap molecules thus combine two types of behavior in one structure; part of the molecule is attracted to water and the other part is attracted to oil. This feature underlies the function of these materials as surface active agents, or surfactants. Soaps are one class of surfactants. The other classes generally are called detergents. See **Detergents.**

All surfactants, including soaps, demonstrate a common physical property—when they dissolve they preferentially concentrate at solution surfaces. These surfaces are known as the *interfacial regions* or regions where one continuous phase, such as water, stops and another, such as oil, begins. By their presence at the interface, surfactants lower the total energy associated with maintaining that boundary and thereby stabilize it. Without surfactants, a mixture of oil and water will soon separate into two distinct phases where the total surface area across which water and oil contact each other will be minimal. Adding soap to the water reduces its surface tension—the energy needed to maintain contact between the oil and the water. The oil then can be broken into microscopic droplets which are dispersed in the water. Creation of these droplets, however, is accompanied by a huge increase in the interfacial contact area between oil and water. The dispersion of the oil in water is only possible and only can be maintained over a period of time because the surfactant reduces the energy associated with the large surface over which oil and water are in contact with each other. This phenomenon is the basis for the cleansing action of soaps and other detergents. Stabilization of the interface between the water used to cleanse and oils and other water-insoluble soils facilitates the dispersion of these materials into the water.

Although soaps and synthetic detergents have similar physical properties, several factors distinguish between them. Soap is generally made from natural fats and oils (oleochemicals). Some important synthetic detergents are also derived from oleochemicals, but almost no ordinary soaps are produced from petrochemicals. Fats and oils are triglycerides which contain three fatty acids, the basic structural unit of soaps, chemically linked to a glycerine backbone. As the "soap" chemical structure basically exists in natural triglycerides, with relatively straightforward processing operations, soap can be obtained from fats and oils.

Another important distinguishing feature of soaps is that they form a curdy, insoluble compound in hard water due to interaction between the carboxylate soap structure and calcium and magnesium ions in the water. Synthetic detergents, which generally are based on sulfate or sulfonate chemical structures for the water-attracting portion of the molecule, have less affinity for these metals and thus work well in all types of water. In addition, since these synthetics maintain their surfactancy, they also function to disperse objectionable curd. For these reasons the synthetic detergents have generally replaced soaps in heavy-duty cleaning (laundry, floors, woodwork). Soaps, however, remain popular for mild cleaning and particularly for personal cleansing.

Personal Cleansing Soap Products

The major soap-based products which one commonly encounters are soap bars. Two broad categories of bar soaps may be defined: *ba-*

sic cleaning bars, which are natural soaps without extra ingredients and comprise about 20% of the market; and bars with special ingredients to provide a benefit beyond fundamental cleansing. The latter category may be further subdivided into *deodorant soaps* and *skin care bars*. Generally most of these bars command a higher retail price than basic cleaning bars, with skin care bars priced above deodorant soaps.

Deodorant soaps add impactful fragrances which are partially substantive to the skin and mask body odors, and antimicrobial agents. The antimicrobials, such as *Triclocarban*®, are deposited on the skin and inhibit bacterial growth and associated malodors.

Skin care bars are formulated with ingredients for which specific skin benefits are claimed. Consumers generally recognize and are concerned that personal cleansing products can dry the skin, leaving it feeling rough, itchy, and tight, and looking powdery and scaly. To counter these effects, particularly during the dry winter months, they may elect to use a cleansing bar containing a moisturizer, as well as increasing their use of body oils and hand and body lotions. Skin care claims for these products are based on the inclusion of moisturizers such as glycerin, cocoa butter, lanolin, cold cream and vitamins to the soap.

The mildness of soap bars toward the skin can also be enhanced by the process of *superfatting*. In superfatting, excess fatty acid is added to the soap during processing. This water-insoluble material functions as an emollient, significantly improving the mildness and the lathering of the bar.

Manufacture of Soap

Ingredients. The primary materials used in the manufacture of bar soaps are natural fats and oils. The performance and physical properties of soap bars can be varied by altering the blend of fats and oils used to make the neat soap. The most common materials used are top-quality animal tallows and coconut oil with blends ranging from 50% to 85% tallow. Generally it is found that bars containing higher proportions of coconut soap are physically harder, more brittle, lather more, and are more expensive to produce due to the higher cost of coconut oil. It is therefore common practice to vary the blend of tallows and coconut oil to meet the desired properties and price of each product.

These basic materials eventually are converted to their neutral salts by use of some alkaline material, such as sodium hydroxide. Additional, minor ingredients are added, e.g. sodium silicate or magnesium sulfate, to control alkalinity, odor, and aging stability.

The basic process is that of reacting fat stocks with alkali to form soap (direct saponification) and glycerin, followed by washing to remove the glycerin. Two methods of direct saponification are in common use (*kettle method* and *continuous saponification*). An alternative method is splitting fat stocks with water (hydrolysis) to form fatty acids and glycerine, followed by neutralization of the fatty acids with alkali.

Kettle Method. The pioneers used a simplified kettle process when they boiled animal fat and wood ashes (for alkalinity) for several hours in a large pot. The modern soap kettle has a capacity of 60,000–300,000 pounds (27,216–136,080 kg) and is equipped for heating, settling, and blending the fats, alkali, salt, and water.

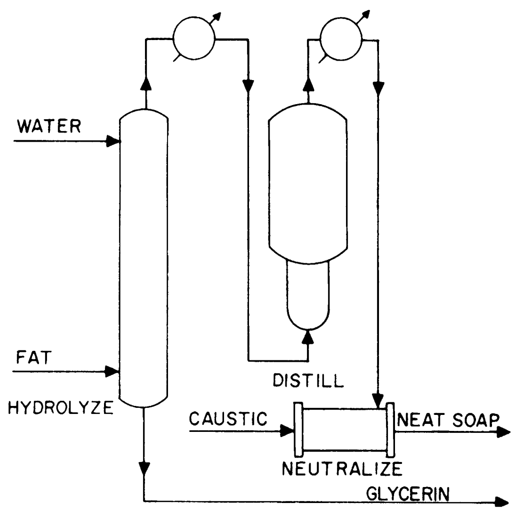
The kettle first is charged with fat and a sodium hydroxide solution. Then follows a sequence of heating, separating, and washing to convert the raw materials to *finished base soap* and to separate the impurities and byproducts. The process normally takes several days for any single kettle. Although there have been improvements in handling and purification such as continuous centrifugation, the basic kettle process of saponifying fats directly with caustic remains unchanged.

Continuous Saponification. Fat stocks plus caustic and salt solutions are fed continuously into an autoclave operating under pressure at typically about 250°F (120°C). A recycle stream provides sufficient soap concentration to solubilize the fat stream for good contacting with the caustic. The soap-lye-glycerin mix moves to a mixer/cooler to complete saponification. The cooler temperature reduces the solubility of soap in the lye and aids separation.

Glycerin and excess caustic are removed by several stages of countercurrent washing with fresh washing solution. The washing and separation stages usually take the form of a series of mixers and centrifugal separators or a continuous countercurrent contactor, such as a rotating disk contractor (RDC) in a vertical column. The mix from the saponifier is fed near the bottom of the RDC and washing solution near the top. The lower-density soap rises through the falling wash solution. Washed soap exits at the top while spent lye (glycerin plus lye solution) exits out the bottom of the RDC column. Spent lye is processed to recover the glycerin.

The washed soap is converted to finished base soap (neat soap) by a final composition adjustment called *fitting*. Fitting is accomplished by adding water (plus salt as needed), which causes a phase separation. Depending on the salt concentration the separated phase is either a lye or niger phase. A centrifuge or kettles can be used to separate the two phases.

Hydrolyzer Process. The development of continuous hydrolysis provides basic improvements in the processing of fats into soap. There are several advantages over the kettle process: (1) better quality soaps can be made from darker fats, (2) glycerin recovery is simplified because no salt is needed and the resulting finished glycerin is of higher quality, (3) a single hydrolyzer unit produces about the same quantity of soap as 10 kettles, thus effecting savings in manufacturing space and a reduction of in-process inventory, and (4) greater flexibility is possible in controlling the chemical and physical properties of the finished soap. The hydrolyzing process consists essentially of (1) hydrolysis, (2) fatty acid distillation, (3) post-hardening (optional), (4) neutralization, and (5) glycerin recovery. The basic hydrolyzer process is shown in the accompanying figure.



Basic hydrolyzer process used in soap manufacture.

Hydrolysis. Development of continuous hydrolyzing was the key step toward this continuous soap making process. In this reaction, fat and water react to form fatty acid and glycerin:

$$(RCOO)_3C_3H_5O + H_2O \leftrightharpoons 3RCOOH + C_3H_5(OH)_3$$

where R is an alkyl of C_8 or larger. This equation represents the complete hydrolysis. Actually, the reaction takes place in a stepwise fashion, forming intermediate diglyceride and monoglyceride.

The reaction can be accomplished only through intimate contact between water and fat molecules. High temperature makes it possible to dissolve an appreciable quantity of water in the fat phase and to obtain this intimate contact. At room temperature, water and fat are essentially insoluble. At elevated temperature, the solubility of water increases to 12–25%, depending upon the type of fat. At the higher temperatures, high pressures also are necessary to keep the water from flashing into steam.

The reaction is reversible. In order to make it proceed to the right, the proportion of water to fat can be increased or the glycerin can be removed. Removal of glycerin is used as the reaction-forcing method. The required combination of high temperature, high pressure, and continuous glycerin removal is accomplished in a countercurrent hydrolyzer column. Fat stocks, blended in the proper formula, are mixed with dry zinc oxide catalyst. The mixture is maintained at about 212°F (100°) to ensure dryness and to keep the catalyst in solution. Hot water for the hydrolysis reactions is put under high pressure by piston-type feed pumps with adjustable drives so that the rates and proportions of fat to water can be accurately controlled. The fat and water are heated to the hydrolyzing temperature by direct steam injection or by heat exchangers. The fats are pumped into the column near the bottom, and the water enters near the top. Thus, a countercurrent flow of water downward through rising fatty material is obtained.

The hydrolysis occurs in a two-phase reaction system. The fats and fatty acids flow continuously with droplets of water falling through them. Glycerin from the hydrolysis is dissolved in the excess water falling through the column. The rate limiting factor is the transfer of glycerin into the water droplets. Zinc oxide catalyzes the reaction of forming zinc soap, which increases the glycerin transfer across the oil-water interface. Fresh water entering the column at the top reduces the glycerin to the lowest possible point, while a glycerin-water seat maintained at the bottom of the column (where the glycerin content is highest) prevents fat from washing out.

The fatty material passes upward through the column with about 99% completeness in splitting. The fatty acids, saturated with water, are discharged through an orifice into a flash tank. The dissolved water vaporizes, cooling the fatty acids and blanketing them with steam. The fatty acid contains the zinc soap catalyst and the remaining unsplit fat.

The column, pumps, and piping in contact with the hot fatty acid are made from corrosion-resistant stainless steel. The column is a hollow vessel, containing no baffles, trays, or packing material of any kind. The quality of the hydrolyzing operation is determined by the degree of split obtained on the fat. The fatty acid stream contains very little free glycerin, if any.

Distillation. The second key step in continuous soapmaking is distillation. Originally, fatty acids made in hydrolyzers were acid washed to split out the zinc soap and then bleached to improve color, but continuous distillation of the hydrolyzer fatty acids results in lighter soap from darker stocks at lower cost.

The fatty acids from the hydrolyzer are collected in the still feed tank and vacuum-dried to reduce moisture to low levels. Then they are flash-distilled at an absolute pressure of 2–5 mm Hg. The still bottoms are recirculated through heat exchangers back to the still to carry the heat necessary for vaporizing the fatty acids. The still bottoms, which contain the zinc soap catalyst and unsplit fat, are removed from the system, acidulated to remove the zinc, and frequently used in animal feeds. The fatty acid vapors from the still pass to several water condensers in series. The condensed fatty acids drop to a surge tank for posthardening or directly for neutralization.

The two prime objectives of this process, maintenance of good odor and color in the distillate and proper bottoms yield, are achieved by effective control over vacuum, temperature, and distillation rate.

Posthardening. Not shown in the figure is an optional further treatment of the fatty acids known as *posthardening*. This operation involves hydrogenation of some of the unsaturated carbon-carbon bonds on the fatty acid molecules. Originally, the purpose of this step was to improve color and odor. As such, the hardening was intended only to eliminate polyunsaturates, leaving the majority of the monounsaturates unaffected. A greater amount of hardening can be performed, however, to tailor some of the physical properties of the finished bar characteristics. The fatty acids from distillation are heated and passed with a metered hydrogen supply through hardening tubes which contain a fixed bed of granular nickel catalyst where the hydrogenation takes place. The hardened fatty acids flow through a filter to remove traces of catalyst. The filtered stock drops to a flash tank, where excess hydrogen is removed. Hardening is controlled by temperature, pressure, hydrogen flow, residence time, and catalyst age. The fatty acids then are cooled for neutralization.

Neutralization. The saponification reaction between alkaline solutions and fatty acids is almost instantaneous:

$$RCOOH + NaOH \rightarrow RCOONa + H_2O$$

Each reactant is metered accurately into the neutralizer, where intimate mixing occurs and the reaction takes place. Soap from the neutralizer is discharged at about 200°F (93°C) to a blend tank equipped with agitation and recirculation to ensure uniform composition of the soap. This base soap (or *neat soap*) is stored until required for subsequent processing into finished bars.

The characteristics of the neat soap are controlled easily by accurately governing the composition of the alkaline solution used. Normal hydrolyzer neat soap contains about 69% actual soap, 30% water, and less than 1% NaCl, plus other stabilizers. Neat soap is a uniform, translucent, white, viscous fluid at 180–200°F (82–93°C).

Glycerin Recovery. The glycerin water stream from the hydrolyzer is concentrated by evaporation, purified, and subsequently sold or used in other processes.

Milled Bar Soap Manufacture. Milled soap is a high-grade soap in which critical crystal-phase changes have been brought about through the use of mixers, milling rolls, and plodders. The milled soap is made by drying a good grade of neat soap to about 15% moisture content, breaking up the crystalline structure that develops during drying and cooling, plasticizing and converting a sufficient portion of the soap to a desirable phase condition, deaerating and compacting the resulting mass, and forming it into bars. Perfume, coloring matter, preservatives, and special additives are incorporated prior to the milling operation. A milled bar is particularly hard, dense, and smooth, and it lathers freely without forming excessive soft soap on the surface of the bar.

Drying. Liquid base soap is dried from a 30% water liquid form to a solid of about 15% water content. If desired, some minor ingredients may be blended into the soap stream prior to drying. Methods of drying used in common practice are (1) chip drying, (2) atmospheric flash drying, and (3) vacuum flash drying.

Chip Drying. Sometimes called *ribbon drying*, this process involves spreading a thin layer of hot base soap on a large chilled drum which cools and firms up the soap. Drying is promoted primarily by the difference in water vapor pressure between the soap chips and the air surrounding them. No attempt is made to increase drying rate by heating the soap itself.

Atmospheric Flash Drying. A tower similar to a synthetic-granules spray-drying tower is used. The heat for drying, however, is put into the soap by heating it under high pressure before flashing it into the tower. During flashing, the pressure on the soap is abruptly relieved and soap moisture flashes to steam. Air to the tower is used for cooling. The soap temperature as it enters the flashing nozzles determines the final moisture of the dried soap.

An alternative method involves flashing the soap from the nozzle onto the surface of a chilled drum. The resultant solid soap is scraped off in flake form. This process called *chill flake drying*, is the method of choice for drying sticky soap/synthetic combination formulas.

Vacuum Flash Drying. In this most recent technique, drying takes place in a vacuum vessel similar to an atmospheric tower but smaller. The soap is similarly heated before flashing but under less pressure, so that boiling (actually drying of the soap) occurs in the heat exchangers. Since there is boiling in the heaters, the moisture of the dried soap depends primarily upon soap flow rate, soap pressure, and steam pressure to the heater and to a minor extent on the absolute pressure in the vacuum chamber. The final temperature of the soap depends entirely upon the absolute pressure in the vacuum chamber.

Mixing. After drying, the soap noodles or flakes are mixed with all additional ingredients required by the final product formula. Mixing is done in batch processes or continuously. These ingredients include dye, perfume, preservatives, deodorants, opacifier, and special purpose items. The type and proportion of these materials is largely what makes one brand of milled soap different from another.

In batch mixing, dried soap and additives are measured and dumped into a dry blender, where macro-mixing occurs. The batch process is cumbersome and slow, and it is difficult to maintain uniform quality. Continuous mixing operations for improving economy and efficiency of mixing include precision metering devices to measure the additives into the soap noodles as they are pulverized and conveyed through the mixer. Although these ingredients constitute but a small portion of the total product, their effect on the physical properties, e.g., softness, resistance to cracking, lathering, and resistance to dissolving, are considerable.

Milling. The three objectives of milling are (1) thorough and intimate final mixing of the soap, perfume, and other ingredients without overheating, (2) crushing lumps of overdried soap and pulverizing them into pieces too small to appear as lumps or hard specks in the finished bar, and (3) conversion of a sufficient portion of the soap into the waxy, plastic phase of cold working. Soap is milled by forcing it through a series of rolls, thus subjecting it to a strong shearing action. This cold working at the proper moisture content changes the crystalline structure or phase of the soap. Temperature control during milling is important. If the temperature is too low, the wrong crystal structure will be formed, resulting in soft soap or a hard, brittle structure prone to cracking. If the temperature is too high, the soap will become sticky and difficult to process further.

Another method for complete mixing and working uses multiple plodding and screening, in which the soap and additives are pushed together through finer and finer mesh screens.

Plodding. After milling, it is necessary to form the soap into a shape for making the final bar. This usually is accomplished with a plodder, which essentially is a large-size meat grinder with a barrel that terminates in a cone. The plodder functions to compact the pellets or flakes of soap into a solid mass, squeeze out any pockets of entrapped air, and extrude it as a firm, uniform, and continuous strip.

Operations which follow include cutting, stamping, wrapping, and packing.

Transparent Bar Manufacture. Most milled bars are opaque and contain a whitening agent (titanium dioxide) to create a uniform appearance. By eliminating this whitener and carefully controlling processing conditions, a bar which is transparent can be produced. This transparency results when the soap crystals are reduced to microscopic size which then allows light to pass through the structure. It is also important to achieve the correct soap phase. The control of soap phase is a function of the ratio of tallow to coconut soaps, the milling temperature, and soap moisture which must be maintained within very rigid limits.

Floating-Bar Manufacture. Base soap made from the desired blends of fat and oil first is flash-dried to a moisture content of about 22%. It then enters a mechanical mixer called a *crutcher*, where it is thoroughly mixed with perfume, preservatives, and air. The amount of air controls the density of the final product, giving the bar a density of less than one and making it floatable.

From the crutcher, the mix goes to a freezer to reduce the temperature of the soap to the point where it will hold its shape when extruded. In the earlier steps, the soap mix is in liquid form. Rapid chilling is required to put it into a solid state. The machine is similar to a commercial ice-cream freezer, consisting of a horizontal cylinder surrounded by a jacket and housing a rotating shaft (mutator) on which scraping blades are mounted. The liquid soap mix from the crutcher is pumped into one end of the cylinder. A refrigerated brine solution is circulated through the jacket to chill the soap. The scraping blades on the mutator remove the chilled soap from the cylinder walls and maintain uniformity of the mix. The nose of the freezer is equipped with an oblong orifice through which the soap is extruded, after chilling, in the form of a continuous ribbon which has the same cross section as the final bar. There follows a series of cooling, storing, stamping, and packaging operations.

Additional Reading

Bailey, A. E.: "Industrial Oil and Fat Products," Wiley, New York, 1964.
Dahlgren, R. Marc, and John N. Kalberg, Ivory Technical Center, The Procter & Gamble Company, Cincinnati, Ohio.
Shreve, R. N.: "Chemical Process Industries," 3rd Edit., 543–559, McGrawHill, New York, 1967.
Woolatt, E.: "The Manufacture of Soaps, Other Detergents, and Glycerin," Wiley, New York, 1985.

R. Marc Dahlgren and John N. Kalberg,
Ivory Technical Center, The Procter & Gamble Company,
Cincinnati, Ohio.

SOCIETY (Ecology). A group of individuals of the same species living together for mutual benefit, with some division of labor. The society is a high expression of colonial organization in the animal kingdom, and is not sharply separated from simpler forms of colonies.

In the simplest type of colony, as found among the 1-celled animals, the associated individuals are similar and each is capable of complete existence in itself. In the same group a slight division of labor appears, accompanied by structural differentiation of the individuals for different tasks. This form of organization persists in the coelenterates, bryozoans, and ascidians, with varying degrees of structural continuity in the colony and varying degrees of differentiation and division of labor.

Among the more highly organized animals the possibility of association of individuals in a complex society involving division of labor is expressed only among the social insects and man. The insect society, or colony as it is often called, continues the associated principle of structural specialization of the individual for its particular duties, with the resulting castes exemplified in a simple form by the queen, drone, and worker honeybees. Among the termites and ants the differentiation is much more extreme and complex. To a moderate degree the honeybee colony also shows specialization of behavior among the workers, which may engage in various activities within their powers according to the requirements of the colony at different times.

The human society differs from that of insects in the restriction of inherent fitness for special duties to less evident details of organization. While inherent fitness undoubtedly exists among men, they are structurally of approximately the same form and their specialization is largely a result of training. In other words, specialization of the individual in human society is conspicuously a specialization of behavior. Lack of structural specialization is compensated by the use of tools.

In all cases, the society is an extension of the prevailing biological principle that biological units of any degree of complexity can be associated together as component parts of a larger coordinated unit.

SODALITE. An isometric mineral, a sodium aluminum silicate containing sodium chloride, with the chemical composition $Na_4Al_3(SiO_4)_3Cl$, potassium sometimes replacing a small amount of sodium. It is commonly found as dodecahedrons or simply massive. When observed sodalite has a dodecahedral cleavage; conchoidal to uneven fracture; brittle; hardness, 5.5–6; specific gravity, 2.14–2.30; luster, vitreous to greasy; color grayish to greenish or yellowish, may be white. It is often a beautiful blue and may sometimes be red. It is transparent to translucent; streak, white. Sodalite is found in igneous rocks of nephelite-syenite type which have been produced from soda rich magmas. Sodalite also has been found in the lavas of Vesuvius. Common minerals associated with it are nephelite and cancrinite. It occurs in the Ilmen Mountains of the former U.S.S.R.; at Vesuvius and Monte Somma, Italy; in Norway and Greenland. In Canada, in British Columbia and in Ontario, beautiful blue sodalite is found; and in the United States similar material comes from Kennebec County, Maine. The mineral derives its name from the fact of its soda content.

SODA NITRE. The mineral soda nitre or Chile saltpeter is naturally occurring sodium nitrate, $NaNO_3$. Its hexagonal crystals are rare, this mineral usually being found in crystalline aggregates, crusts or masses. It is soft; hardness, 1.5–2; specific gravity, 2.266; vitreous luster; colorless or white to yellow or gray; transparent to opaque. Soda nitre is a most important mineral commercially, being used in the manufacture of nitric acid, other nitrates and fertilizers. The chief soda nitre deposits of the world are those found in the Atacama and Tarapaca deserts of northern Chile, although others exist in the Argentine and Bolivia. Some small deposits have been found in California, New Mexico and Nevada. The origin of these nitrate deposits is far from being well understood. They have been regarded as nitrates formed originally by oxidation of organic matter and subsequently leached out. Guano, the excrement of birds, might be the original source of the nitrates. Ground water and ancient marine deposits have been suggested as well as the possibility of derivation from nitric acid produced in the atmosphere during electrical storms. Some investigators consider that the nitrates may have come from volcanic sources.

SODIUM. Chemical element, symbol Na, at. no. 11, at. wt. 22.9898, periodic table group 1 (alkali metals), mp 97.82°C, bp 882.9°C, density 0.971 g/cm^3 (solid at 0°C), 0.9268 g/cm^3 (liquid at mp). Elemental sodium has a face-centered cubic crystal structure.

Sodium is a silvery-white metal, can be readily molded and cut by knife, oxidizes instantly on exposure to air, and reacts with water violently, yielding sodium hydroxide and hydrogen gas, consequently is preserved under kerosene, burns in air at a red heat with yellow flame. Discovered by Davy in 1807.

There is only one naturally occurring isotope, ^{23}Na. There are five known radioactive isotopes, ^{20}Na through ^{22}Na, and ^{25}Na, all with short half-lives except ^{22}Na with a half-life of 2.6 years. See also **Radioactivity.** In terms of abundance, sodium ranks sixth among the elements occurring in the earth's crust, with an average of 2.9% sodium in igneous rocks. In terms of content in seawater, the element ranks fourth (due mainly to excellent solubility of its compounds), with an estimated 50,000,000 tons of sodium per cubic mile of seawater. First ionization potential 5.138 eV. Oxidation potential $Na \rightarrow Na^+ + e^-$, 2.712 V.

Other important physical properties of sodium are given under **Chemical Elements.**

Sodium does not occur in nature in the free state because of its great chemical reactivity. Sodium occurs as sodium chloride in the ocean (1.14% Na), in salt deposits (salt, halite, NaCl), e.g., in Michigan, New York, Louisiana, in Great Britain, in Germany, in salt lakes, e.g., the Dead Sea (3% Na), Great Salt Lake; in common rocks (average of the solid shell of the earth 2.75% Na) as sodium nitrate (Chile saltpeter, $NaNO_3$) in Chile; as sodium borate (rasorite, kernite, $Na_2B_3O_7 \cdot 4H_2O$, in California; tinkal, $Na_2B_4O_7 \cdot 10H_2O$, in Tibet); as sodium carbonate Na_2CO_3 and sulfate Na_2SO_4 in certain salt lake areas. See also **Sodium Chloride.**

Although sodium metal was isolated in 1807, it remained a laboratory curiosity until Oersted discovered in 1824 that sodium metal will reduce aluminum chloride to produce pure aluminum metal. This discovery led to the development of a commercial process for the manufacture of sodium. The first cell was designed by Castner in 1886 and a plant was built in Niagara Falls, N.Y., because of availability of low-cost electric power, for the electrolysis of fused NaOH. This process was made obsolete in 1921 by introduction of the Downs process in which a mixture of fused sodium chloride and calcium chloride is electrolyzed to produce metallic sodium. The modern cells have four anodes (graphite) surrounded by a steel cathode. Wire mesh diaphragms extend down into the electrolysis zone to prevent recombination of product sodium and chlorine. The use of calcium chloride in the cell significantly lowers the melting point of the mix. Sodium chloride has a mp 800°C, calcium chloride, mp 772°C, the two-salt eutectic, mp 505°C. Calcium has limited solubility in sodium. The excess calcium reacts with the sodium chloride present, $Ca + 2NaCl \rightarrow 2Na + CaCl_2$, and thus does not contaminate the sodium metal to a large degree. The sodium, which is saturated with calcium, is cooled in a riser pipe. This reduces the solubility of Ca in Na, precipitating Ca, which falls back into the cell, where it reacts to form more Na. The Na that overflows at the top of the riser pipe contains 1% or less of Ca. The Na is further purified by filtration at a temperature near its melting point, reducing the Ca content to about 0.05%. The cells operate at about 8 V, with groups of 25 to 40 cells connected in series.

Uses. Like so many of the chemical elements, the compounds of sodium are far more important than elemental sodium—by several orders of magnitude.

Among the attractions of molten sodium metal as a heat-transfer medium are (1) low density compared with other metals and combinations of salts, contributing to low cost per unit volume and thus relative ease of pumping, sodium being about one-half that of the more commonly applied nitrate-nitrite heat-transfer salts, (2) relatively low vapor pressure even at temperatures as high as 550°C, (3) greater heat capacity than most common metals in liquid form, the thermal conductivity being 5 to 10 $\times$ greater than the conductivities of lead or mercury and 50 $\times$ higher than for most organic heat-transfer media, and (4) the viscosity of molten sodium is quite low. Despite these fine qualifications, however, the use of sodium as a heat-transfer medium has enjoyed a mixed reception over the years, partially attributable to a lack of marketing thrust in its behalf. Sodium is fifth among the metals in terms of electrical conductivity—hence bus bars are constructed from steel pipe filled with sodium. The characteristic yellow sodium light, created by the passage of an electric current through sodium vapor, is used for commercial and industrial lighting. Sodium is used to modify aluminum-silicon alloys. Normally coarse and brittle, such alloys can be transformed into fine-grained alloys with good casting properties through the addition of a fraction of 1% of sodium. Sodium also has been used as a hardening agent in bearing metals. When added with an alkaline-earth metal, such as calcium, sodium increases the hardness of lead. The German alloy "Bahnmetal" is an alloy of this type.

Generally, plain carbon steel containers are sufficient for handling metallic sodium at temperatures not in excess of the metal's boiling point. All-welded pipeline construction and bellows-sealed packless valves are usually used. Because of the metal's violent reactivity with H_2O, conventional fire extinguishers, including CO_2 and chlorinated hydrocarbons, should not be used. The preferred fire-retarding agents are salt, graphite, and soda ash, but they must be dry. Sand usually is not recommended because it is difficult to obtain perfectly dry sand in an emergency. In manufacturing operations involving sodium, particularly at reasonably high temperatures, an apron, leggings, and a complete face covering should be used. At normal temperatures, or where only small quantities of the metal are required, as may be the case in a research laboratory, conventional protective gear and goggles and gloves usually suffice.

Chemistry and Compounds: Sodium metal is obtained by electrolysis of fused sodium chloride or hydroxide out of contact with air. Its uses are limited in extent, but important in particular cases, as in the liberation of a metal from its chloride by reaction of sodium to form sodium chloride, and in certain reactions of organic chemistry.

The ionization potential of sodium (5.138 eV) is second to that of lithium and higher than those of the other alkali metals. However, the measured value of its oxidation potential against a normal aqueous solution of its ion is 2.712 V, the lowest of the group. Potassium is more electropositive in many of its reactions, even with water; though both react vigorously to produce the hydroxide and hydrogen, the reaction of potassium is more vigorous. With bromine, sodium reacts only slowly without heating and with iodine scarcely at all even on heating; potassium reacts violently with bromine, and with iodine on heating.

Because of the ease of removal of its single 3*s* electron (5.138 eV) and the great difficulty of removing a second electron (47.29 eV), sodium is exclusively monovalent in its compounds, which are electrovalent. Some experimental work indicates that the sodium alkyls may be covalent, but even they form conducting solutions in other metal alkyls.

Sodium Atoms Confined. In an interesting experiment conducted at the National Bureau of Standards in 1985, Migdall and colleagues trapped slow-moving netural Na atoms in a magnetic field that created an energy well for the atoms. Robinson (1985) reported that approximately 10^5 sodium atoms in a trap volume of 20 cubic centimeters were stopped for a brief instant of time. For many years, spectroscopists have visualized the ideal sample where a collection of atoms or molecules would reside motionless in space for a period of time. Because of this experiment, researchers are closer to this goal. In the experiment, it was found that the particles gradually leak out, with time constants ranging from 0.1 to 1 second. Similar experiments have been conducted at AT&T Bell Laboratories (Holmdel, New Jersey). The theoretical trapping time in a perfect vacuum has been estimated as greater than 1000 seconds. The importance of these experiments is explored in considerable detail by Robinson in the reference listed.

Like lithium, sodium and its compounds have been studied extensively in solution in liquid NH_3. Sodium metal in such solutions slowly or with catalysis forms the amide, $NaNH_2$. The solution of the metal is a powerful reducing agent, reacting with metallic salts to free the metal, with which it may form an intermetallic compound

$$Na + AgCl \rightarrow NaCl + Ag$$
$$9Na + 4Zn(CN)_2 \rightarrow 8NaCN + NaZn_4$$

Sodium chloride also forms the amide, or at low temperatures the pentammoniate, $NaCl \cdot 5NH_3$.

Like the other alkali metals, sodium forms compounds with virtually all the anions, organic as well as inorganic. These compounds are remarkable for their great variety and for the fact that the reactivity of sodium bicarbonate with many metallic oxides permits preparation of many compounds that are unstable in aqueous solution. While other alkali bicarbonates react similarly, the general discussion of these compounds, and the inorganic alkali salts generally, is appropriately given in this book under this entry for sodium, of which such a great number of inorganic (as well as organic) salts have been prepared.

Thus, normal (ortho) sodium arsenates $Na_3AsO_4{\cdot}xH_2O$ and acid arsenates exist both in solution and in the solid state, whereas the meta- and pyroarsenates exist only as solids, but are readily prepared by heating arsenic pentoxide, As_2O_5, and sodium bicarbonate in correct proportions to produce the primary and secondary sodium arsenates, whence the meta- and pyroarsenates are obtained by heating

$$NaH_2AsO_4 \xrightarrow{\text{Heat}} NaAsO_3 + H_2O$$
$$2\,Na_2HAsO_4 \xrightarrow{\text{Heat}} Na_4As_2O_7 + H_2O$$

Similarly, the boron salts include metaborates, $NaBO_2{\cdot}xH_2O$, tetraborates, $Na_2B_4O_7{\cdot}xH_2O$, other polyborates, $Na_2B_{10}O_{16}{\cdot}xH_2O$, at least one orthoborate, Na_3BO_3, and peroxyborates, such as $NaBO_3{\cdot}H_2O$. See also **Boron.** Other important sodium salts include the carbonates, cyanides, cyanates, hexacyanoferrates, $Na_4Fe(CN)_6$ and $Na_3Fe(CN)_6$, halides, polyhalides, hypohalites, halites, halates, perhalates, permanganates, ortho-, pyro-, meta-, fluoro-, and peroxyphosphates, hyposulfites, sulfites, sulfates, thiosulfates, peroxysulfates, polythionates, tungstates, vanadates, uranates, etc.

In addition to the simple compounds, sodium forms double salts of various types, although because of the relatively small size of the Na^+ ion, the number of sodium alums (see **Alum**) is relatively small.

In addition to the inorganic salts, sodium forms such binary compounds as a phosphide, Na_3P, by direct union with phosphorus, a nitride, Na_3N, by direct union with nitrogen when activated electrically (which decomposes partly to give sodium amide, NaN_3, also obtained by heating sodium nitrate with sodium amide) and the oxides. Sodium monoxide, Na_2O, is obtained by heating the nitrite with the metal, displacing the nitrogen. Sodium peroxide, Na_2O_2, is the most stable oxide, obtained by reaction of the elements. Sodium superoxide is known, NaO_2, and one other oxide, Na_2O_3, has been reported. Sodium hydroxide, NaOH, is very soluble in H_2O and soluble in alcohol. It is almost completely ionized in water at ordinary concentrations, although its basic character is less than those of the higher elements in the group ($pK_B = -0.70$).

The detailed chemistry and applications of some of the more important compounds, other than those already discussed, follow.

Aluminate: Sodium aluminate, $NaAlO_2$, white solid, (1) by reaction of aluminum hydroxide and NaOH solution, (2) by fusion of aluminum oxide and sodium carbonate, the solution reacts with CO_2 to form aluminum hydroxide. Used as a mordant, and in water purification. See also **Aluminum.**

Aluminosilicate: Sodium aluminosilicate is used as a water softener for the removal of dissolved calcium compounds.

Amide: Sodamide, sodamine, $NaNH_2$, white solid, formed by reaction of sodium metal and dry NH_3 gas at 350°C, or by solution of the metal in liquid ammonia. Reacts with carbon upon heating, to form sodium cyanide, and with nitrous oxide to form sodium azide, NaN_3.

Bromide. Sodium bromide, NaBr, white solid, soluble, mp 755°C. Used in photography and in medicine. See also **Bromine.**

Carbonates: Sodium carbonate (anhydrous), soda ash, Na_2CO_3, sodium carbonate decahydrate, washing soda, sal soda, $Na_2CO_3{\cdot}10H_2O$, white solid, soluble, mp 851°C, formed by heating sodium hydrogen carbonate, either dry or in solution. Commonly bought and sold in quantity on the basis of oxide Na_2O determined by analysis (58.5% Na_2O equivalent to 100.0% Na_2CO_3).

Soda ash is a very-high-tonnage chemical raw material and approaches a production rate of 10 million tons/year in the United States. About 40% of soda ash is used in glassmaking; approximately 35% goes into the production of sodium chemicals, such as sodium chromates, phosphates, and silicates; nearly 10% is used by the pulp and paper industry; the remainder going into the production of soaps and detergents and in nonferrous metals refining. The first process for preparing soda ash was developed by Leblanc during the first French Revolution. In the Leblanc process, sodium chloride first is converted to sodium sulfate and subsequently the sulfate is heated with limestone and coke: (1) $Na_2SO_4 + 2C \rightarrow Na_2S + 2CO_2$; (2) $Na_2S + CaCO_3 \rightarrow Na_2CO_3 + CaS$. During the mid-1800s, the Solvay process was introduced. In this process, CO_2 is passed through an NH_3-saturated sodium chloride solution to form sodium bicarbonate, then followed by calcination of the bicarbonate: (1) $NH_3 + CO_2 + NaCl + H_2O \rightarrow HNaCO_3 + NH_4Cl$; (2) $2HNaCO_3 + \text{heat} \rightarrow Na_2CO_3 + CO_2 + H_2O$. A large proportion of soda ash now is derived from the natural mineral trona, which occurs in great abundance near Green River, Wyoming. Chemically trona is sodium sesquicarbonate, $Na_2CO_3{\cdot}NaHCO_3{\cdot}2H_2O$. After crushing, the natural ore is dissolved in agitated tanks to form a concentrated solution. Most of the impurities (boron oxides, calcium carbonate silica, sodium silicate, and shale rock) are insoluble in hot H_2O and separate out upon settling. Upon cooling, the filtered sesquicarbonate solution forms fine needle-like crystals in a vacuum crystallizer. After centrifuging, the sesquicarbonate crystals are heated to about 240°C in rotary calciners whereupon CO_2 and bound H_2O are released to form natural soda ash. The crystals have a purity of 99.88% or more and handle easily without abrading or forming dust and thus assisting glassmakers and other users in obtaining uniform and homogeneous mixes.

Chlorate: Sodium chlorate, chlorate of soda, $NaClO_3$, white solid, soluble, mp 260°C, powerful oxidizing agent and consequently a fire hazard with dry organic materials, such as clothes, and with sulfur; upon heating oxygen is liberated and the residue is sodium chloride; formed by electrolysis of sodium chloride solution under proper conditions. Used (1) as a weedkiller (above hazard), (2) in matches, and explosives, (3) in the textile and leather industries.

Chloride: Sodium chloride, common salt, rock salt, halite, NaCl, white solid, soluble, mp 804°C. See also **Sodium Chloride.**

Chromate: Sodium chromate, $Na_2CrO_4{\cdot}10H_2O$, yellow solid, soluble, formed by reaction of sodium carbonate and chromite at high temperatures in a current of air, and then extracting with water and evaporating the solution. Used (1) as a source of chromate, (2) in leather tanning, (3) in textile dyeing, (4) in inks.

Citrate: Sodium citrate, $Na_3C_6H_5O_7{\cdot}5\frac{1}{2}H_2O$ white solid, soluble, formed (1) by reaction of sodium carbonate or hydroxide and citric acid, (2) by reaction of calcium citrate and sodium sulfate or carbonate solution, and then filtering and evaporating the filtrate. Used in soft drinks and in medicine.

Cyanide: Sodium cyanide, NaCN, white solid, soluble, very poisonous, formed (1) by reaction of sodamide and carbon at high temperature, (2) by reaction of calcium cyanamide and sodium chloride at high temperature, reacts in dilute solution in air with gold or silver to form soluble sodium gold or silver cyanide, and used for this purpose in the cyanide process for recovery of gold. The percentage of available cyanide is greater than in potassium cyanide previously used. Used as a source of cyanide, and for hydrocyanic acid.

Dichromate: Sodium dichromate, $Na_2Cr_2O_7{\cdot}2H_2O$, red solid, soluble, powerful oxidizing agent, and consequently a fire hazard with dry carbonaceous materials. Formed by acidifying sodium chromate solution, and then evaporating. Used (1) in matches and pyrotechnics, (2) in leather tanning and in the textile industry, (3) as a source of chromate, cheaper than potassium dichromate.

Dithionate: Sodium dithionate, "sodium hyposulfate," $Na_2S_2O_6{\cdot}2H_2O$, white solid, soluble, formed from manganese dithionate solution and sodium carbonate solution, and then filtering and evaporating the filtrate.

Fluorides: Sodium fluoride NaF, white solid, soluble, formed by reaction of sodium carbonate and hydrofluoric acid, and then evaporating. Used (1) as an antiseptic and antifermentative in alcohol distilleries, (2) as a food preservative, (3) as a poison for rats and roaches, (4) as a constituent of ceramic enamels and fluxes; sodium hydrogen fluoride, sodium difluoride, sodium acid fluoride, $NaHF_2$, white solid, soluble, formed by reaction of sodium carbonate and excess hydrofluoric acid, and then evaporating. Used (1) as an antiseptic, (2) for etching glass, (3) as a food preservative, (4) for preserving zoological specimens.

Fluosilicate: Sodium fluosilicate, Na_2SiF_6, white solid, very slightly soluble in cold H_2O, formed by reaction of sodium carbonate and hydrofluosilicic acid. Used (1) in ceramic glazes and opal glass, (2) in laundering, (3) as an antiseptic.

Formate: Sodium formate, $NaCHO_2$, white solid, soluble, formed by reaction of NaOH and carbon monoxide under pressure at about 200°C. Used (1) as a source of formate and formic acid, (2) as a reducing agent in organic chemistry, (3) as a mordant in dyeing, (4) in medicine.

Hydride: Sodium hydride, NaH, white solid, reactive with water yielding hydrogen gas and NaOH solution, formed by reaction of sodium and hydrogen at about 360°C. Used as a powerful reducing agent.

Hydroxide: Sodium hydroxide, caustic soda, sodium hydrate, "lye," NaOH, white solid, soluble, mp 318°C, an important strong alkali, not as cheap as calcium oxide (a strong alkali) nor sodium carbonate (a mild alkali), but of wide use. Formed (1) by reaction of sodium carbonate and calcium hydroxide in H_2O, and then separation of the solution and evaporation, (2) by electrolysis of sodium chloride solution under the proper conditions, and evaporation. Commonly bought and sold in quantity on the basis of oxide Na_2O determined by analysis (77.5% Na_2O equivalent to 100.0% NaOH). Used (1) in the manufacture of soap, rayon, paper ("soda process"), (2) in petroleum and vegetable oil refining, (3) in the rubber industry, in the textile and tanning industries, (4) in the preparation of sodium salts, (a) in solution, (b) upon fusion. See accompanying illustration.

Triple-effect evaporator used in concentrating soda solutions and preparation of solid sodium hydroxide.

Hypochlorite: Sodium hypochlorite, NaOCl, commonly in solution by (1) electrolysis of sodium chloride solution under proper conditions, (2) reaction of calcium hypochlorite suspension in water and sodium carbonate solution, and then filtering. Used (1) as a bleaching agent for textiles and paper pulp, (2) as a disinfectant, especially for water, (3) as an oxidizing reagent.

Hypophosphite: Sodium hypophosphite, $NaH_2PO_2 \cdot H_2O$, white solid, soluble, formed (1) by reaction of hypophosphorous acid and sodium carbonate solution, and then evaporating, (2) by reaction of NaOH solution and phosphorus on heating (poisonous phosphine gas evolved).

Hyposulfite: Sodium hyposulfite, sodium hydrosulfite (not sodium thiosulfate), $Na_2S_2O_4$, white solid, soluble, formed by reaction of sodium hydrogen sulfite and zinc metal powder, and then precipitating sodium hyposulfite by sodium chloride in concentrated solution. Used as an important reducing agent in the textile industry, e.g., bleaching, color discharge.

Iodide: Sodium iodide, NaI, white solid, soluble, mp 651°C, formed by reaction of sodium carbonate or hydroxide and hydriodic acid, and then evaporating. Used in photography, in medicine and as a source of iodide.

Manganate: Sodium manganate, Na_2MnO_4, green solid, soluble, permanent in alkali, formed by heating to high temperature manganese dioxide and sodium carbonate, and then extracting with water and evaporating the solution. The first step in the preparation of sodium permanganate from pyrolusite.

Nitrate: Sodium nitrate, nitrate of soda, Chile saltpeter, "caliche," $NaNO_3$, white solid, soluble, mp 308°C, source in nature is Chile, in the fixation of atmospheric nitrogen HNO_3 is frequently transformed by sodium carbonate into sodium nitrate, and the solution evaporated. Used (1) as an important nitrogenous fertilizer, (2) as a source of nitrate and HNO_3, (3) in pyrotechnics, (4) in fluxes.

Nitrite: Sodium nitrite, $NaNO_2$, yellowish-white solid, soluble, formed (1) by reaction of nitric oxide plus nitrogen dioxide and sodium carbonate or hydroxide, and then evaporating, (2) by heating sodium nitrate and lead to a high temperature, and then extracting the soluble portion (lead monoxide insoluble) with H_2O and evaporating. Used as an important reagent (diazotizing) in organic chemistry.

Oleate: Sodium oleate, $NaC_{18}H_{33}O_2$, white solid, soluble, froth or foam upon shaking the H_2O solution (soap), formed by reaction of NaOH and oleic acid (in alcoholic solution) and evaporating. Used as a source of oleate.

Oxalates: Sodium oxalate, $Na_2C_2O_4$, white solid, moderately soluble, formed (1) by reaction of sodium carbonate or hydroxide and oxalic acid, and then evaporating, (2) by heating sodium formate rapidly, with loss of hydrogen. Used as a source of oxalate; sodium hydrogen oxalate, sodium binoxalate, sodium acid oxalate, $NaHC_2O_4 \cdot H_2O$, white solid, moderately soluble.

Palmitate: Sodium palmitate, $NaC_{16}H_{31}O_2$, white solid, soluble, froth or foam upon shaking the H_2O solution (soap), formed by reaction of NaOH and palmitic acid (in alcoholic solution) and evaporating. Used as a source of palmitate.

Permanganate: Sodium permanganate, permanganate of soda, $NaMnO_4$, purple solid, soluble, formed by oxidation of acidified sodium manganate solution with chlorine, and then evaporating. Used (1) as disinfectant and bactericide, (2) in medicine.

Phenate: Sodium phenate, sodium phenoxide, sodium phenolate, $NaOC_6H_5$, white solid, soluble, formed by reaction of sodium hydroxide (not carbonate) solution and phenol, and then evaporating. Used in the preparation of sodium salicylate.

Phosphates: Trisodium phosphate, tribasic sodium phosphate, $Na_3PO_4 \cdot 12H_2O$, white solid, soluble, formed (1) by reaction of sodium hydroxide and the requisite amount of phosphoric acid, and then evaporating, (2) by reaction of disodium hydrogen phosphate plus sodium hydroxide, and then evaporating. Used (1) as a cleansing and laundering agent, (2) as a water softener, (3) in photography, (4) in tanning, (5) in the purification of sugar solutions; disodium hydrogen phosphate, dibasic sodium phosphate, $Na_2HPO_4 \cdot 12H_2O$, white solid, soluble, formed (1) by reaction of dicalcium hydrogen phosphate and sodium carbonate solution, and then evaporating the solution, (2) by reaction of sodium carbonate and the requisite amount of phosphoric acid, and then evaporating. Used (1) in weighting silk, (2) in dyeing and printing textiles, (3) in fireproofing wood, paper, fabrics, (4) in ceramic glazes, (5) in baking powders, (6) to prepare sodium pyrophosphate; sodium dihydrogen phosphate, monobasic sodium phosphate, $NaH_2PO_4 \cdot H_2O$, white solid, soluble, formed (1) by reaction of sodium carbonate and the requisite amount of phosphoric acid, and then evaporating, (2) by reaction of calcium monohydrogen phosphate and sodium carbonate solution, and then evaporating the solution. Used (1) in baking powders, (2) in medicine, (3) to prepare sodium metaphosphate; sodium pyrophosphate, $Na_4P_2O_7 \cdot 10H_2O$, white solid, soluble, mp about 900°C, formed by heating disodium hydrogen phosphate to complete loss of water, followed by crystallization from water solution. Used in electroanalysis; sodium metaphosphate, $NaPO_3$, white solid, soluble, mp 617°C, formed by heating sodium dihydrogen phosphate or sodium ammonium phosphate to complete loss of water, is an easily fusible phosphate forming colored phosphates with many metallic oxides, e.g., cobalt oxide. The hexametaphosphate, $(NaPO_3)_6$, is an important water-conditioning agent forming soluble complex compounds with many cations, e.g., Ca^{2+}, Mg^{2+}. Many polyphosphate compounds are known; their various uses include water softening and ion exchange. They are widely formulated in detergents, as are several of the simpler phosphates.

Phosphites: Disodium hydrogen phosphite, $Na_2HPO_3 \cdot 5H_2O$, white solid, soluble, formed by reaction of phosphorous acid and sodium carbonate, and then evaporating at a low temperature, mp of anhydrous salt is 53°C, at higher temperatures yields sodium phosphate and phosphine gas; sodium dihydrogen phosphite, $NaH_2PO_3 \cdot 2\frac{1}{2}H_2O$, white solid, soluble, formed by reaction of phosphorous acid and NaOH cooled to −23°C when the crystalline salt separates.

Salicylate: Sodium salicylate, $NaC_7H_5O_3$, white solid, soluble, formed by reaction of sodium phenate and CO_2 under pressure. Used as a source of salicylate and for salicylic acid.

Silicate: Sodium silicate, sodium metasilicate, "water glass," Na_2SiO_3, colorless (when pure) glass, soluble, mp 1,088°C, formed by reaction of silicon oxide and sodium carbonate at high temperature; solution reacts with CO_2 of the air, or with sodium carbonate solution or ammonium chloride solution, yielding silicic acid, gelatinous precipitate. Sodium silicate solution is used (1) in soaps, (2) for preserving eggs, (3) for treating wood against decay, (4) for rendering cloth, paper, wood noninflammable, (5) in dyeing and printing textiles, (6) as an adhesive (e.g., for paper boxes) and cement. Sold as granular, crystals, or 40° Baumé solution.

Silicoaluminate: (See aluminosilicate, above.)

Silicofluoride: (See fluosilicate, above.)

Stearate: Sodium stearate, $NaC_{18}H_{35}O_2$, white solid, soluble, froth or foam upon shaking the water solution (soap), formed by reaction of NaOH and stearic acid (in alcoholic solution) and evaporating. Used as a source of stearate.

Sulfates: Sodium sulfate (anhydrous), "salt cake," Na_2SO_4, sodium sulfate, decahydrate, "Glauber's salt," $Na_2SO_4 \cdot 10H_2O$, white solid, soluble, formed by reaction of sodium chloride and H_2SO_4 upon heating with evolution of hydrogen chloride gas. Used (1) in dyeing, (2) along with carbon in the manufacture of glass, (3) as a source of sulfate, (4) to prepare sodium sulfide; sodium hydrogen sulfate, sodium bisulfate, sodium acid sulfate, "nitre cake," $NaHSO_4$, white solid, soluble, formed by reaction of sodium nitrate and H_2SO_4, upon heating, with evolution of HNO_3. Used (1) as a cheap substitute for H_2SO_4. (2) in dyeing, (3) as a flux in metallurgy; sodium pyrosulfate, $Na_2S_2O_7$, white solid, soluble, formed by heating sodium hydrogen sulfate to complete loss of H_2O.

Sulfides: Sodium sulfide, Na_2S, yellowish to reddish solid, soluble, formed (1) by heating sodium sulfate and carbon to a high temperature. Used (1) as the cooking liquor reagent (along with sodium hydroxide) in the "sulfate" or "kraft" process of converting wood into paper pulp, (2) as a depilatory, (3) in sheep dips, (4) in photography, engraving and lithography, (5) in organic reactions, (6) as a source of sulfide, (7) as a reducing agent; sodium hydrogen sulfide, sodium bisfulfide, sodium acid sulfide, NaHS, formed in solution by reaction of NaOH or carbonate solution and excess H_2S.

Sulfites: Sodium sulfite, Na_2SO_3, white solid, soluble, dilute solution readily oxidized in air, but retarded by mannitol (carbohydrates), formed by reaction of sodium carbonate or hydroxide solution and the requisite amount of SO_2, at high temperature yields sodium sulfate and sodium sulfide. Used (1) as a source of sulfite, (2) as a reducing agent, (3) to prepare sodium thiosulfate, (4) as a food preservative, (5) as a photographic developer, (6) as a bleaching agent and antichlor in the textile industry; sodium hydrogen sulfite, sodium bisulfite, sodium acid sulfite, $NaHSO_3$, white solid, soluble, formed by reaction of sodium carbonate solution and excess sulfurous acid. Uses similar to those of sodium sulfite.

Tartrate: Sodium tartrate, $Na_2C_4H_4O_6 \cdot 2H_2O$, white solid, soluble, formed by reaction of sodium carbonate solution and tartaric acid. Used in medicine; sodium potassium tartrate, Rochelle salt, $NaKC_4H_4O_6 \cdot 4H_2O$, white solid, soluble. Used (1) in medicine, (2) as a source of tartrate.

Thiosulfate: Sodium thiosulfate, "Hypo" $Na_2S_2O_3 \cdot 5H_2O$, white solid, soluble, formed by reaction of sodium sulfite and sulfur upon boiling, and then evaporating. Used (1) in photography as fixing agent to dissolve unchanged silver salt, (2) as a reducing agent and antichlor. See also **Sodium Thiosulfate.**

Tungstate: Sodium tungstate, $Na_2WO_4 \cdot 2H_2O$, white solid, soluble, by reaction of NaOH solution and tungsten trioxide upon boiling, and then evaporating. Used (1) in fireproofing fabrics, (2) as a source of tungsten for chemical reactions.

Uranate: Sodium uranate, uranium yellow, Na_2UO_4, yellow solid, insoluble, formed by reaction of soluble uranyl salt solution and excess sodium carbonate solution. Used (1) in the manufacture of yellowish-green fluorescent glass, (2) in ceramic enamels, (3) as a source of uranium for chemical reactions.

Vanadate: Sodium vanadate, sodium orthovanadate, Na_3VO_4, white solid, soluble, formed by fusion of vanadium pentoxide and sodium carbonate. Used (1) in inks, (2) in photography, (3) in dyeing of furs, (4) in inoculation of plant life.

The larger number of organic compounds of sodium are for great part derivatives of oxygen-containing compounds such as salts of organic acids (several of which are discussed above), alcoholic and phenolic compounds (carboxylates, alkoxides, phenoxides, etc.). However, in some cases, sodium derivatives of nitrogen-containing compounds, as sodium benzamide, $C_6H_5C(O)NHNa$, and sodium anilide, C_6H_5NHNa, contain sodium-nitrogen bonds, while even sodium-boron bonds exist in certain boron-containing compounds, as sodium triphenylborene, $NaB(C_6H_5)_3$, and others; and in a number of compounds sodium is carbon-connected, as in methylsodium, CH_3Na, ethylsodium, C_2H_5Na, cyclopentadienylsodium, C_5H_5Na, and sodium triphenylmethane, $NaC(C_6H_5)_3$.

The organometallic compounds of sodium may be divided into two groups, differing in properties. One group, e.g., ethylsodium, consists of compounds that are colorless, insoluble in organic solvents, and that electrolyze readily in diethylzinc solution. Another group, e.g., benzylsodium, $C_6H_5CH_2Na$, are colored, and soluble in organic solvents.

Like all the alkali metals, sodium coordinates with salicylaldehyde. Its tetracovalent compounds, with those of potassium, are the more stable of the group, for the following reasons: (1) Increasing ionic size carries with it increasing electropositiveness and ease of ionization, which diminishes the tendency to coordinate. (2) The increasing distance of the nucleus from the coordinating electrons with increasing atomic volume makes it less likely that additional electrons will be held with ease. (3) On the other hand, there is an increase in the maximum coordination number with the elements of higher atomic number. These factors are in keeping with a maximum stability for the tetracovalent compounds occurring with sodium.

Sodium in Biological Systems: Sodium is essential to higher animals which regulate the composition of their body fluids and to some marine organisms. The several important roles played by the sodium cation in biological systems, frequently in concert with the potassium cation are described in the entry on **Potassium and Sodium (In Biological System).**

Additional Reading

Hammond, C. R.: "The Elements" in *Handbook of Chemistry and Physics*, 67th Edition, CRC Press, Boca Raton, Florida, 1986–1987.

Meyers, R. A.: "Handbook of Chemicals Production Processes," McGraw-Hill, New York, 1986.

Migdall, A. L., Prodan, J. V., et al.: *Phys. Rev. Lett.*, **54**, 2596 (1985).

News: "Four Groups Build More Efficient Atom (Sodium) Traps," *Science*, **237**, 26–27 (1987).

Prodan, J., Migdall, A., et al.: *Phys. Rev. Lett.*, **54**, 992 (1985).

Robinson, A. L.: "Sodium Atoms Stopped and Confined," *Science*, **229**, 39–41 (1985).

Sax, N. I.: "Dangerous Properties of Industrial Materials," 6th Edition, Van Nostrand Reinhold, New York, 1984.

Sax, N. R., and R. J. Lewis, Sr.: "Dangerous Properties of Industrial Materials," 8th Edition, Van Nostrand Reinhold, New York, 1992.

Sittig, M.: "Sodium," in McGraw-Hill Encyclopedia of Chemistry, McGraw-Hill, New York, 1983.

Staff: "ASM Handbook—Properties and Selection: Nonferrous Alloys and Pure Metals," ASM International, Materials Park, Ohio, 1990.

Staff: "Handbook of Chemistry and Physics," 73rd Edition, CRC Press, Boca Raton, Florida, 1992–1993.

SODIUM CHLORIDE. NaCl, formula weight 58.44, white solid, cubic crystal structure, mp 800.6°C, bp 1,413°C, sp gr 2.165. Commonly called "salt," the mineral name for rock salt is *halite*. See also **Halite.** The compound is soluble in H_2O (35.7 g/100 g H_2O at 0°C; 39.8 g/ 100 g H_2O at 100°C), only slightly soluble in alcohol, and insoluble in HCl. Sodium chloride is produced in nearly all nations of the world, but some only have a sufficient supply for local needs. The leading salt-producing nations include the United States, China, the former U.S.S.R., West Germany, France, the United Kingdom, India, Italy, Canada, and Mexico. In 28 states of the United States and in several provinces of Canada, salt occurs as bedded or domed deposits. Most of the rock salt produced in the United States comes from Michigan, New York, Texas, Ohio, Louisiana, and Kansas. Purity ranges from 97% NaCl for Kansas salt to 99% purity and higher for Louisiana salt. The main impurities are calcium sulfate (0.5–2%), dolomite, quartz, calcite, and traces of iron oxides. Natural rock salt is mined much as coal and usually marketed without purification, after crushing and screening. For most industrial and consumer requirements, the impurities are harmless. There is no evidence that bacteria exist in rock salt. Additionally, there is some solar salt production in the Great Salt Lake area of Utah and on the west coast. Salt deposits date back to past geologic ages and are believed to be the results of evaporated impounded sea water.

Purified salt for table and industrial processing requirements of a special nature is made by dissolving raw sodium chloride in H_2O and then evaporating the H_2O to form a final product. There are several types of evaporated salt, including *granulated salt* in which each crystal is a tiny cube, and *grainer* or *flake salt*, made up of irregularly shaped crystals, often thin and flaky and unusually soft. A process for producing evaporated salt is shown in Fig. 1. Holes are drilled into the salt deposits, after which H_2O is pumped into the beds to create a brine which then is brought to the surface for refining. In this method, all insolubles are left in the bed. After some pretreatment to remove hardness and dissolved gases, the semipure brine is evaporated in multiple-effect vacuum pans. The salt crystallizes as perfect cubes of NaCl. In the system shown, each vacuum pan performs not only as an evaporator, but also as a boiler. The vapors from a preceding pan are used to heat the contents of the following pan. This system of heat economizing is possible because each succeeding pan in the series is under less pressure—hence the contents boil at a lower temperature. See Fig. 2. The lower pressure in succeeding pans results from condensation of the vapors as well as assistance from vacuum pumps. Crystal size is controlled by evaporation rate, the latter depending on the degree of vacuum, temperature, and agitation maintained. When grown to proper size, the crystals drop to the bottom of the pans and fall into the salt legs, from which they are drawn continuously in the form of a slurry. After washing, filtering, cooling, and screening, they are packaged. See also **Evaporation.**

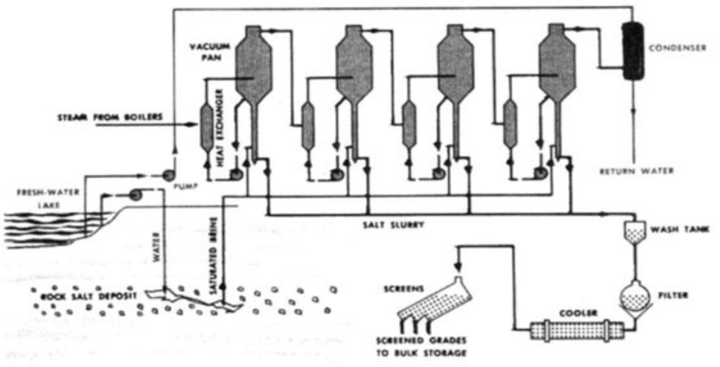


Fig. 1. Multiple-effect vacuum pans used in production of sodium chloride from brine. The saturated brine is formed by pumping fresh water directly into the rock salt deposit, leaving insoluble materials in the deposit.

Grainer salt is made by surface evaporation of brine in flat pans open to the atmosphere. Heat usually is furnished by steam pipes located a few inches below the tank bottom. Crystals form at the surface of the brine and are held there temporarily by surface tension. Thus, they grow laterally for awhile and form thin flakes. But, as they grow, they tend to sink and this process imparts a peculiar, hollow pyramid-like structure to them. Such crystals are called *hopper crystals*. Ultimately, the crystals sink to the bottom where they are scraped to one end of the pan. The crystals are fragile and during handling they break up, finally assuming a flake-like shape. Thus, the term *flake salt*.

In the *recrystallizer process* for making salt, advantage is taken of the fact that the solubility of NaCl increases with temperature whereas the solubility of the principal impurity, $CaSO_4$, decreases with temperature. In *solar* facilities, the raw brine is pumped into concentrating ponds

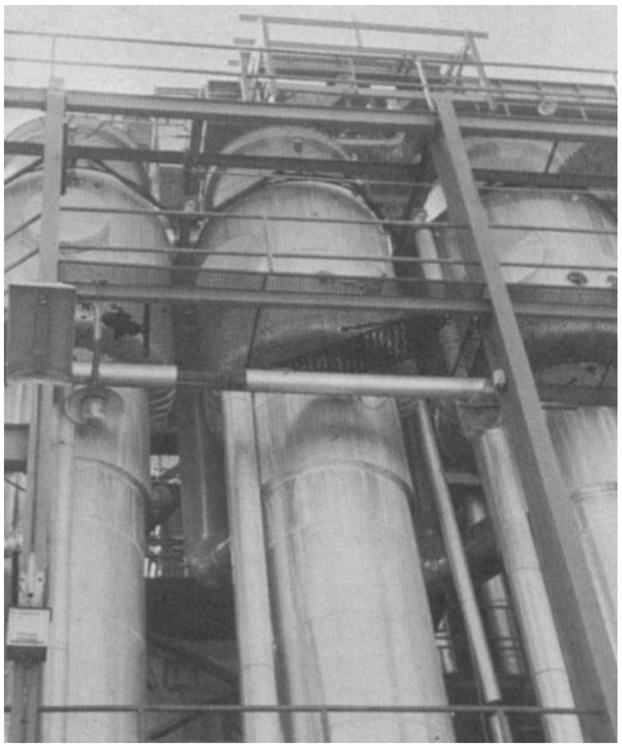


Fig. 2. Portion of train of evaporator bodies in a multiple-effect vacuum evaporation system used in production of sodium chloride.

where most of the H_2O is evaporated. Some of the impurities are precipitated out in this stage, after which the saturated brine is transferred to crystallization ponds where the salt crystallizes out at a high degree of purity. Since evaporation occurs at the surface of the ponds, hopper crystals are formed as in the grainer process, with flake salt being the final product.

Uses. Sodium chloride is a very high-tonnage material. In addition to its familiar use in the diets of man and animals, representing a small part of total production, large quantities are used by highway departments to control icy road conditions, in agriculture, and as a basic chemical raw material. The chemical industry consumes about two-thirds of the salt produced, the majority of it going to electrolytic plants. Some of the basic inorganic chemicals that require salt as a starting material include soda ash, calcium chloride, caustic soda, sodium sulfate, sodium bisulfate, HCl, sodium cyanide, sodium hypochlorite, and chlorine. See also **Chlorine; Sodium.**

Salt and the Diet

In food processing, the preservative and organoleptic qualities of salt are well established and it is fully appreciated why use of salt even to excess is attractive to food processors. Excessive usage is also habitual among people who "salt first and taste later." Dickinson in 1979 reported that just over 1 million tons (900,000 metric tons) of salt are used in foods and in connection with eating in the United States in a given year. Nearly an additional 2 million tons (1.8 million metric tons) are used in the agricultural field, much of which is consumed by livestock. The total daily intake of the North American consumer as of 1980 is estimated to be in the range of 10–12 grams of salt, which reduces to a range of 3900–4700 milligrams of sodium. Highly salted snack foods, the consumption of which has increased markedly in many parts of the world during recent years, accounts for a significant consumption of salt. In addition, certain other food ingredients, such as monosodium glutamate and soy sauce, sometimes used in excess, also contribute to the average intake of sodium.

Sodium and chloride are not normally retained in the body even when there is a high intake. See also **Chloride.** Amounts consumed in excess of need are excreted, so that the level in the body is maintained within very narrow limits, as is also the chloride, regardless of intake. The primary route of excretion is via the urine, with substantial amounts also lost in sweat and feces. About 50% of the sodium in the human body is located in the extracellular body fluids; 10% inside the cells; and 40% in the bones. Chloride is found mainly in the gastric juice and other body fluids.

Essential though sodium is to the normal functioning of the human body, there has been considerable concern over the last few years about the amount of salt in the diet, this concern centering mainly on possible relationship between salt and hypertension (high blood pressure).

Hypertension afflicts more than 20% of the world population, with an estimated 24 million cases in the United States as of 1980. In 1876, Marx reported that in about 90% of these cases, the actual causes of hypertension cannot be pinpointed—this in face of the fact that research on the possible role of sodium in essential hypertension has been underway for 60 years or longer. Tests of unmedicated persons with essential hypertension have been found to indicate a lowering of blood pressure when sodium intake is restricted below one gram per day—and that the blood pressure rises again if additional sodium is taken. However, in other studies, some persons have retained a normal blood pressure level even when fed substantially increased amounts of salt (or other sodium-ion-furnishing substances). In 1976, Freis reported positive correlations between estimated average salt consumption of various ethnic populations and their incidence of hypertension. But such studies are complicated by many factors, including the inability to control or eliminate other possible causes of hypertension, such as obesity, genetic predisposition, general nutritional status, and potassium intake. It also has been generally proved extremely difficult to determine differences between individuals within these cultures.

Nevertheless, the concern remains on the part of a large number of professional people who feel that one day a definitive correlation will be made. And, with considerable awareness of the lay public in this regard, very definite pressures are being exerted on food processors to reduce salt usage and to more accurately label their merchandise in this regard.

The physiology of the sodium-potassium relationship is explained in some detail in the entry on **Potassium and Sodium (In Biological Systems).**

Concerning the sodium content (much of which is derived from salt), the following composition data may be of interest. The figures in parentheses are milligrams of sodium per 100 grams of food.

Meats: Canadian bacon (2555), bacon (1077), cured ham (860), beef liver (136), pork chops (60), ground beef (48).

Cheeses: Parmesan (1848), process (1421), blue (1396), brick (557), cream cheese (294).

Other dairy products: Ice cream (83), whole milk (50), sherbet (45).

Miscellaneous foods: Pretzels (7800), soy sauce (regular) (6082), dill pickles (4000-5000), soy sauce (mild) (3569), green olives (2400), soda crackers (1100), salted peanuts (groundnuts) (418), eggs (122).

Vegetables: Beet greens (130), celery (126), dandelion greens (76), kale (75), spinach (60), beets (60), watercress (52), turnips (49), carrots (47), artichokes (43), collards (43), mustard greens (32), Chinese cabbage (20). Other common vegetables range between (10) and (18).

Sodium Chloride and Energy

As pointed out by Wick (*Oceanus*, **22**, 4, 28, 1980), most of the energy in the oceans is bound in thermal and chemical forms. Although thermal energy is presently commanding the most attention, within the past few years another, rather unusual, form has received notice. Where rivers flow into the oceans a completely untapped source of energy exists—represented by a large osmotic pressure difference between fresh and salt water. If economical ways to tap these salinity gradients could be developed, large quantities of energy would be available. See the entry on **Solar Energy.**

SODIUM THIOSULFATE. $Na_2S_2O_3 \cdot 5H_2O$, formula weight 248.19, white crystalline solid, decomposes above 48°C, sp gr 1.685. Also known as "hypo" and sometimes misnamed "hyposulfite," sodium thiosulfate is very soluble in H_2O (301.8 parts in 100 parts H_2O at 60°C), soluble in ammonia solutions, and very slightly soluble in alcohol. When sodium thiosulfate is added to an acid, thiosulfuric acid $H_2S_2O_3$ may be formed, but only for an instant, immediately decomposing into sulfur and SO_2.

Sodium thiosulfate is used (1) to dissolve silver chloride, bromide, iodide in the photographic "fixing" bath, soluble sodium silver thiosulfate being formed plus sodium chloride, bromide, iodide, (2) in reaction with iodine in solution, sodium tetrathionate and sodium iodide being simultaneously formed, or with ferric salt solution, sodium tetrathionate and ferrous being simultaneously formed, (3) in reaction with chlorine as an "antichlor" forming sulfate and chloride. Sodium thiosulfate reacts with silver nitrate solutions yielding silver sulfide, brown precipitate, and with permanganate yielding manganous. Sodium amalgam changes sodium thiosulfate to sodium sulfide plus sodium sulfite.

Sodium thiosulfate is formed (1) by reaction of sodium sulfite solution and sulfur upon warming, (2) by reaction of sodium sulfite solid and sulfur upon heating, (3) by complex reaction of sulfur and sodium hydroxide solution upon warming—sulfur yields sodium sulfide plus sodium sulfite and the latter reacts with excess sulfur forming sodium thiosulfate, and the sodium sulfide present may be converted into sodium thiosulfate by passing in SO_2 until the solution changes from yellow to colorless.

There are numerous other thiosulfates, including potassium, magnesium, calcium, barium, mercury, lead, and silver. All are soluble in H_2O except Ba, Pb, and Ag thiosulfates.

Thiosulfates are commonly identified as follows:

1. Dilute acids precipitate sulfur from thiosulfates (difference from sulfides and sulfites).
2. Zinc sulfate and sodium hexacyanoferrate(II) give no color (difference from sulfites).

SOFTWARE (Computer System). The totality of programs, procedures, rules, and (possibly) documentation used in conjunction with computers, such as compilers, assemblers, narrators, routines, and subroutines. References are made to software and hardware parts of a system where the hardware comprises the physical (mechanical and electronic) components of the system. In some machines, the instructions are microprogrammed in a special control storage section of the machine using a more basic code which is actually wired into the machine. This is contrasted with the situation where the instructions are wired into the control unit. The microprogram technique permits the economic construction of various size machines which appear to have identical instruction sets. However, microprograms are not generally considered as software. They are sometimes called firmware. See also **Digital Computer;** and **Microprogram;** and terms listed under alphabetical index.

SOFTWOODS. See **Wood.**

SOIL. All consolidated earth material over bedrock. Soil is approximately equivalent to regolith.[1] Agriculturally, soil is any one of many varied natural media that support or can support land plant growth outdoors; or, when in containers, indoors. The lower limit of *topsoil* is normally the lower limit of biologic activity, which usually coincides with the common rooting of native perennial land plants. The word *soil* is derived from the Latin *solum* for "ground."

The upper part of the regolith is divided into topsoil and subsoil. The topsoil is usually a relatively thin layer or zone of the more highly decomposed mineral constituents of the regolith and contains a varying proportion of organic material called *humus*. This soil zone is the habitat of the shallow-rooted plants, such as most grasses. The topsoil usually passes gradationally into the subsoil which supplies some of the moisture and food for the deeply rooted plants and trees. The subsoil may or may not pass gradationally into the underlying bedrock. Topsoil is easily destroyed by erosion when not protected by a mantle of vegetation.

Soil serves: (1) As a foundation for holding plants in place, whether tiny grasses or huge trees; (2) as a protective covering for the root structures of plants; (3) as a source and/or medium of exchange for supplying plants with nutrients; (4) and as a reservoir for moisture upon which

[1]A general term for the entire layer or mantle of fragmental and loose, incoherent, or unconsolidated rock material, of whatever origin (residual or transported) and of a much varied character, that nearly everywhere forms the surface of the land and overlies or covers the more coherent bedrock. It includes rock debris (weathered in place) of all kinds, volcanic ash, glacial drift, alluvium, loess and eolian deposits, vegetal accumulations, and soils.

growing plants can draw. Soil also must be capable of allowing excessive moisture to pass through its pores and drain to a lower level so that the soil will not remain excessively wet. Properties such as permeability and strength are not only of importance to the agriculturist, but to civil engineers and construction people who excavate, drill through, and handle soil in connection with buildings, roads, tunnels, etc.

Soil is a subsystem that interacts as part of a four-element system: (1) The *climatic or environmental subsystem* prevails immediately above the ground level and thus is the microclimate for a particular location. The principal variables of this subsystem are temperature, humidity, precipitation, and solar radiation—all of which interact constantly with soil. (2) The characteristics and patterns of the *hydrologic subsystem* determine essentially how water reaches the soil, both from above and below, and how water is carried away or drained from the soil. See also **Hydrology.** (3) The *plant subsystem* reduces the nutrient and moisture content of the soil, depending upon the particular uptake characteristics of a given plant. The plant subsystem also contributes in a major way to hold the soil in place and to protect it from disintegration and destruction by water and wind erosion. The plant subsystem also distributes moisture over the top of the soil so that all porous paths of the soil can be used to transport water rather than overloading and hence enlarging only some of the pores. The plant subsystem also protects the top of the soil against drying into a hard crust during periods of drought. Once disturbed, the characteristics of soil are difficult to replicate—a problem which arises when large projects, such as strip-mining, remove vast amounts of soil. See **Revegetation.** (4) The *soil subsystem*, the properties of which are described briefly in this entry.

Lack of sufficient attention to the long-term protection of soil has caused innumerable problems and losses over the years. Although warnings of gross problems arising from soil destruction and land mismanagement were given in North America as early as the latter part of the 1600s, it required the rudest of awakenings to precipitate national interest and action in soil conservation. This came in the early 1930s in the form of the Great Dust Bowl, a national disaster that affected some 96 million acres (38.4 million hectares) of farmland in the southern part of the Great Plains region, involving parts of Kansas, Oklahoma, Texas, New Mexico, and Colorado. During just a few years of severe droughts, accompanied by frequent high winds, literally billions of tons of soil were removed. Organic matter, clay, and silt were lifted and carried for great distances. There were times when the heavily laden skies as far east as the Atlantic coast were darkened. Sand and silt dunes from 4 to 10 feet (1.2 to 3 meters) in height were formed in many locations. In some parts of the Dust Bowl, as much as 80% of the land suffered from wind erosion. Parallel situations have occurred in several other areas of the world.

The Soil Conservation Service of the U.S. Department of Agriculture was established in 1935. Concentrated and participative programs with land users in the Great Plains region from the mid-1950s to the present time have provided impressive improvements: (1) 2.4 million acres (1 million hectares) of permanent vegetative cover have been established; (2) 1.0 million acres (0.4 million hectares) of field and wind stripcropping have been introduced; (3) 169 thousand (68 thousand hectares) of grasslands have been reestablished; (4) 41 thousand acres (16.8 thousand hectares) of trees or shrubs have been placed as windbreaks; (5) 81 thousand miles (150 thousand kilometers) of terraces have been constructed; (6) 5.4 million acres (2.2 million hectares) of brush control have been provided; and (7) 9 thousand miles (16.7 thousand kilometers) of pipelines to provide water for livestock grazing lands have been installed. See also **Erosion (Geology).**

Soil Characteristics and Classification

A soil is a naturally occurring three-dimensional body with morphology and properties resulting from effects of climate, flora and fauna, parent rock materials, topography, and time. A soil occupies a portion of the land surface, is mappable and is composed of horizons that parallel the land surface. A vertical section downward through all the horizons of the soil is called a *soil profile*. See Fig. 1.

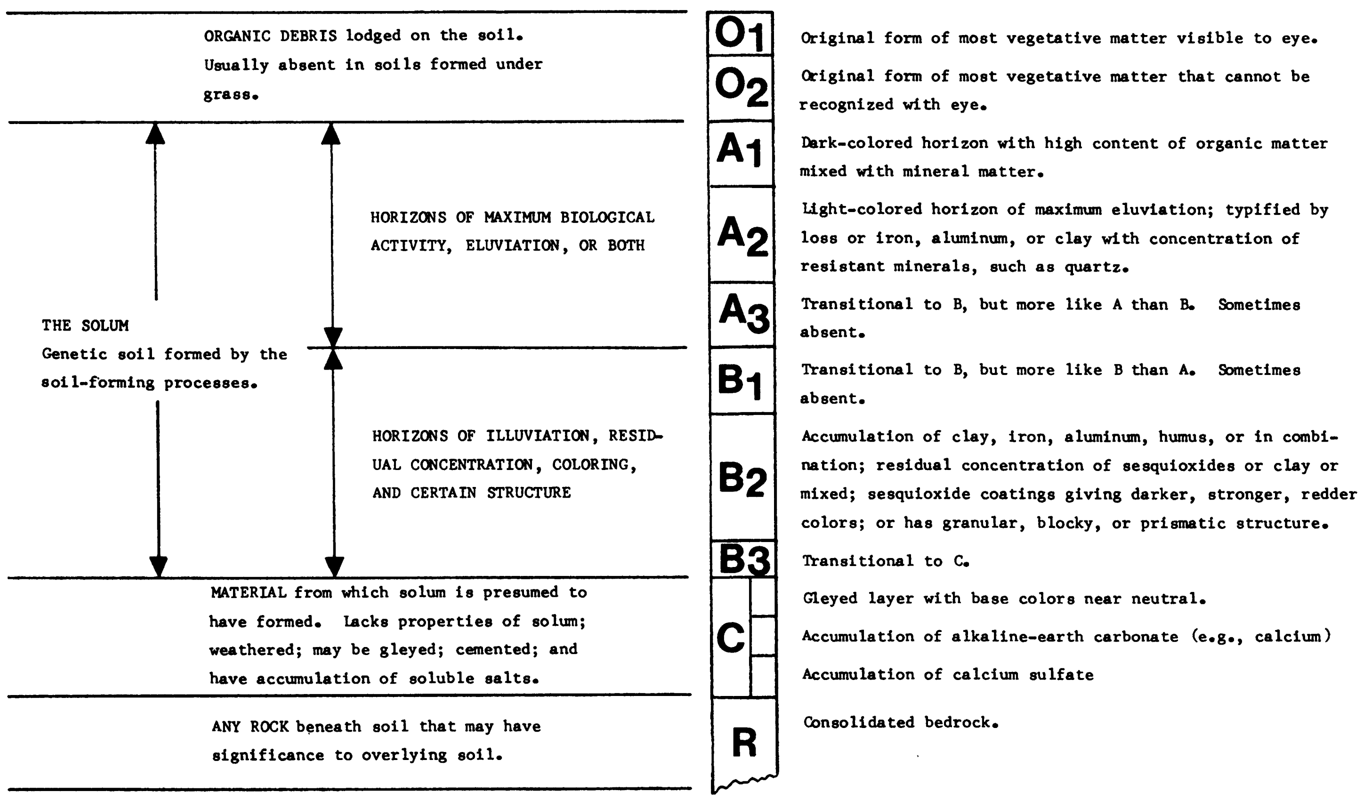

ORGANIC DEBRIS lodged on the soil. Usually absent in soils formed under grass.
THE SOLUM Genetic soil formed by the soil-forming processes.
HORIZONS OF MAXIMUM BIOLOGICAL ACTIVITY, ELUVIATION, OR BOTH
HORIZONS OF ILLUVIATION, RESIDUAL CONCENTRATION, COLORING, AND CERTAIN STRUCTURE
MATERIAL from which solum is presumed to have formed. Lacks properties of solum; weathered; may be gleyed; cemented; and have accumulation of soluble salts.
ANY ROCK beneath soil that may have significance to overlying soil.
O1
O2
A1
A2
A3
B1
B2
B3
C
R
Original form of most vegetative matter visible to eye.
Original form of most vegetative matter that cannot be recognized with eye.
Dark-colored horizon with high content of organic matter mixed with mineral matter.
Light-colored horizon of maximum eluviation; typified by loss or iron, aluminum, or clay with concentration of resistant minerals, such as quartz.
Transitional to B, but more like A than B. Sometimes absent.
Transitional to B, but more like B than A. Sometimes absent.
Accumulation of clay, iron, aluminum, humus, or in combination; residual concentration of sesquioxides or clay or mixed; sesquioxide coatings giving darker, stronger, redder colors; or has granular, blocky, or prismatic structure.
Transitional to C.
Gleyed layer with base colors near neutral.
Accumulation of alkaline-earth carbonate (e.g., calcium)
Accumulation of calcium sulfate
Consolidated bedrock.

Fig. 1. Hypothetical soil profile that has all principal horizons. Not all horizons shown are present in any given profile, but every profile has some of them. Terms used in diagram: *Eluviation* is the downward movement of soluble or suspended material in a soil from the A horizon to the B horizon by groundwater percolation. The term refers especially, but not exclusively, to the movement of colloids, whereas the term *leaching* refers to the complete removal of soluble materials. *Illuviation* is the accumulation of soluble or suspended material in a lower soil horizon that was transported from an upper horizon by the process of eluviation. *Gleying* is soil mottling, caused by partial oxidation and reduction of its constituent ferric iron compounds, due to conditions of intermittent water saturation. Process is also called *gleization*. (*Adapted from USDA diagram.*)

SOIL CLASSIFICATION SYSTEM (U.S.D.A.)

Order and Suborders	Definitions and Properties
Entisols	Weakly developed soils on freshly exposed rock or recent alluvium without genetic horizons. While alluvium may be rich in plant nutrients, entisols often are too shallow, too wet, or too dry for agricultural purposes.
E1	*Aquents*. Seasonally or perennially wet.
E2	*Orthents*. Loam or clay texture, often shallow to bedrock.
E3	*Psamments*. Sand or loamy sand texture.
Vertisols	Clay soils that have deep wide cracks during periods of moisture deficiency. During rainfall, vertisols swell, slide and produce warping.
V1	*Uderts*. Usually moist, with cracks open less than 90 days/year.
V2	*Usterts*. Dry and cracked more than 90 days/year.
Inceptisols	Soils that are beginning to show development of genetic horizons. Inceptisols lack evidence of weathering and usually are found in humid climates where leaching is active.
I1	*Andepts*. Soils containing amorphous or allophanic clay, often associated with volcanic ash and/or pumice.
I2	*Aquepts*. Seasonally or perennially wet.
I3	*Ochrepts*. Soil with thin, light colored surface horizons.
14	*Tropepts*. Continuously warm or hot.
15	*Umbrepts*. Dark surface horizons; medium to low base supply.
Aridisols	Soils which contain little organic matter or nitrogen. They are usually dry for more than 6 months/year. In numerous areas, salts accumulate on or near the soil surface. Since the nutrient content, except nitrogen, of ardisols is often high, these soils can be productive with irrigation and nitrogen application. Salt accumulation can be a problem with some crops.
D1	*Undifferentiated aridisols.*
D2	*Argids*. Soils with horizons of clay accumulation.
Mollisols	Soils with dark, thick, organic-rich surface horizon, high base supply. Mollisols are highly fertile and can support a variety of crops.
M1	*Albolls*. Soils with seasonally high water tables.
M2	*Borolls*. Cool or cold soils.
M3	*Rendolls*. Soils with subsurface accumulations of calcium carbonate, but no clay.
M4	*Udolls*. Temperate or warm, usually moist.
M5	*Ustolls*. Temperate or hot. Dry more than 90 days/year.
M6	*Xerolls*. Cool to warm. Moist in winter and continously dry more than 60 days/year.
Spodosols	Soils found primarily in cool and humid forested regions. Spodosols have subsurface accumulations of amorphous materials, mainly iron and aluminum oxides. These soils are usually strongly leached, but can be used for crop support with addition of lime and fertilizer.
S1	*Undifferentiated spodosols.*
S2	*Aquods*. Seasonally wet.
S3	*Humods*. Soils with subsurface accumulations of organic matter.
S4	*Orthods*. Soils with subsurface accumulations of organic matter, iron, and aluminum.
Alfisols	Soils of middle latitudes and degraded grasslands soils. Alfisols are strongly weathered, with gray to brown surface horizons, a subsurface clay accumulation, and a medium-to-high base supply. With adequate lime and fertilizer, the alfisols will continue to produce a variety of crops.
A1	*Boralfs*. Cool soils.
A2	*Udalfs*. Temperate to hot. Usually moist.
A3	*Ustalfs*. Temperate to hot. Dry more than 90 days/year.
A4	*Xeralfs*. Temperate to warm. Moist in winter and continuously dry more than 60 days in summer.
Ultisols	Strongly weathered soils of the middle and low latitudes. Ultisols are usually moist and low in organic matter. These soils have experienced a high degree of mineral alteration and extensive leaching. With fertilizer additions and good management, ultisols can support crops.
U1	*Aqults*. Seasonally wet.
U2	*Humults*. Temperate or warm. Moist all year. High content of organic matter.
U3	*Udults*. Temperate to hot. Usually moist.
U4	*Ustults*. Warm or hot. Dry more than 90 days/year.
Oxisols	The predominant soils of the Tropics. Oxisols have experienced the greatest degree of mineral alteration and horizon development of any soil. The humus breakdown is rapid and the soils are usually deep and porous. Oxisols require fertilization to support continued crop production.
O1	*Orthox*. Hot and nearly always moist.
O2	*Ustox*. Warm or hot. Dry for long periods, but moist for at least 90 days/year.
Histosols	Bog or peat soils composed primarily of vegetative debris in various stages of decomposition.
Mountain	Soils with various temperature and moisture parameters. Altitude, aspect, steepness of slope, and relief cause these soils to vary greatly within short distances. In many places, soil will be entirely absent.
Soil-absent	Rugged mountains and icefields.

NOTE: Further details can be obtained from "Soil Classification, A Comprehensive System, 7th Approximation," Soil Conservation Service, U.S. Department of Agriculture, Washington, D.C. (Published periodically).

Characterization of a soil requires selection of a representative profile that is described as quantitatively as possible, utilizing comparative charts for color, structure, and other properties, and accurately measuring soil horizons. Soils are collected from horizons and analyzed for particle-size distribution, pH, organic carbon, nitrogen, free iron oxide, calcium carbonate equivalent, moisture tension, cation-exchange capacity, extractable cations (calcium, magnesium, hydrogen, sodium, potassium), base saturation, and bulk density, among other factors.

Soil classification has been oriented to soil properties in recent years, but still is tempered with concepts of soil genesis, with external associations, and with the use of the soil. The first systematic classification was by Dokuchaiev in Russia in 1882. Based upon field and laboratory characteristics, soils were grouped into three categories—*normal soils* of the dry-land vegetative zones and moors, *transitional soils* of washed or dry land sediments; and *abnormal soils*. The system involved properties of the soil with external associations of climate and vegetation. Later, an associate (Sibirtsev) renamed the highest classes *zonal*, *intrazonal*, and *azonal*.

A traditional classification of soils includes three categories: (1) *Young soils*. These usually show their relationship to the parent material and are typical flood plain and hilly land deposits, when the soil surfaces are constantly being replenished or disturbed. (2) *Mature Soils*. These usually cover relatively flat lands where there are good drainage conditions but relatively little erosion. The development of these soils has gone so far in some cases, particularly in semi-arid regions, that little relation is shown to the parent material and their nature has therefore been principally determined by climatic and organic factors. (3) *Old Soils*. These usually cover old flat surfaces which have not been disturbed by erosion or sedimentation for a long time. Such soils, due to the dominance of climatic factors in their formation, have lost many of their original characteristics and have, therefore, developed abnormal features. When soils are intensively cultivated their mineral and organic constituents are rapidly depleted and must be replenished by rotation of crops and the application of natural fertilizers. The method of allowing the land to remain fallow is now known to be inefficient. The complete removal of vegetable cover, such as may result from overgrazing, deforestation, or dry farming, exposes the soil to rapid erosion and destruction.

A more scientific classification of soils, adapted by the U.S. Department of Agriculture, is given in the accompanying table. Systematic classification of soils in the United States began with the work of Coffey in 1912 and resulted in the first comprehensive system by Marbut in 1936. Considering the size of the earth and the large number of soils represented, the detailed cataloging of soils for any country is a tremendou task. To simplify the task to some extent and to make findings more meaningful from a practical viewpoint, the European Commission on

Agriculture (Working Party on Soil Classification and Survey) in 1966 correlated types of soils (soil units) with several regions designated by geography, geology, and climatology, and, to some extent, by the traditional use of the soils. Mixed criteria enter into soil classification schemes simply because the various physical or chemical characteristics, considered separately or together, do not fully identify a soil. The principal categories adopted by the European Commission include: Lowlands, Mountainous Areas and Highlands, Volcanic Areas, Zones of the Tundras and Fields, Zones of the Boreal Forests of Conifers and Birch, Zones of Mixed Forests of Conifers and Broadleaved Trees, Zones of the Central European Beech Forests and Oak Forests, Zones of the Oak Forests and of the Atlantic Heaths, Zones of the Continental Oak Forest, Zones of the European Grassland, Zones of the Mediterranean Sclerophyll Forests, Zones of the Juniper Forests and the Mediterranean Steppes, Zones of the Montane Mediterranean Forests, Zones of the Subalpine Mediterranean Forests, and Zones of the Arabo-Caspian Steppes.

Soil Genesis. The origin and processes of soil formation usually are inferred by relating measured morphological, physical, and chemical properties of a part to other parts of a given soil. And, during the last several decades, laboratory experimentation has revealed a better understanding of many of these processes. A factor to be stressed is that, in general, these processes occur over very long periods of time and frequently under multivariate conditions—conditions that are extremely difficult to duplicate and speed up in the laboratory. In the late 1950s, an interesting group of experiments (Thorpe et al., 1957) revealed information concerning the formation of something similar to podzolic soil. Organic and distilled water leachates from tree leaves were passed through columns of different soil materials. Bleached surface layers and subjacent layers of stronger color formed in the columns. Effluent solutions from the base of the columns contained detectable amounts of calcium, magnesium, iron, manganese, phosphorus, potassium, and sodium. Very fine silicate clays, e.g., illite, montomorillonite, vermiculite, and chlorite, also were suspended in the effluent. Removal, transfer, and transformation were demonstrable experimentally. Examination of the columns showed that clay was partially removed from the bleached layers and was deposited in voids in the lower layers. The experiments showed removal (*eluviation*) and addition (*illuviation*) actually occurring and at a much accelerated pace.

Organic matter probably best illustrates additions to a soil and is formed in the biological decomposition of plant and animal residues by soil microorganisms. Plants supply most of the organic matter as a dry material added to the soil surface and as roots in the subsurface. It has been estimated that short grass prairie in semiarid regions may annually add 0.7 ton/acre (1.6 metric tons/hectare) of dry matter; tall grass prairie in subhumid regions, 0.8 to 1.7 tons/acre (1.8 to 3.8 tons/hectare); pine forest in more humid areas, 2.1 tons/acre (4.7 metric tons/hectare); and tropical rain forest, from 45 to 90 tons/acre (101 to 202 metric tons/hectare). Under bluegrass roots, additions may amount to 2.4 tons/ acre (5.4 metric tons/hectare) in the top 4 inches (10 centimeters) of soil.

During decomposition, plant materials are converted to carbon dioxide, water, mineral elements, and other chemically altered substances. Less-resistant materials are consumed first by soil microbes—so that more resistant plant materials remain with the new organic compounds that are synthesized by the organisms. At any time, the organic matter at a place in the soil reflects an equilibrium state of the addition of new material to the system, removal of more readily decomposable materials, and transformation to other forms by microorganisms and other agents. Organic matter also may be transferred within the soil by physical and physicochemical processes. Burrowing animals, worms, and insects turn over the soil and physically mix adjacent portions. Freezing and thawing and wetting and drying also assist in the process. Colloidal organic matter may be flushed downward or laterally and coagulate as coatings on structural aggregates in the soil.

The more unstable organic compounds are rapidly oxidized to carbon dioxide and water by various biochemical processes, while the more stable fractions accumulate. Conjugated ring compounds containing carbon, hydrogen, oxygen, nitrogen, phosphorus, and sulfur and other elements in small quantities accumulate in relatively stable organic and organomineral colloidal complexes. Ligninlike, phytinlike, and nucleoproteinlike compounds are included. Sorption of the organic matter on mineral colloid surfaces, particularly layer silicates, such as montmorillonite, helps to stabilize the organic matter against biochemical oxidation. In tropical soils, high stability of soil organic matter is imparted by coatings of aluminum hydroxide and red ferric oxide. Organic and iron oxide colloids, when fairly abundant, stabilize the soil into porous aggregates through which ample air and water can circulate.

Localized spots of decomposing organic matter are important in reducing small but important quantities of iron to ferrous form and manganese to divalent form so that they become available to plants. Moderately to highly alkaline soils sometimes have inadequate activity of the reduced forms of iron and manganese, particularly in the absence of sufficient organic matter.

Radiocarbon dates of organic matter from the surface horizons of soils not only reflect the equilibrium status, but point out the turnover of the system. One example of research, for example, has shown that in the Edina soil in southern Iowa, organic matter is 410 ± 110 years old in the 6-inch (15-centimeter) top layer. In the next subjacent layer, the age is 840 ± 220 years. At depths of 23 to 25 inches (58.4 to 63.5 centimeters), the organic carbon is 1545 ± 110 years old. The entire soil has been estimated as 14,000 years old.

The four kinds of changes that develop soil horizons are dependent upon many basic processes, such as hydration, oxidation, reduction, solution, precipitation, freezing, thawing, wetting, drying, among others. These processes, in turn, are dependent upon the four fundamental factors of soil formation: (1) nature of the parent material; (2) topography; (3) climate; and (4) biological activity that occurs in the upper strata of the soil. To these factors must be added time and imposed manual and mechanical manipulation (as by tilling, planting, etc.) and chemical manipulation (as by fertilizing and use of various control chemicals that seep into the soil).

Soil is destroyed by two principal processes—*water erosion* and *wind erosion*. The word *erosion* is the physical removal of all or part of established soil by washing or blowing away. Erosion, in some instances, also brings soil to convenient locations, but usually in so doing, unless carried out over long periods as in the development of bottomlands and deltas where crops can be grown, the new muddy, fine, highly unconsolidated and disintegrated soil causes more problems than immediate benefits. The bringing in or transfer of soil by water is commonly referred to as *sedimentation*. Much research has been carried out

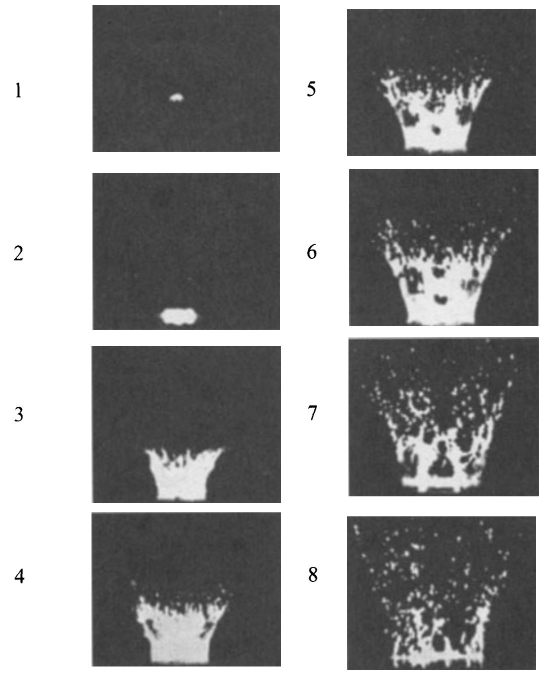


Fig. 2. Interaction of raindrops with the soil surface is an important component of the erosion process. Frames shown here were made by a 16-millimeter movie camera capable of speeds from 150 to 8000 frames per second. In the experiment, water drops are released from a tower 40 feet (12 meters) high and strike a plate glass target table. Drops from 5 to 6 millimeters in diameter are produced. Target plate is covered with water approximately 0.5 millimeter deep. (*North Central Soil Conservation Research Center, U.S. Department of Agriculture, Morris, Minnesota.*)

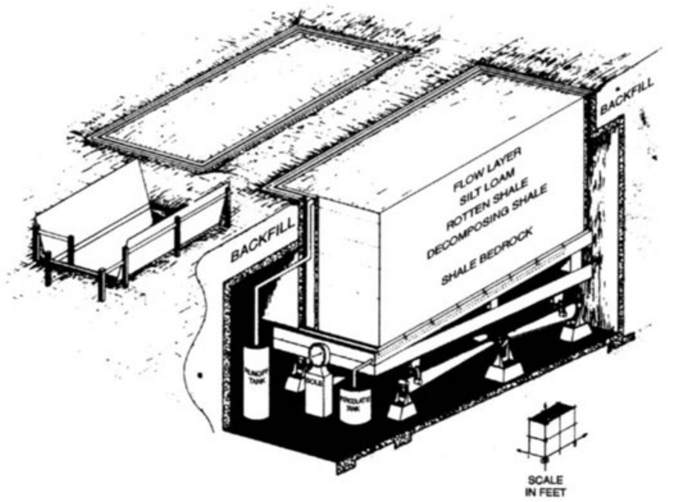


Fig. 3. A lysimeter which provides a means for isolating soil masses and recording weight changes and water percolation. Such instruments provide accurate assessments of moisture behavior in soil. The lysimeter shown here represents $\frac{1}{500}$ acre (0.0008 hectare) and is 8 feet (2.4 meters) deep. The soil weighs 65 tons (58.5 metric tons), yet can be weighed to a precision of 5 pounds (2.3 kilograms). Soil scientists use lysimeters to study evapotranspiration, moisture consumption by crops, precipitation, water movement, and pollution. (*USDA diagram.*)

in connection with water and wind erosion. Typical of fundamental research is the study of splash patterns as shown in Fig. 2. The lysimeter, as shown in Fig. 3, also has been effectively used. The effects of wind erosion have been extensively studied, as exemplified by Figs. 4 and 5.

Additional Reading

Bowies, J. E.: "Foundation Analysis and Design," 4th Edition, McGraw-Hill, New York, 1988.
Carroll, R. C., and J. H. Vandermeer: "Agroecology," McGraw-Hill, New York, 1990.
FAO: "Soil Maps of the World," Food and Agriculture Organization of the United Nations, Rome, Italy. (Revised periodically.)
Franklin, J. A., and M. B. Dusseaultz: "Rock Engineering Applications," McGraw-Hill, New York, 1991.
Goldman, S., et al., Editors: "Erosion and Sediment Control Handbook," McGraw-Hill, New York, 1986.
Hausmann, M. R.: "Engineering Principles of Ground Modification," McGraw-Hill, New York, 1990.
Lal, R.: "Soil Erosion in the Tropics," McGraw-Hill, New York, 1990.
Larson, W. E., et al.: "The Threat of Soil Erosion to Long-Term Crop Production," *Science*, **219**, 458–465 (1983).
Levy, R.: "Chemistry of Irrigated Soils," Van Nostrand Reinhold, New York, 1984.
Rendig, V. V., and H. M. Taylor: "Principles of Soil-Plant Interrelationships," McGraw-Hill, New York, 1989.
Schroth, M. N., and J. G. Hancock: "Disease-Suppressive Soil and Root-Colonizing Bacteria," *Science*, **216**, 1376–1381 (1982).
Tucker, C. J., et al.: "African Land-Cover Classification Using Satellite Data," *Science*, **227**, 369–374 (1985).
USDA: "Soil Conservation Reports," National Soil Survey Laboratory, U.S. Department of Agriculture, Washington, D.C. (Published periodically.)

SOIL BEARING VALUE. See **Foundations.**

SOIL PERMEABILITY. See **Hydrology.**

SOL. A word sometimes used to describe the solar day on the planets and their satellites. A sol or solar day is the interval between two suc-

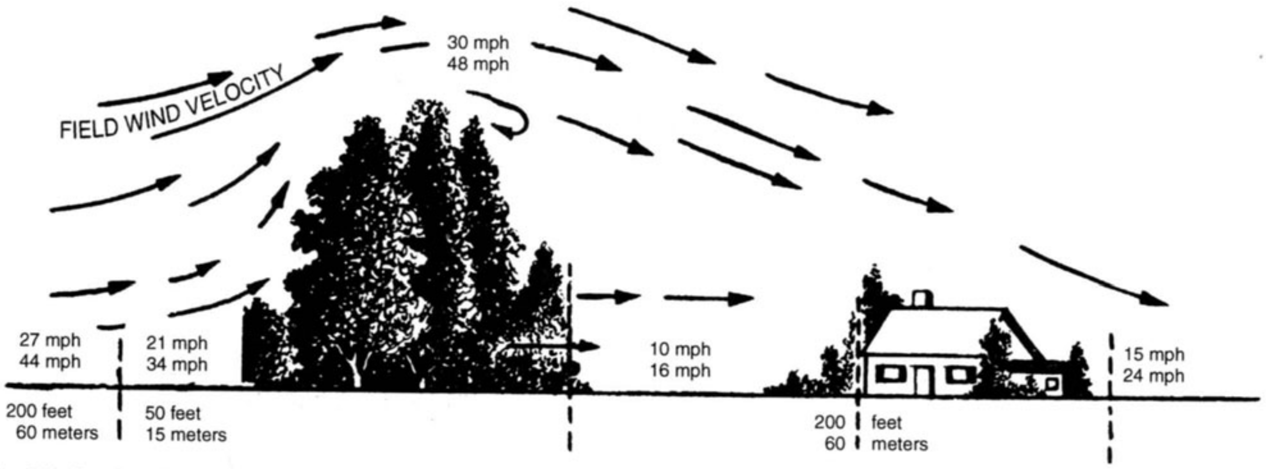


Fig. 4. Windbreaks reduce wind currents. Part of the air current is diverted over the top of the trees and part of it filters through the trees. Breaks like this reduce wind erosion of soil. Farmstead, livestock, and wildlife windbreaks should be relatively dense and wide to permit maximum protection close to the trees. Field, orchard, and garden-type windbreaks need not be so wide and dense. (*USDA diagram.*)

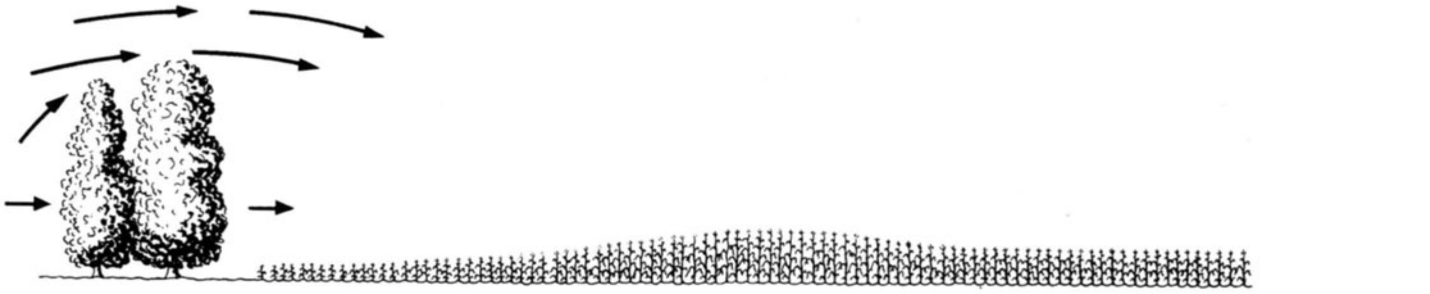


Fig. 5. Sometimes crop yields are lowest next to a tree windbreak. A common error made by some growers is to observe only the immediately adjacent area. The greatest gains are from a distance of 30 to 45 feet (9 to 13.5 meters) from the windbreak. (*USDA diagram.*)

cessive transits of the sun past a given meridian. The length of the solar day ranges widely from one planetary body to the next.

SOLANACEAE (Potato Family). A relatively small plant family of some 1,700 species, in about 80 genera. The members include herbs, shrubs, and a few trees, most of the last being found in the tropics. The family has many important food and medicinal plants. All of its members contain poisonous alkaloids, as solanin, capsaicin, nicotine, and atropine.

Members of this family have leaves which are usually alternate and variously shaped, often lobed or dissected or pinnately compound. The variability frequently occurs in the leaves of a single plant. The flowers occur either singly or in cymes, and are regular with an inferior five-lobed calyx, a five-lobed corolla of various shapes; often large and conspicuously colored, five stamens and a pistil composed of two united carpels, a long style and a single terminal stigma. The ovary contains many ovules. The fruit is either a berry (potato and tomato) or a capsule (tobacco).

The most important member of this family is the Potato, *Solanum tuberosum*, commonly called the Irish potato, common potato, White Potato, or English potato, to distinguish it from the sweet potato. The plant is a branched herb, the branches tending to spread out more or less, and growing 2–4 feet in height. The green stems are annual, but the tubers, which are modified stems, give to the plant a perennial nature. The leaves are pinnately compound and rather irregular. The flowers are $1–1\frac{1}{2}$ inches in diameter, white, often with blue or purple tones or stripes, and with a tubular corolla. The fruit is a globular berry containing many small seeds embedded in a green pulp. Common names given these potato berries are potato balls, potato apples or seed balls. When mature, they are either green or brown; often they contain few seeds, but since they are rarely used in propagating the potato, this is of little consequence. They are used experimentally in the production of new varieties. Once a desirable variety is found, it is perpetuated vegetatively by means of the tubers. These tubers are formed at the tips of modified underground branches, or rhizomes, which radiate outward from the basal portion of the stem and to the casual observer resemble roots. The length of these stems varies from a few inches to a foot or more, depending somewhat on the variety. Rhizome formation and tuber development start soon after the tops appear above ground, and are advanced by darkness and low temperatures. The tubers continue to grow throughout the growing life of the plant. A mature tuber has the structure of a stem, but very much modified. Externally, one may recognize nodes and internodes, the nodes being determined by the eyes, depressions in the surface of the potato, each depression or eye containing a tiny bud. Internally the tuber is largely parenchymatous tissue with the cells filled with starch-grains. Near the surface of the potato, cross sections show a faint dark line which is the vascular tissue, very much reduced. Potato tubers vary greatly in shape, as well as in size; the better varieties are oblong or oval, with smooth skin and rather shallow eyes. In color, tubers range from brown through yellow to red, according to variety. The length of time required to mature a marketable tuber also varies greatly, some early varieties reaching a desirable size in about two months, while late varieties require five or six months.

The tubers are the principal means of propagation. For this purpose a tuber is cut into irregular pieces, each of which contains two or three eyes. To prevent disease the cut pieces are treated with fungicides and allowed to dry slightly, after which they are planted. Sprouting starts at once, indeed frequently occurs while the potato is still in the storage bin. If the bin is dark the sprouts formed will be long, slender and white. Sprouts formed in light are short, stout and dark green.

The potato is a native plant of cool upland regions of South America, where it has been long cultivated by the natives, who use the tubers for food. From America it was carried to the southern part of Europe, where at first it was grown largely as a curiosity. From Europe the plant was brought back to America and introduced into what is now the United States, thus explaining its name of English or Irish Potato. It is now extensively grown in regions having a cool climate, Maine and Idaho being particularly noted for their potato crops.

In all green parts of the potato a poisonous alkaloid, solanin, occurs. This substance may also occur in the tubers, particularly if the latter are exposed to light long enough to become green in color.

Another important food plant of this family is the tomato, *Lycopersicum esculentum*, and related species. This is a native of South America, in which continent the tomato is still found wild. The tomato is a coarse branching perennial herb which is usually grown as an annual. It is an important food which has become immensely popular. Large quantities are consumed raw or canned. Tomato juice is important in many diets. Greater greenhouse production and improved rapid transportation from southern states makes it possible to enjoy fresh tomatoes in any season.

The tomato was carried to Europe by the early Spanish explorers and became an important food plant in southern European countries. In England, and also in the United States, it was widely grown before 1830 as a garden ornamental known as "Love Apple," which however was not eaten, but was held by many to be rankly poisonous. The Italians seem to be the first to have braved the danger of tasting the delicious fruit of this "poisonous" plant.

A third member of the family providing food to people is the egg plant, *Solanum melongena*, a coarse, somewhat woody, branching herb native to India. The plant has a rought stem, 2 or 3 feet tall, large sinuate-lobed ovate leaves and purplish flowers. See Fig. 1. The fruit is a berry, very large in some varieties. It is eaten either baked or sliced and fried.

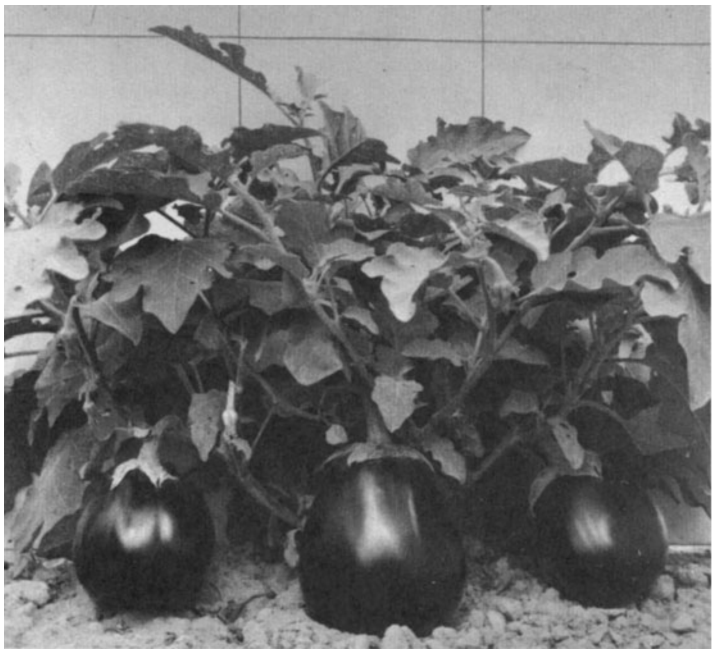


Fig. 1. Large-fruited *Black Beauty* variety eggplant, the most popular of eggplant varieties in the United States. (*USDA photo.*)

Serving rather as a condiment than a food, but still used as a food stuff in some of its varieties, is the pepper, *Capsicum annuum*, another native plant of tropical America. Peppers are either annual or biennial plants, with branching stems 1–3 feet tall, smooth shining leaves and white flowers. There are many varieties, which bear fruits of a variety of sizes and shapes (Fig. 2), as well as degrees of pungency. Some varieties are known as sweet peppers, and are used while yet green in salads or stuffed and baked; others are hot peppers. The pungent taste is due to an acrid compound, capsaicin, which in hot peppers occurs throughout the fruit, but in sweet peppers is largely restricted to the immediate region of the seeds; since only the fleshy pericarp is eaten, the seed being removed, this pungent substance is lost.

Peppers are used in many ways other than as food. Small hot peppers are a frequent component of mixed pickles, and also are used in salads. The whole fruit of some varieties is ground up to a powder, and becomes Cayenne pepper, an extremely pungent condiment. Small smooth fruits of var. *conoides* are preserved whole in brine or vinegar,

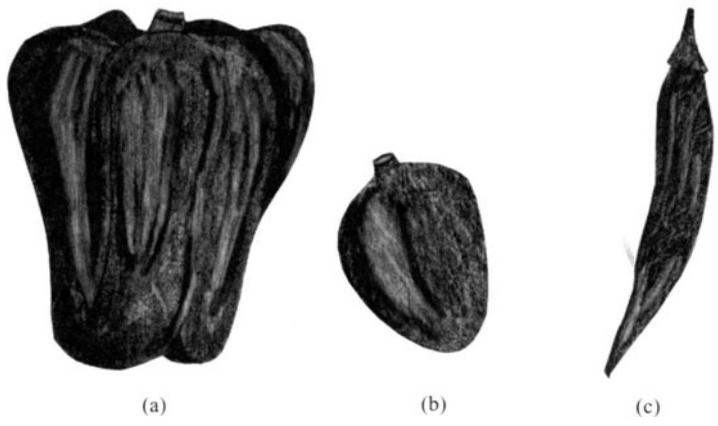


Fig. 2. Comparative sizes and shapes of common peppers: (a) bell or bullnose pepper; (b) pimiento; (c) Cayenne or chili pepper.

and known as Tabasco sauce or Tabasco peppers. Red peppers are used medicinally. Many varieties of pepper are grown as ornamental plants, the brilliant fruit offering a startling contrast to the dark green leaves.

More widely grown for its ornamental properties is the *Petunia*, introduced from South America.

Many members of the Potato Family are important drug plants. Here belong *Atropa belladonna*, yielding atropine, and *Datura stramonium*, the Thorn Apple, called also Jimson Weed or Jamestown Weed. The latter is a tall, rather coarse, branching annual having broad leaves with sinuate margins and large trumpet-shaped white flowers, borne singly in the axils of the leaves (a related species, *D. Tatula*, has purple flowers). The fruit is a prickly capsule. In all parts of the plant, but especially in the seed, are found several drugs. The most abundant is hyoscyamine, with small amounts of atropine and scopolamin present. The powdered leaves and seeds are used medicinally, chiefly in treating asthma.

Less important medicinally are several plants such as *Hyoscyamus niger, Solanum dulcamara*, the Bittersweet, and *Solanum niger*, the garden Nightshade, all held to be poisonous plants if eaten in sufficient quantity by man or domestic animals.

Another member of this family, ranking along with the potato in importance, is the tobacco plant, *Nicotiana tabacum*, also a native plant of tropical America. It is an annual plant growing 3–6 feet or more in height, and stout. The leaves are alternate, simple and rather large and, like the stem, covered with sticky hairs. The rather large flowers are borne in terminal racemes and have a funnel-shaped corolla which is yellow, white, pink or purple in different species and varieties. The fruit is a capsule containing many small seeds.

SOL AND SOLATION. See **Colloid System.**

SOLAR CELLS. See **Solar Energy.**

SOLAR ENERGY. The vast quantity of energy received by Earth from the sun and the potential for converting that energy into more useful forms for society has intrigued scientists, engineers, and social planners for decades. This interest was sharpened by the oil embargo of the 1970s and, for about a decade after that, tremendous interest was displayed by the scientific and lay community alike in alternative energy sources, including a turn to solar energy. Energy from the sun was considered by many people as a relatively low-cost and essentially pollution-free source, particularly in contrast with polluting, nonrenewable, so-called fossil fuels and with nuclear fuels, which many people consider in a negative light. During the 1970s, but tapering off in the 1980s, many, many millions of dollars were invested by governments worldwide and by private institutions, architectural and solar equipment firms, and energy supply firms toward the development of practical, economically competitive solar energy systems.

As a result of that activity, progress in designing passive solar energy systems into office and factory buildings has been impressive, but not nearly so extensive as once estimated. Active solar energy systems, in which solar radiation is converted into another energy form (usually electrical) has also progressed, but the number of outstandingly successful installations is relatively limited and essentially these are presently regarded as still in an experimental phase. In contrast, considerable research continues to be directed toward solar cells (solid-state devices that convert solar radiation into electric power), but it should be immediately stressed that solar cells for communication and other satellites and space vehicles are vitally needed because they provide an energy source difficult to obtain in other ways. Cost, in this instance, is not supercritical, but one finds that solar cells for building, etc. heating and power still are essentially noncompetitive. Some relatively low-cost, small solar-powered devices designed mainly for public consumption have appeared in recent years.

To put solar energy into perspective as the year 2000 nears, one should review other energy resources as well. Fortunately, throughout this encyclopedia, such energy information is available. See list of articles at end of this description and also consult alphabetical index.

Availability of Solar Energy

Not to be confused with insulation, the word *insolation* (acronym for "incoming solar radiation") defines the rate at which direct solar radiation is incident upon a unit horizontal surface at any point on or above the surface of the earth. The unit of insolation is the Langley, named after Samuel Pierpoint Langley (1834–1906), an American astronomer, physicist and pioneer in the utilization of solar energy:

$$\begin{aligned}1 \text{ langley} &= 1 \text{ gram calorie per square centimeter}\\ &= 3.687 \text{ Btu per square foot}\end{aligned}$$

Fundamental to the practical application of real-time solar energy systems is the amount of energy received from the sun at any given location at any particular time. The energy received varies with the geometry of the sun-earth system—and thus varies with latitude, season of the year, time of day, as well as with local weather conditions. On a typical day in June, anywhere from 500 to 700 langleys of solar energy can be expected in most parts of the United States, whereas in December, with shorter days and more inclement weather, only from 100 to 300 langleys can be expected. In June, for example, both Saskatoon, Canada and Tampa, Florida receive about 600 langleys of solar energy daily, but in December, the amount of such energy received in Saskatoon drops to 75 langleys per day (only 12% of the June value at that location), whereas, in Tampa, more than 50% of the June amount is received in December. In terms of equipment costs, this translates into needing four times the solar-energy collector surface in Saskatoon as that required in Tampa in order to supply the same amount of power year-round. Maps of the type shown in Fig. 1 can be helpful in this regard.

The solar constant (intensity of solar radiation outside the earth's atmosphere at the mean distance between the earth and the sun) has been determined by measurements from satellites and high-altitude aircraft and is 1.353 kilowatts per square meter. This extraterrestrial radiation, which corresponds closely to that of a blackbody at 5762°K, is 7% in the ultraviolet range (wavelength less than 0.39 micrometer) and 47% in the visible range (wavelengths from 0.38 to 0.78 micrometer), with the balance in the near-infrared (largely with wavelengths of less than 3 micrometers). Radiation is depleted as it passes through the atmosphere by a combination of scattering and absorption; the radiation that reaches the ground—the raw material of this energy source—can vary from almost none under heavy cloud cover to 85–90% of the solar constant under *very clear* skies.

Solar Energy for Building and Residence Comfort

The application of solar energy for residences and commercial and public buildings tends to fall into three categories of increasing complexity: (1) heating only; (2) cooling only; and (3) combined heating and cooling.

Simple Heating System with Hot Storage. As shown in Fig. 2, the basic elements of this system, exclusive of pumps, valves, and controls, are: (a) a solar collector; (b) an auxiliary heating device; (c) hot storage system; and (d) heater element fan and air duct system. The solar collector absorbs heat energy from the sun and transfers it to a heat-transfer fluid which conveys the heat to a hot storage sys-

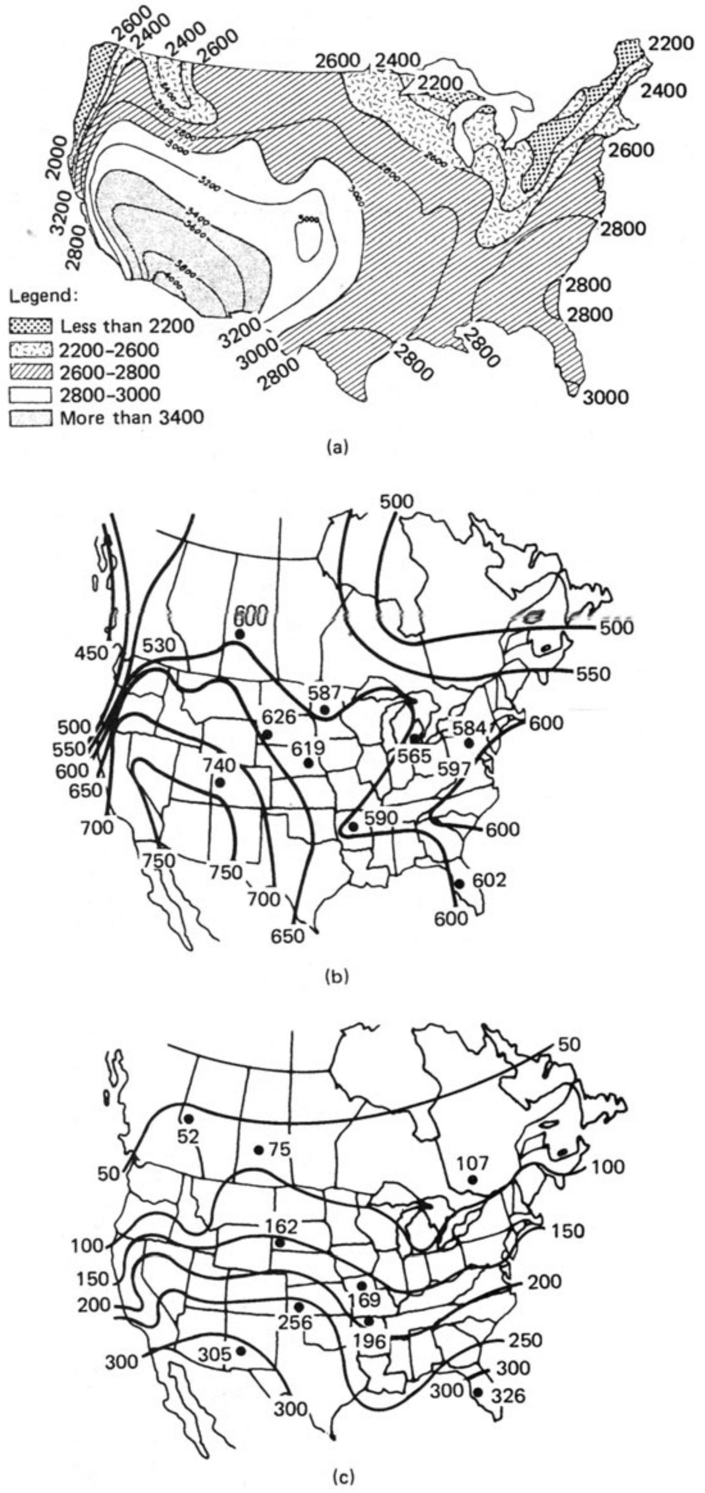


Fig. 1. Availability of solar energy (insolation): (a) Average number of hours of sunshine per year (United States); (b) median daily insolation in langleys (North America in June); (c) in December. (*National Oceanic and Atmospheric Administration.*)

tem. From the hot storage, heat is withdrawn from storage through a heater coil, where an air-circulating fan carries heat from the coils into an air duct system. When the solar collector cannot provide an adequate amount of heat to maintain the hot storage at a minimum temperature, the auxiliary heating device, such as a fuel burner, electric resistance heating, or an electric driven heat pump, comes on. This auxiliary heat could be added to hot storage as shown in Fig. 2, or used to directly heat the room air. For the case of electric heating, heat addition to storage will provide the opportunity to limit auxiliary heat addition to nonpeak hours.

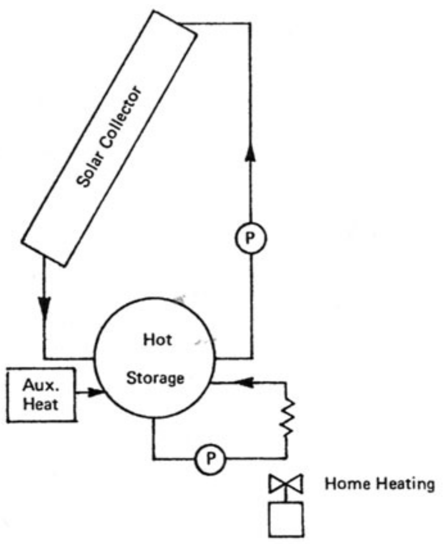


Fig. 2. Basic elements of solar heating system.

Solar Cooling System. As shown in Fig. 3, the basic elements of this system are: (a) a solar collector; (b) an auxiliary heating device; (c) a cold storage system; and (d) a heat-actuated refrigeration loop (absorption cycle). The solar collector absorbs heat energy and transfers it to a heat-transfer fluid, which in turn conveys the solar heat to the generator or boiler of the heat-actuated refrigeration loop. This loop can also be driven by auxiliary heating when the solar heat input is not adequate. If the heat-actuated refrigeration loop were of the Rankine-cycle type, rather than an absorption cycle, it might also be possible to drive the refrigeration loop with auxiliary power rather than auxiliary heating—a more favorable situation if the auxiliary is electric rather than fuel. The refrigeration loop cools the cold storage reservoir from which home cooling is supplied upon demand.

Combined Solar Heating and Cooling System. A system of this type is shown in Fig. 4. This is only one of a variety of possible systems. The major elements of this system are: (a) a solar collector; (b) an aux-

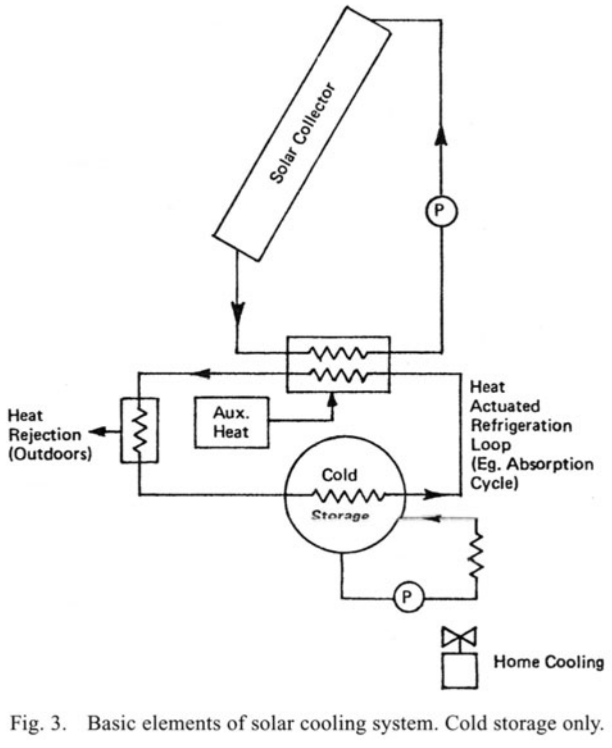

Solar Collector
P
Heat Rejection (Outdoors)
Aux. Heat
Heat Actuated Refrigeration Loop (Eg. Absorption Cycle)
Cold Storage
P
Home Cooling

Fig. 3. Basic elements of solar cooling system. Cold storage only.

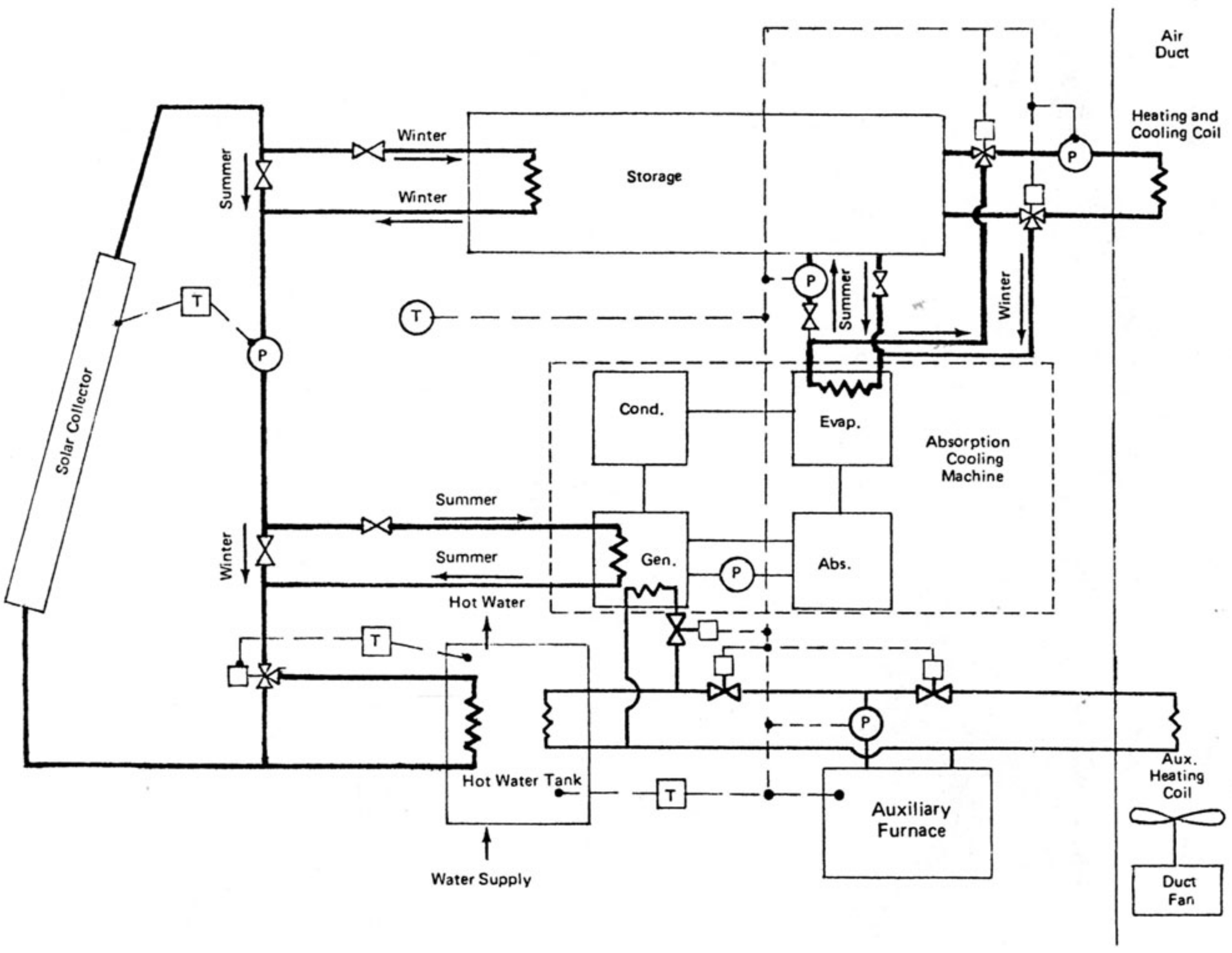


Fig. 4. Combined solar climate heating and cooling system.

iliary furnace with heating coils; (c) a storage system (hot in winter; cold in summer); (d) absorption refrigeration cycle; and (e) necessary valving and controls. The system is designed to provide both heating and cooling upon demand. The heat energy generated by the solar collector is directed to either the hot storage tank or the refrigeration cycle generator, according to the seasonal mode of operation. When solar energy (either direct or stored) cannot supply the required heating or cooling load, auxiliary heating or cooling can be used.

In the winter mode of operation, solar energy is gathered at the collector and is pumped directly to the storage system. From the storage system, heat is extracted according to household needs through the heating/cooling coil in the main air duct. This coil is controlled by 3-way valves which are open to heating and shut to cooling. Heat is then extracted from the coil by air fans and carried into the house. At certain times, the solar collector and storage system will not be able to provide enough heat to maintain the heat needs of the household. In these cases, the auxiliary furnace will assume the heating load until the solar system is able to provide heating. This furnace will also provide auxiliary heat for domestic hot water when solar energy cannot provide this function. At times, while on winter mode operation, there will be days when cooling may be required. When this condition exists, the auxiliary furnace will drive the absorption cycle. The two 3-way valves on the heating/cooling coil will be actuated to permit chilled water from the absorption machine to circulate through this coil.

In the summer mode of operation, solar energy is gathered at the collector and pumped in the form of heat directly to the absorption refrigeration machine and to the domestic hot water heater. The collected energy serves as the main driving force for the cooling system. When the collector is unable to provide the necessary energy for the cooling system, the auxiliary furnace is activated to supplement the energy load. Once the cooling cycle is activated, the cooling produced is directed to the same storage system used for storing heat in the winter. Cooling is extracted per household needs from the storage system through the heating/cooling coil. In this mode, the 3-way valves of the coil are open to storage system and shut to direct cooling from the cooling cycle. Should the storage system not be able to provide the cooling, these valves would be closed to the storage system and open to direct cooling. While on the summer mode, the auxiliary furnace can be used to heat the house on occasional cold nights.

The foregoing systems are representative of general concepts and not necessarily of final designs or optimum arrangements. The final detail system design will, in particular, be dependent upon whether fuel or electricity is used for auxiliary heating. For the near term, natural gas or fuel oil may be preferable to electric heating. However, in the longer term, as developments in solar collection and heat storage reduce the amount of auxiliary heat required, and as heat pump technology is improved, electricity may become more attractive.

Architectural and Building Factors. Solar climate control systems will have to be integrated with different building designs. New buildings can be designed to fit the requirement of a solar climate control system while applications to existing buildings will have to be determined individually. Collectors can be installed on flat roofs of buildings, or designed to fit the sloping roofs of a wide variety of buildings. Collectors could serve a single building or a cluster of buildings. As previously pointed out, the geometry of the collector installation varies with location latitude. Economics to a large degree will be determined by availability of reasonably sustained sunshine.

Flat-Plate Collectors. Essentially since the outset of solar energy technology for environmental heating and cooling of living and working enclosures, the flat-plate collector has dominated the field. Within recent years, however, some of the initial needs for collectors of this

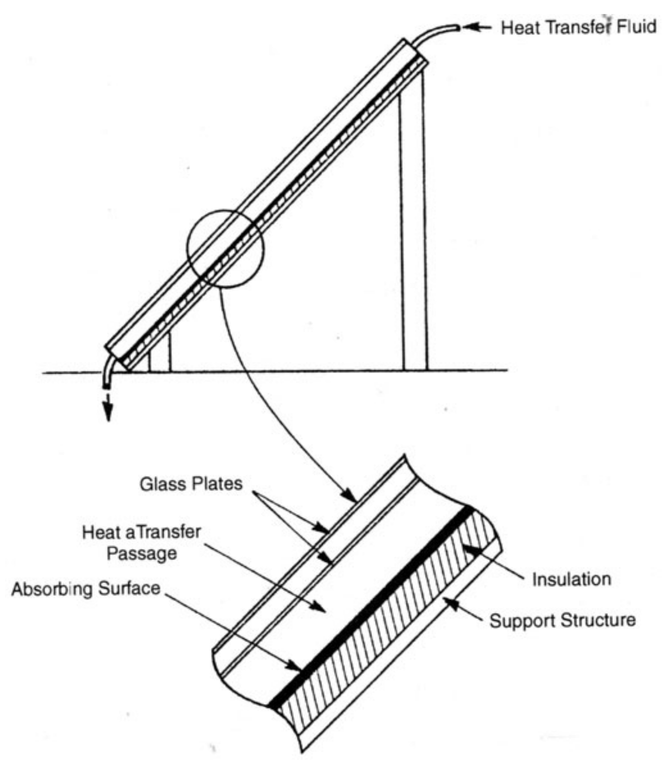


Fig. 5. Basic elements of a flat-plate solar collector.

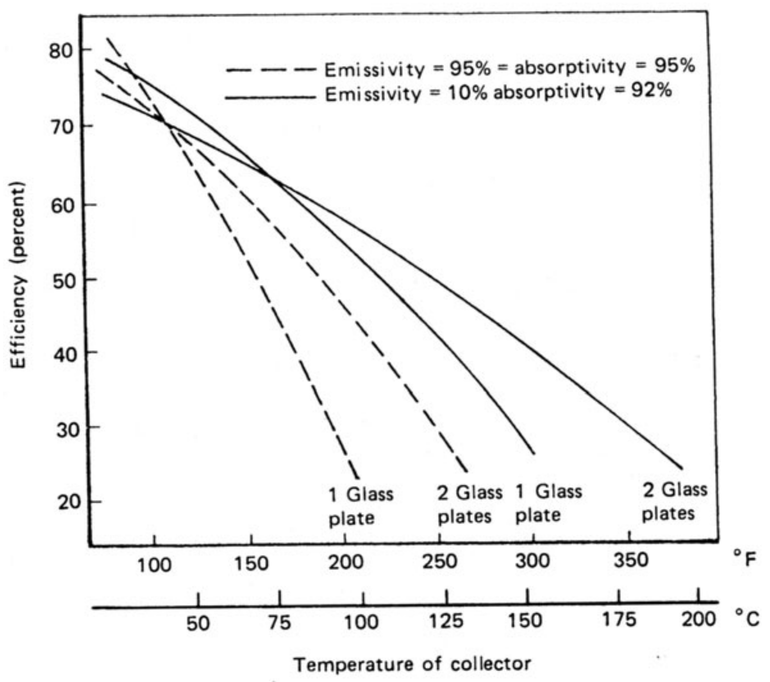


Fig. 6. Relative solar collection efficiency of plate-type solar collectors with different emissivities.

nature have been obviated by more attention being given to the design of passive solar collectors, described a bit later. Also, for very large commercial installations, there has been a trend toward the use of nonfocusing or trough or line-focusing concentrators, also described later.

The essential features of a flat-plate solar collector are shown in Fig. 5. A blackened receiver surface covered by one or more special glass plates is used. Since the glass is transparent to the incident solar radiation, but opaque to the reradiated energy, the solar collector, like a greenhouse, serves to trap solar energy. The working fluid used to remove the heat from the collector can be either air flowing between the blackened surface and the glass plate or water (or some other liquid) flowing in tubes attached to the blackened plate. Solar collection efficiency is defined as the ratio of usable energy collected per unit time to the incident solar flux. Efficiency, η, may be calculated as:

$$\eta = \alpha\tau - \frac{q_L}{q_{in}}$$

where: α is the absorptivity for sunlight; τ is the transmittance of the glass plate; q_L is the heat loss from the collector; and q_{in} is the incident solar flux. The heat loss q_L is, in turn, dependent upon the emissivity, ϵ, for low-temperature radiation. Typical performance characteristics for a solar energy collector are shown in Fig. 6 which is a plot of collector efficiency versus temperature of the absorber plate for an incident radiation of 300 Btu/(hour) (square feet); (814 kcal/(hour) (square meter). Note that the efficiency falls off as absorber temperature rises. Efficiency, of course, also drops off rapidly as the incident radiation is reduced, since the heat loss term is a function of absorber temperature only.

A problem that must be faced by architects and engineers is the need to integrate collectors into building and residence design in a way to maximize thermal performance and, at the same time, provide an esthetically satisfactory structure. A major variable, depending upon energy requirements and the average solar insolation over a year, is the amount of collector area needed. Obviously, the larger the energy requirements and the less favorable the insolation, the greater the problem. Because collectors must be oriented within rather narrow limits if they are to maximize their capture of solar radiation, the problem of retrofitting many existing structures sometimes renders a project impractical. Thus, the general pattern has been one of concentrating on new construction, particularly for low-rise, flat-roofed buildings, such as schools and shopping centers, where cooling may be more important than heating. Where surrounding area is available, as, for example, an adjacent parking area or facility, the collectors can be installed apart from the structure which is to be heated and/or cooled. Also, by turning to advanced collector designs which provide greater efficiency, even at greater initial cost, architects and builders can better cope with the problem of collector area required. Passive heating of buildings is also a possibility in many cases.

Evacuated-Tube Collectors. In this type of collector, an inner glass cylinder, blackened to absorb solar radiation, is enclosed within an outer protective cylinder. The space between the two cylinders is evacuated. The inner cylinder is usually coated with material that reduces energy loss through reradiation. Transfer of heat is accomplished by a fluid (air or liquid) that flows through the inner cylinder. These collectors are similar to flat-plate collectors in that they can use both direct and diffuse light. However, the evacuated-tube collectors operate better during the early and late parts of the day. The vacuum provides such excellent insulation that they are less affected by high winds and cold weather than the flat-plate collectors. The output of the evacuated-tube collectors is essentially independent of ambient temperature and their efficiency is generally 40–50%. Ordinarily these collectors operate at about 180°F (82°C) for space-heating applications and certain process heating uses. Equipped with reflectors, they can operate up to 240°F (116°C), which is sufficient to drive absorption air conditioners. Cylindrical evacuated tube collectors can absorb radiation coming from any direction (360° aperture). Usually, they are mounted in arrays with a spacing of about one cylinder diameter between tubes and with a reflective material behind them.

Large-scale collectors, with or without focusing (radiation concentrators) are described a bit later.

Heat Storage. A comparison of heat-storage capacity on a volumetric basis between various storage media shows that water can store 62.5 Btu per cubic foot per degree Fahrenheit (311.5 kcal per cubic meter per degree Celsius. Rocks, bricks and gravel can store about 36 Btu/cu. ft/degree F (179.4 kcal/cu. meter/degree C). In addition to fluid media and solid media, advantage can be taken of the latent heat of a phase transition. Some salts which melt in the desired temperature range can store about 60 Btu/cu. ft/degree F (299 kcal/cu. meter/degree C) as sensible heat and 9500 Btu/cu. ft./degree F (47,348 kcal/cu. meter/degree C) at the melting point as heat of fusion. However, these salts tend to undercool rather than crystallize during the cooling cycle. While undercooling can be prevented by use of nucleating agents, the fixed rate of crystal growth is very slow in most salt hydrates. Heat cannot be withdrawn more rapidly than it can be supplied by the growth of crystals. This is a serious problem which can only be partially overcome by the design of the storage container to provide a large heat-transfer area. Based upon these alternatives, an insulated tank of water to store heat is one of the most efficient solutions. Early work concentrated on $Na_2SO_4 \cdot 10H_2O$ which undergoes a phase transition when heated at 32°C (89.6°F). Because phase separation of this hydrate occurs on cycling, other chemical systems are being sought which can undergo thousands of cycles without loss of storage capacity.

Typically optimum storage capacity varies from 1 to 3 typical winter days' solar energy supply, depending upon the site, and for a medium-size residence or small commercial building would be in the range of 1000–2000 gallons (38–76 hectoliters). A section of a comparatively small solar-heated building is shown in Fig. 7.

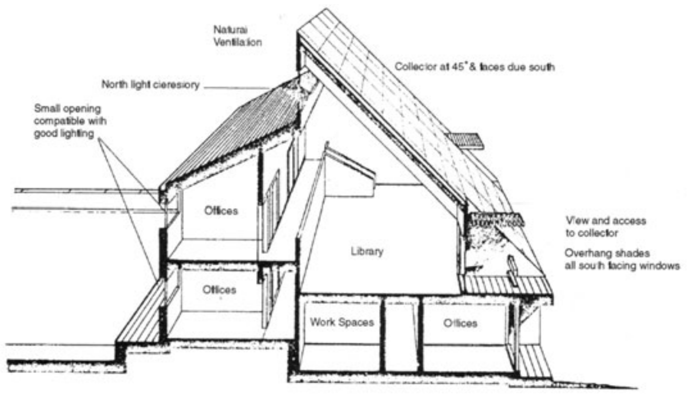

Natural Ventilation
North light clerestory
Small opening compatible with good lighting
Collector at 45° & faces due south
Offices
Offices
Library
Work Spaces
Offices
View and access to collector
Overhang shades all south facing windows

Fig. 7. Section of solar-heated building. Solar collector has an area of 3500 square feet (325 square meters) facing south at an angle of 45 degrees. There are about 8000 square feet (743 square meters) of working space. Estimates of heat loss indicate heat demand is in range of 40,000-70,000 Btu (10,080-17,640 kcal) per day. Located in the northeastern United States, the building was designed to furnish between 65 and 75% of total seasonal heating load.

Collectors also can be used as energy dissipaters by designing them to lose heat by convection and radiation to the clear night sky. The role of the collector is thus fully reversed for the cooling cycle. This requires a system for moving insulation, unless design compromises in collector design are made for both the heating and cooling cycles.

Solar collectors also can be combined with heat pumps. The latter can serve as an independent (auxiliary) source of heating energy, or the collector-storage system can serve as the energy source for the evaporator of the heat pump. The latter system has apparent advantages of lowering mean collector temperature and raising the mean evaporator temperature of the heat pump, thus improving the performance of each. See Fig. 8; see also the entry on **Heat Pump.**

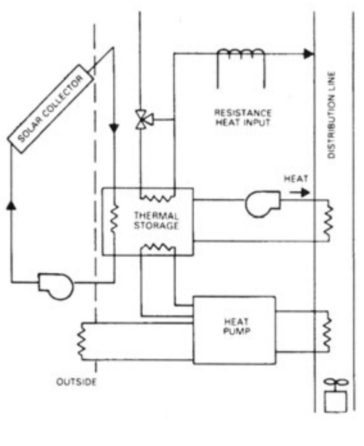


Fig. 8. Solar heating system with heat pump auxiliary.

In addition to close cycle absorption cooling, open cycles are of potential interest. Desiccants can be used to absorb water vapor from room air, which then can be evaporatively cooled. The desiccant is regenerated and recycled. Löf has suggested the use of triethylene glycol as a desiccant, with solar-heated air for regeneration. Lithium chloride also has been proposed as a desiccant.

Heat-Actuated Refrigeration. A variety of heat-actuated refrigeration cycles has been proposed for solar air conditioning. These can be divided into heat engine types, such as the Rankine and Stirling cycles, and the absorption machines. Most successful to date have been the lithium bromide-water and the ammonia-water absorption cycles. Regardless of type, operating temperature is a tradeoff when coupling a solar collector to a heat-actuated refrigeration machine. The efficiency of solar collectors decreases with temperature. On the other hand, the coefficient of heat-actuated refrigerators increases with generator temperature. Figure 9 shows Carnot coefficient of performance as a function of generator temperature for an evaporator temperature of 45°F (7°C) and for heat rejection temperatures of 120°F (49°C) (typical of air cooling) and 100°F (38°C) (warm water cooling, such as might be achieved with well water). For the case of absorption machines, it is assumed that absorber and condenser operate at a common heat-rejection temperature. Using these plots, the collection efficiency–Carnot coefficiency of performance product for the simple single-pane black collector reaches a maximum rate at about 175°F (80°C) and falls off rapidly at higher temperatures. For a more refined collector, such as a single pane with selective surface collector, the efficiency coefficient of performance-product is maximum and nearly constant in the range of 200 to 300°F (93 to 149°C).

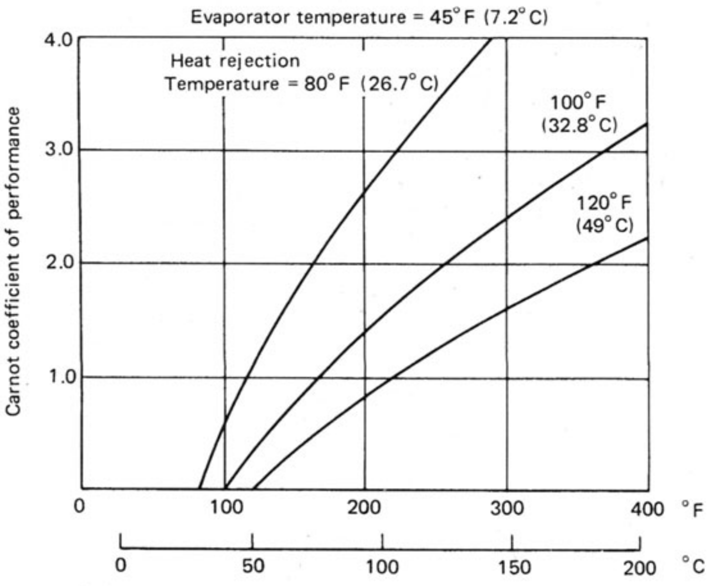


Fig. 9. Carnot coefficient of performance of heat-actuated refrigerators.

To achieve temperatures above the boiling point of water, concentrators are required as well as collectors. One of the largest high-temperature solar energy systems for building heating and cooling commenced operation in late 1978 at an eight-story office building (Minneapolis) which houses about 500 employees and has over 100,000 square feet (929 square meters) of working space. During an average year, the system was designed to generate about 50% of heating needs, 80% of cooling energy needs, and 100% of heat for hot water needs. On the roof of a parking ramp adjacent to the building, 252 troughlike collectors track the sun to focus its rays on liquid-filled pipes. The liquid, which may reach 177°C (350°F), is pumped to an isolation heat exchanger that allows different fluids, pressures, and flow rates to be used in the two-loop system, dictated because of the cold winter temperatures. Excess heat is stored underground in two 18,000-gallon (681 hectoliters) tanks until required. Each row of solar collectors is under the control of a local system that uses a photosensitive sun tracker and bidirectional electric drive ac motor. Wind, solar isolation and other safety control circuits are continuously monitored to protect the collector field from damage by weather excesses.

Passive Solar Heating and Cooling. Although not always the practical solution, one of the most sensible approaches to the utilization of solar energy does not require pumps, fans, etc., but utilizes the building or residence structure per se as the solar radiation absorber (or insulator). An example of the modern approach to passive approaches is, ironically, incorporated in Montezuma's Castle, built around 700 A.D. by the cliffdwelling Indians of Arizona. The basic philosophy is to design the structure to capture and retain heat during winter and to remain cool during summer. For example, the windows that face north are made small or largely eliminated. These windows do not contribute much to heat collection during summer or winter and thus, if small, diminish heat leakage in either direction during all seasons. South-facing windows are made large because they are required as radiation collectors during winter. However, they are protected during the summer season by an overhanging roof. These features alone, of course, are common-sense ideas that have been used by some designers for many years. Passive systems also take advantage of heavy masonry walls (or other sources of thermal mass) which can absorb solar radiation during the day and reradiate some of it at night. Such an arrangement could be called a "concrete collector." The designer can improve the effectiveness of extensive south-facing windows by constructing an interior masonry wall adjacent to the windows (lighting becomes a problem for special design). In an experimental building (Wallasey School, Liverpool, England), the south wall of the two-story concrete structure is made up essentially of double-glass windows with a heat-storage (or insulating) wall. The only supplementary heat required is derived from body heat of the students and heat radiated by electric lights. A structure of this type, of course, is subject to wider interior temperature variations than those to which much of society has become accustomed. Temperature swings can be reduced by partially decoupling thermal storage from the living and working space. In this concept, solar radiation entering the south-facing windows is absorbed by a masonry wall (sometimes called a Trombe wall) or by a water wall in which water-filled drums are placed. This wall insulates the building from high temperatures during daytime and transmits stored energy to the structure for warming during nighttime. There is an office building and warehouse in northern New Mexico which incorporates a water wall passive system and which provides 95% of the energy required for heating.

Some designers use a "roof pond," in which plastic bags filled with liquid are exposed to the sun during the day. They are covered with an insulating panel at night so that they can radiate stored heat downward to the structure. The cycle can be reversed to provide cooling during summer.

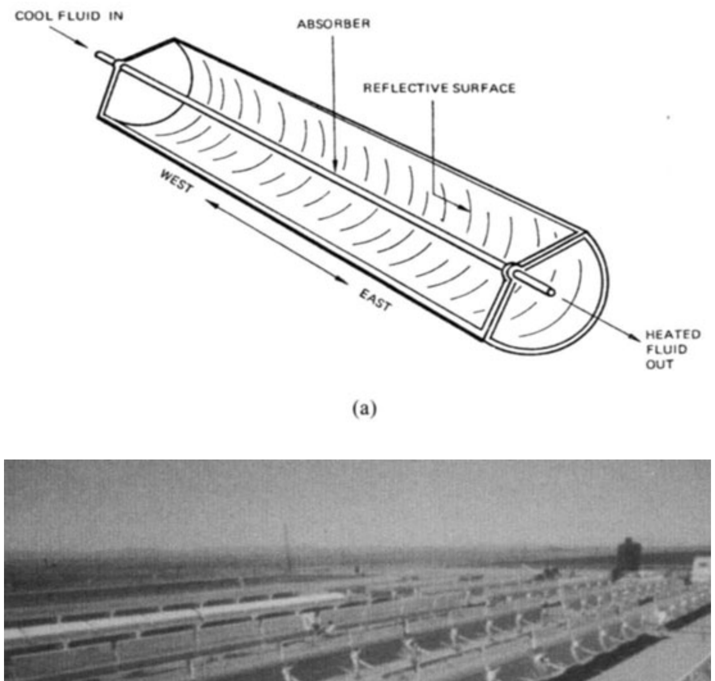
Investigations indicate that the optimum thickness for concrete thermal storage walls is about 30–40 centimeters (1–1.3 feet). Innovations in passive systems are appearing at a rapid rate. Some of these include movable insulation for shielding glass areas at night, and more compact thermal storage systems, such as ceiling tiles which have been developed by the Massachusetts Institute of Technology. These tiles contain a material that undergoes a phase change at 75°F (24°C), storing heat as the material melts and later releasing the heat to the room as the material solidifies. It is expected that, as passive systems improve, there will be a considerable impact on the traditional flat-plate collector.

Large-Scale, High-Temperature Solar Energy Systems

Systems in this category require concentration of solar radiation prior to its collection and utilization. Concentrators fall into three categories:

(1)*Nonfocusing concentrators* have the advantage that they do not have to continuously track the sun and thus do not require optical precision. Also, they can utilize both diffuse and direct radiation and thus are partially operable on cloudy days. They do not, however, operate as efficiently as focusing types, particularly during the early and late periods of the day. A simply designed nonfocusing concentrator essentially will consist of a stationary mirror or reflector located next to a flat-plate collector. In another approach, placing reflectors behind evacuated-tube collectors also accomplishes a modest degree of concentration. An advanced nonfocusing concentrator, known as the *compound parabolic concentrator* (CPC) was developed by the Argonne National Laboratory as the result of experience gained from designing light-concentrating devices for use in high-energy physics experiments. The device incorporates a parabolic surface designed to provide the maximum amount of radiation to an absorber for a given concentration ratio. In one configuration, the radiation is concentrated 1.8 times and, when operated in conjunction with a stationary collector, will reach temperatures as high as 250°F (121°C). Units with even higher concentration ratios (3× up to 6×) have been developed for use with evacuated-tube collectors and can achieve temperatures between 300 and 450°F (150 and 232°C). At a concentration ratio of 6, it is necessary to reorient the collector once each month. Currently, manufacturing costs are relatively high, but the CPC holds promise for a number of future applications.

(2) *Trough or line-focusing concentrators* track the sun by focusing in one direction only. Concentrations in this category of device range from 10× to 100× and they are capable of achieving temperatures between 200 and 600°F (93 and 316°C). On average, these devices deliver a minimum of 50% of the solar energy available to the heat-transfer medium in the absorber. In one configuration, mirrors form a parabolic trough that focuses radiation onto a linear absorber. Usually the mirrors are constructed of polished metal or coated plastic; the absorbers are blackened metallic pipe or evacuated-glass tubes. See Fig. 10. The entire assembly or array tracks the sun. Although normally considered in terms of relatively large thermal capacities, versions have been offered for residence and small building applications. Other, more sophisticated versions, operating at the high-temperature range, are used to drive Rankine-cycle heat engines for pumping irrigation water in a number of locations in the southwestern United States. Other installations include water heating for industrial processes. In another type of line-focusing concentrator, the optics are altered to utilize plastic Fresnel lenses for focusing the radiation onto the absorbers. Apparently a



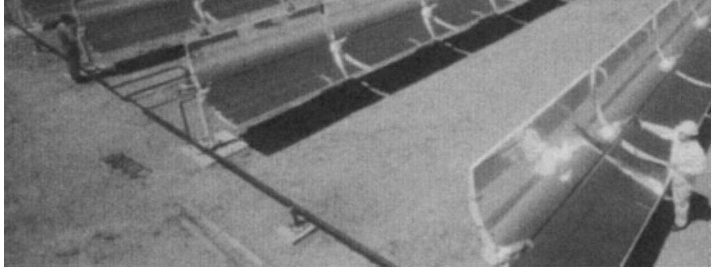


Fig. 10. (a) One configuration of an optical-type concentrator with axial absorber; (b) parabolic troughs under test for various uses, including irrigation pumping for croplands.

similar thermal result can be accomplished, as compared with the parabolic mirror approach, with a smaller optical surface.

(3) In another one-axis tracking concentrator, a fixed trough is made up of flat mirrors and tracking is accomplished by moving the absorber. Known as the *Russell collector*, this unit apparently can reach temperatures up to 900°F (482°C).

Two-axis focusing systems will be described a bit later.

High-Temperature Solar Energy

Prior to the serious consideration of high-temperature solar energy as a source of electric power, much was learned from the design of solar furnaces, the objective of which was the production of extremely high temperatures for materials testing, a very useful research function that continues. Much knowledge was and is continuing to be learned from the operation of solar furnaces—knowledge that is helpful in the design of solar power plants. It is fitting here, as a backdrop to describing the solar power tower concept, to present information on solar furnaces. Historically, solar furnaces have been selected for high-temperature research and development activities where a highly concentrated source of nonpolluting radiant energy is required. Generally such activities can be categorized as (1) high-temperature chemistry involving the formation of very pure or otherwise unique materials; (2) high-temperature processing by which a material is fused, purified or otherwise improved; (3) high temperature property measurements involving the determination of the behavior of a material under conditions which require a noncontaminating environment; (4) determination of the thermal shock resistance or other behavior of materials in a high-temperature, high-heat flux radiant energy environment; and (5) study of high temperature solar-thermal conversion systems. Certain of these applications may be refined further by conducting the operation in an optically transparent vessel or one containing a transparent window such as fused quartz through which the radiant energy may pass and in which the composition and pressure of the atmosphere can be controlled.

A few examples of the types of high temperature studies which have been conducted in the previously described categories are: (1) gas phase reactions to form pyrolytic graphite; (2) production of very high purity fused aluminum oxide and fused silica, the production of stabilized zirconia and the purification of reactive metals in a controlled atmosphere; (3) determination of microwave transmission characteristics of dielectric materials at very high temperatures; (4) study of the thermal shock resistance of materials under high heat flux thermal radiation conditions simulating exposure to the thermal radiation pulse provided by a nuclear explosion; and (5) study of heat exchangers, such as boilers and superheaters for the production of steam for electric power generation.

Although the motivation for the design of such furnaces may be for high-temperature research, much can be learned from them that is applicable to the design of solar energy facilities for power generation. Up to the point of conversion, the problems are essentially parallel.

Solar Furnaces in France. In 1948, under the leadership of Professor F. Trombe, the Centre National de la Recherche Scientifique (CNRS) in Paris undertook the design, construction, and development of the world's first large solar furnace at Montlouis in the French Pyrenees mountains. This furnace was completed in 1952, and provided 50 kilowatts of thermal energy. The Montlouis solar furnace became the prototype design for other large high-temperature solar furnaces. Basically, this design utilized a single large heliostat (array of numerous flat mirror elements) which continuously tracked the sun to direct the sun's rays onto a concentrating reflector (parabolic or spherical) consisting of many smaller mirror elements each of which was contoured to concentrate the incident radiation at a common focal point. In the case of the Montlouis furnace, the heliostat was 43 feet (13.1 meters) wide and 34 feet (10.4 meters) tall and contained 540 flat mirrors each 50 × 50 centimeters. The concentrating reflector was made up of 3,500 mirrors 16 × 16 centimeters arranged in a parabolic configuration 36 feet (11 meters) wide and 30 feet (9.1 meters) high with a focal length of 6 meters. Each of the 3,500 flat mirror elements in the parabolic concentrator was mechanically contoured and aligned to focus the radiation received from the heliostat onto the focal point of the parabola.

The successful performance of Montlouis solar furnace led to the use of its design as the prototype for the next three large single heliostat-concentrator solar furnaces which were to be built during the next twenty years. All three of these furnaces were similar to the Montlouis furnace in size, operation and thermal power level and were constructed by: (1) U.S. Army Quartermaster Corps, Natick, Massachusetts; (2) Tohoku University, Sendai, Japan; and (3) the French Army's Laboratoire, Central de L'Armement, Odeillo, Font-Romeu, France. In 1973 the U.S. Army's solar furnace was moved to the Nuclear Weapon Effects Laboratory, White Sands Missile Range, New Mexico, where it became operational in 1974.

Although the Montlouis solar furnace played a major role in developing applications for high-temperature solar energy and in providing design information for the three other large solar furnaces, its most valuable contribution to the field of high-temperature solar energy was the experience and background which it provided the CNRS Solar Energy Laboratory that led them to design and construct the CNRS 1,000-kilowatt solar furnace.

The CNRS 1,000 kilowatt solar furnace is located at Odeillo, Font-Romeu, altitude of 5,900 feet (1798 meters) about 25 miles (40 kilometers) east of Andorra and 5 miles (8 kilometers) west of Montlouis. At this location the sun shines as many as 180 days a year and solar intensities as high as 1,000 watts per square meter are common. The solar furnace was completed on October 1, 1970, after more than 10 years of construction.

Figure 11 is a schematic of the CNRS 1,000-kilowatt solar furnace. This furnace utilizes 63 heliostats to direct the sun's rays onto the surface of the giant parabolic concentrator.

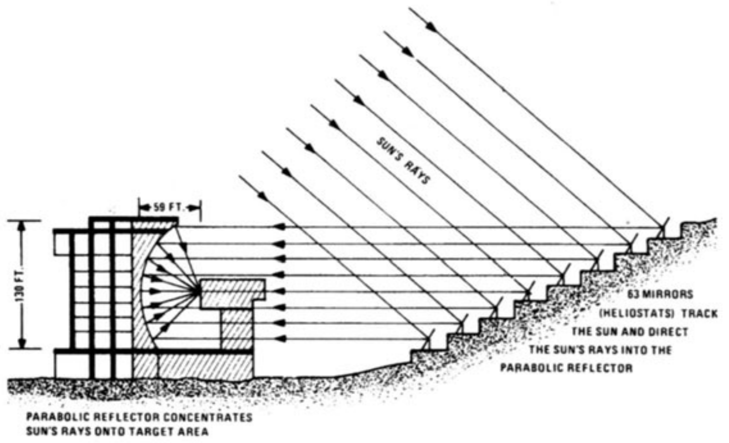

SUN'S RAYS
59 FT.
130 FT.
63 MIRRORS (HELIOSTATS) TRACK THE SUN AND DIRECT THE SUN'S RAYS INTO THE PARABOLIC REFLECTOR
PARABOLIC REFLECTOR CONCENTRATES SUN'S RAYS ONTO TARGET AREA

Fig. 11. Schematic representation of the 1000-kilowatt solar furnace at Odeillo, Font-Romeu, France. (*Centre National de la Recherche Scientifique.*)

The 63 heliostats are each 7.5 meters wide by 6 meters high and contain 180 single flat mirror elements 50 × 50 centimeters. The total area of mirror surface in the 63 heliostats is 2,835 square meters or over one-half the playing area of a football field.

The heliostats are located directly north of the parabola and are arranged on eight terraces. Each terrace corresponds in elevation to one of the floors of the building supporting the concentrating parabola. A solar beam of constant energy is thus directed horizontally and southward from the heliostats to the mirrors which make up the concentrating parabola.

Each heliostat is designed to illuminate a specific area on the parabola and is equipped with a dual optical control system which maintains the proper orientation for each heliostat by means of a dual hydraulic system. This dual system permits each heliostat to be operated in either a "search" or "track" mode. In both cases the optical guidance system uses an optical tube which contains four photodiodes which control the heliostat motion in east-west and up-down direction.

When operating in the "search" mode a short (10-centimeter) optical tube with a 40 degree acceptance angle is used to activate the "fast" hydraulic system which operates in an on-off mode to quickly bring the heliostat to within the operating range of the "track" system. In the "track" mode a 100-centimeter optical tube is used to control a slower acting hydraulic system which operates in a proportional control mode. The size of the sun's image at the base of the 100-centimeter tube is $\frac{1}{2}$-inch (13 mm) in diameter and the accuracy of the control is one minute of arc.

The concentrating parabola has a focal length of 18 meters, is 40 meters high and 54 meters wide, and the focal axis is 13 meters above the first floor. The parabola consists of 9,500 initially flat glass mirrors which were mechanically curved and adjusted to provide a solar image of minimum diameter at the focal point. Almost two years were required to accomplish these two precise adjustments which were completed on 1 October 1970. Figure 12 shows the parabola and the focal building into which the concentrated solar energy is directed. See also Fig. 13.

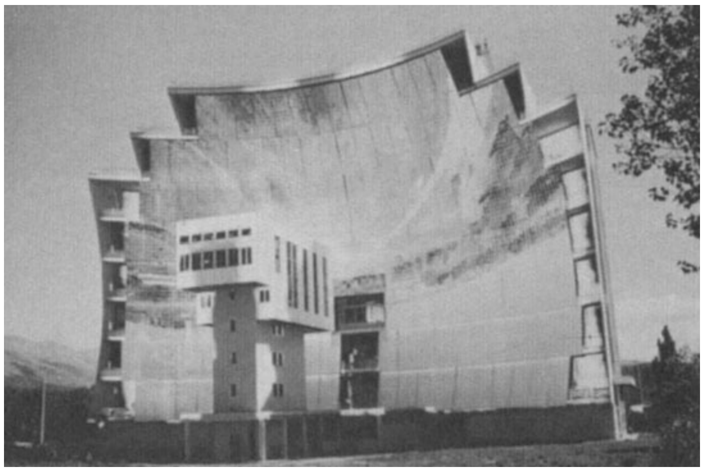


Fig. 12. Large parabolic reflector and focal building in foreground. Concentrated energy is directed at the solar furnace located within the focal building. Installation is at Odeillo, Font Romeu, France. (*Photo by Glenn D. Considine.*)

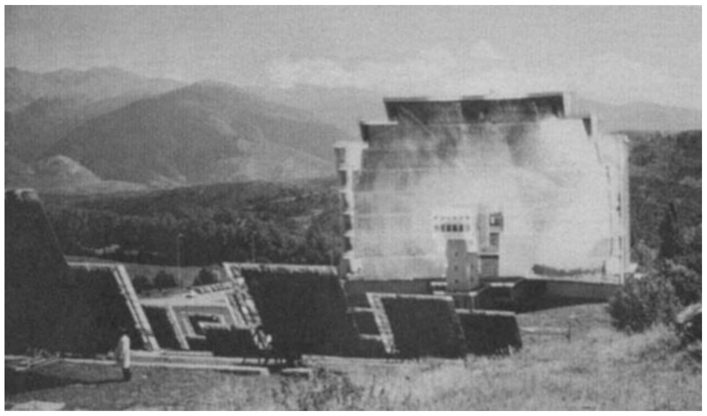


Fig. 13. Field of heliostat-controlled collector-reflector mirrors, which direct their energy to the parabolic reflector at the solar furnace installation at Odeillo, Font-Romeu, France. (*Photo by Glenn D. Considine.*)

The solar energy incident on an area of about 2,000 square meters is concentrated by the parabolic reflector onto an area of less than 0.3 square meters. Sixty percent of the total thermal energy (about 600 kilowatts) is concentrated in an area of less than 0.08 square meter at the center of the focal plane of the parabola.

The diameter of the image of the sun at the focal point is 17 centimeters and 27% of the thermal energy (about 270 kilowatts) is concentrated in this area. Heat flux data in watts per square centimeter in the focal area are presented graphically in Fig. 14. Curve 0 represents the heat flux at the focal plane. Curve d/2 shows the heat flux and temperature at a plane removed one-half the diameter of the solar image (8.5 centimeters) behind the focal plane. Curve d presents the same data on a vertical plane removed one diameter of the solar image (17 centimeters) behind the focal plane.

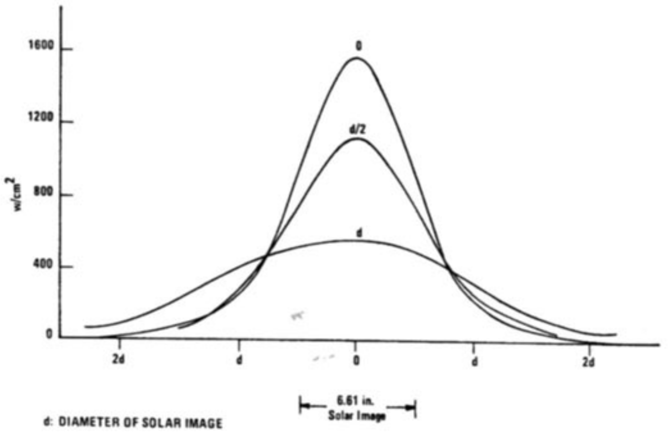


Fig. 14. Solar energy versus distance from focal point in Odeillo solar furnace. (*Georgia Institute of Technology.*)

Solar Tower Energy Collector

Authorities have observed that solar energy can be usefully collected optically from one square mile (2.6 square kilometers) of surface area, or even larger, and concentrated onto a central receiver by an array of heliostats, i.e., independently steered mirrors. By judiciously spacing mirrors over 35% of the area, such a system in the desert southwest of the United States, for example, could collect 2800 megawatt-hours thermal per day in midwinter and almost twice that amount of energy in midsummer. In order that the reflected radiation from this field be efficiently intercepted, the central receiver would have to be several hundred meters high.

Unlike the Odeillo installation, previously described, where a field of heliostats finally focus their energy to a small aperture by way of a huge parabolic reflector, in the solar tower approach, the energy from each mirror is directed to a central receiving tower, located high above the field, as shown schematically in Fig. 15. Shown is a large array of heliostats by which essentially flat mirrors are automatically steered to reflect or redirect the incident solar radiation to a high tower. It is assumed that the terrain of the heliostat field is flat. However, a gentle slope southward would be advantageous. After reflection from the mirrors, the redirected solar energy can be absorbed and converted to heat by a black body receiver placed in the focal region. The heat can be transported down the supporting tower by way of liquid metal and/or steam lines and can be stored or used to operate a conventional turbine generating station. Alternate uses would be direct conversion to electricity by way of high-power-density solar cells placed in the focal region, or use of the heat to produce a fuel thermodynamically. Two-axis control can be obtained by either hydraulically or electrically operated servo-mechanisms that derive a signal from a simple position-sensing element.

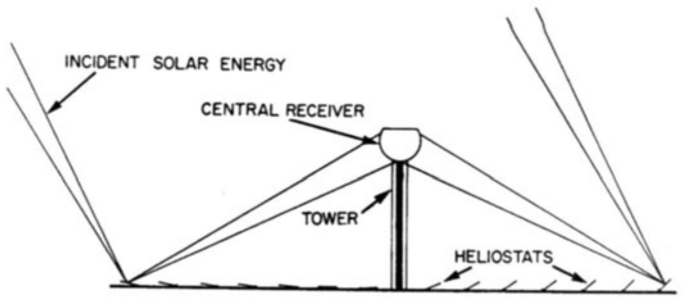


Fig. 15. Schematic diagram of solar tower energy concept.

It requires energy collected from an area like a square mile to be of interest to power utilities. Energy collection from hundreds of such installations would be required to make a significant impact on the energy supply. If one replaces the 300-meter tower with geometrically identical systems using 100-meter towers, it is found that 9 such towers would be required. Although the cost of nine smaller towers would compare with the cost of a single, large tower, additional costs and losses would be incurred in connecting heat-transfer lines to a central generator to handle 9 collectors. Also, heliostats smaller than about 20 square meters are not economical because the cost per heliostat of the support, actuator, and steering systems have a substantial fixed component.

The first choice of heat-transfer fluid would be steam because it would appear not to require any new technology. However, because the flux density which must be absorbed and transferred to the fluid can be appreciably higher than in conventional steam plants, efficient operation may require some new technology. Also, because of the large daily and seasonal heat flux variations, the design of the receiver is not trivial and may ultimately be best accomplished by utilization of liquid metal, such as sodium, for heat transfer from the receiver surface to a steam line. Liquid sodium technology has been developed for nuclear reactors and operating temperatures of 550 to 650°C appear reasonable. Sodium also presents a promising high-temperature thermal storage medium. In general, the thermal cycling due to the intermittent nature of solar energy of this high-temperature system will have to be considered in any detailed design. The black body boiler surface should not deteriorate at high temperature in the presence of air. If deterioration is a problem and if convective losses are appreciable an inverted cavity design can be used in order to avoid the use of vacuum jacketing.

Solar energy may best be utilized at first by a steam generator operating in a solar-only mode with short-term storage of an hour or less to provide operational stability. Because utilities require very high reliability, little new plant capacity credit would be given such a plant because of possible cloudiness. New plant credit could be given if there were possibilities of using liquid fuels, such as liquid petroleum gas or oil for backup. This could be accomplished by adding a simple low-cost, but possibly inefficient burner to add heat. Although liquid fuels may be in short supply, they are easily stored and afford an ideal way of giving a solar plant high reliability as an intermediate plant. Depending on actual operating conditions, it may be that very little liquid fuel is burned. The comparison of such a hybrid plant should be made with a solar plant that has a gas turbine as a backup. An alternate approach might be to store solar energy as sensible heat, perhaps in underground cavities, or as latent heat.

Some authorities believe that close ties between solar power and conventional electric power plants—so-called solar thermal electric designs—represent an ideal approach to the large-scale use of solar energy. In this concept, instead of a solar-powered facility being linked to traditional power-generating facilities by way of the electric grid, a fossil fuel back-up for a solar system would be located at the same site as the solar electric plant. Savings could be realized through the common usage of certain equipment items. In one design, the oil-fired backup would use the same turbine as a power tower system. Some designers have estimated that this additional capability would add only about 0.26% of the total cost of the solar electric facility.

There are other authorities who believe that adapting solar energy to electric utilities will limit the economic potential of solar energy. For example, Asbury and Mueller generally conclude from a study of the topic that conventional electric utility systems and most solar energy systems represent a poor technological match. The basic problem, as envisioned by these investigators, is that both technologies are very capital intensive and that the electric utility, because of the high fixed costs of generation, transmission, and distribution capacity, represents a poor backup for solar energy systems. On the other hand, the solar collection system, because it represents pure, high-cost capital and because of its outage problems, cannot be considered as a part-load source of auxiliary energy for the electric utility system.

Solar Energy Plant at Electric Utility Level

Construction commenced in 1975 on an experimental 10 MWe central receiver pilot plant in a combined effort by two electric utilities in the southwestern United States, Southern California Edison and the Los Angeles Department of Water & Power, who worked in cooperation with the U.S. Department of Energy and the California Energy Commission. The start of continuous electric power production commenced in August 1984 and the plant is now up to its design capacity of 10 MWe. The plant, known as *Solar One*, is located in Daggett, California just off Highway 40 and east of Barstow. A panoramic view of the facility, clearly showing nearly 2000 heliostat-controlled mirrors focusing their collected energy on a boiler atop a 300-foot (91-meter) tower, is given in Fig. 16.

Although large and very impressive, *Solar One* is regarded as an experimental pilot plant for proving and testing technological improvements that can be incorporated in future commercial-size plants. The Daggett plant is a scale model of a 100 MWe generating plant. On its own, *Solar One* is currently furnishing the electricity requirements for a community of about 6000 people. *Solar One* relies on a combination of both old and new solar technology. Certain features not found in typical commercial generating plants allow great flexibility in plant operation. Several different types of solar central receiver plants can be simulated within this one project.

The Basic Concept of Solar One. Computer-controlled mirrors (heliostats) totaling 1818 in number form a circular array around a central tower. Within the receiver, the solar energy is transformed into high-temperature thermal energy in a water-steam heat transport fluid. The thermal energy can be converted to electric power immediately or stored to extend plant operation. See Fig. 17. The collected solar energy is most efficiently put to work as receiver steam to power a turbine-generator (Path A). If the energy is to be stored, receiver steam follows path B and heats oil that is routed to and from the thermal storage tank. Energy is discharged from storage by using hot oil from the tank (path C) to generate steam, which is then sent to the turbine along path D.

The thermal storage system uses oil as both a thermal storage medium and a heat transport fluid. The maximum operating temperature of the storage system is 575°F (300°C). As a result, electricity is generated less efficiently than when 960°F (515°C) receiver-supplied steam is used directly in the turbine.

The operating temperature of the storage system simulates steam generation conditions in industrial plants and the chemical processing industry. Furthermore, because storage-supplied heat can supplement solar-supplied energy, *Solar One* can simulate a plant that uses both conventional fuels and solar energy.

Heliostats. The facility receives 3600 to 4000 hours of sunlight/year (9.8 to 10.9 hours/day). Construction of the 1818 heliostats for the pilot plant demonstrated that prototype designs can be successfully produced in volume quantities with conventional manufacturing techniques. Each heliostat has a reflective area of 430 square feet (39.3 square meters). The heliostat glass is specially formulated to contain a minimum amount of impurities. As a result, 91% of the incident sunlight can be reflected when the mirror surface is clean. A close-up of a set of mirrors (in a vertical position for demonstration purposes) is given in Fig. 18.

The vertical and horizontal movement of the heliostats is directed by a control system—a microprocessor in each heliostat, a controller to regulate groups of up to 32 heliostats, and a central computer. Over 97% of the heliostats are available more than 98% of the time. Operation of the heliostats has suggested areas for further research and development; for example, rain water may be sufficient to maintain the cleanliness of the mirrors, and mechanical rinsing may be required only in dry months.

The heliostats, as shown in Fig. 19, are distributed in a south field (578) and a north field (1240). The mirrors are slightly concave (approximately 1000 foot focal length with $\frac{1}{6}$-inch curvature in a 10-foot length). The total weight of a heliostat, as previously shown in Fig. 18, is 4312 pounds (1956 kg). The heliostats are normally stowed in a vertical position except when high wind conditions exist. During daylight hours, of course, the mirrors are rotated by a drive mechanism to follow the direct solar rays as closely as possible. It is interesting to note that the sun's position is calculated rather than sensed—so that even when clouds briefly cover the sun, maximum energy is reflected.

Receiver. On top of the steel tower rests the cylindrical receiver, a superheated steam boiler that is 14 meters (46 feet) tall and 7 meters (23 feet) in diameter. The receiver weighs almost 50 tons and is positioned over 20 stories above the ground. Feedwater is pumped to the bottom of the receiver, where it is vaporized to superheated steam in a single pass to the receiver's top. The steam is then piped to the turbine-generator at the foot of the tower. This steam can also provide heat to the thermal storage system.

Thermal Storage System. On a clear day, the receiver can generate sufficient steam to simultaneously operate the turbine and also deliver heat to the storage system. The thermal storage can generate power in

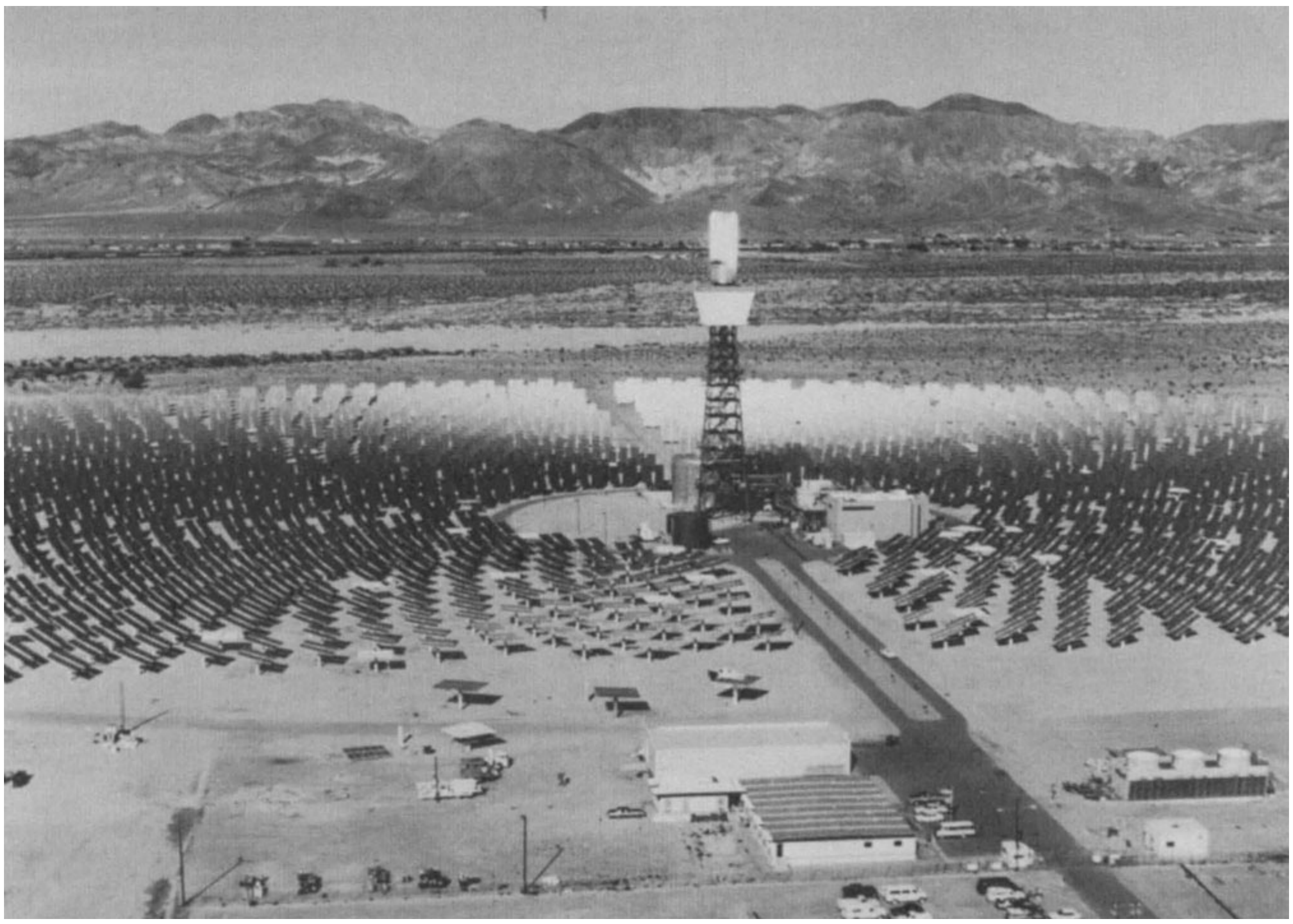


Fig. 16. Panoramic view of the world's largest solar thermal electric power plant, located on the Mojave Desert, near Daggett, California. The 10 megawatt (electric) facility commenced operation in 1982 and achieved design capacity in 1984. The collector tower (receiver upon which solar energy is reflected) is located atop a 300-foot (91-meter) tower. The north field of heliostats (mirrors kept in synchronism with the movement of the sun), 1240 in number, is shown in background; the south field (578 heliostats) is shown in foreground. Operating facilities, turbine generators, and storage tank are shown in circular middle section under the tower. The facility is operated by Southern California Edison and the Los Angeles Department of Water and Power, in conjunction with the U.S. Department of Energy and the California Energy Commission.

the evening or during periods of cloud cover. It also provides steam for start-up in the morning and for keeping selected portions of the plant warm when the plant is not operating. The steel-walled insulated storage tank has a capacity of 3.5 million liters and sits on lightweight insulating concrete. The tank is filled with sand, rocks, and a high-temperature thermal oil. Steam from the receiver is routed through heat exchangers to heat the thermal oil, which is then pumped into the tank to heat the rock and sand. This stored thermal energy can then be transferred to the turbine generator for electrical power production.

Power Generation Station. The turbine-generator is rated at 12.5 megawatts and is sized to handle the full plant system output plus all internal plantloads. The dual-admission turbine has a high-pressure steam inlet for steam produced by the receiver and a low-pressure inlet for steam produced from thermal storage. The rated turbine thermal-to-electric efficiency from receiver to steam is 35%. The efficiency is 25% from the lower quality thermal storage system.

Master Control System. In the morning the operator, through keyboard commands, positions the heliostats at standby operating points, begins water circulation in the receiver, and then issues a command to the system to start up the plant. At this point, a computer takes over and automatically directs heliostats to track the receiver. When receiver steam conditions are correct, steam is routed to the turbine. The operator then synchronizes the turbine to the electrid grid, after which the minimal manual attention is needed. If conditions change, such as a cloud passing over, the control system automatically makes adjustments to keep the plant in the best operating state. If some abnormal event occurs, alarm messages tell the operator which parameters are out of normal operating range. The operator can, at any time, make changes in any plant operating condition. While the pilot plant control system was designed for controlling a water-steam central receiver solar plant, the basic functions and operating philosophy are readily adaptable to other power plants.

Performance. The requirement for production of 10 MWe was exceeded by a peak production of 12.1 MWe. Similarly, the required 7 MWe net generation from storage was exceeded by an output of 7.3 MWe. The plant also has successfully operated down to 0.5 MWe, which is considerably lower than the designed minimum operating production level of 2 MWe. The minimum sunlight threshold for operation was designed as 450 W per square meter, yet the plant has operated in direct solar radiation levels as low as 300 W per square meter. In an endurance test, the receiver and storage system kept the turbine continuously on-line for 33.6 hours and generated 127 MWe net.

Solar One was designed to have 95% of the heliostats available at any one time. Between April 1982 and April 1983, 98% of the heliostats were available for operation. This percentage later increased to 99%. The establishment of the sharp thermal gradient (thermocline) needed for the storage system has been verified. Gradients of 49°C/meter have been measured. Equally important is the very low rate of heat loss from the storage tank. The tank heat loss has been measured at 1.3% day.

Central Receiver Test Facility

The largest facility specifically designed for testing central receiver components and subsystems was built in Albuquerque, New Mexico in the late 1970s. At this facility, a 15-meter (49-foot) diameter concrete-and-steel receiver tower rises some 60 meters (197 feet) above the ground. Within the tower, three test bays at different levels are used for experiments. A huge elevator can transport equipment weighing as much as 100 tons to these test bays. There is a total of 222 heliostats;

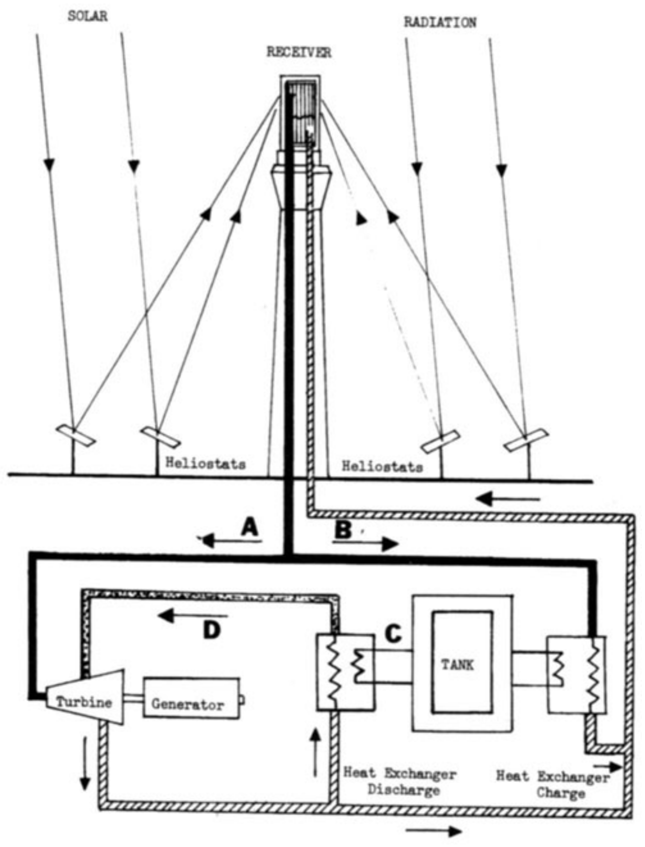


Fig. 17. General concept of *Solar One*. Within the receiver, the solar energy is transformed into high-temperature thermal energy in a water-steam heat transport fluid. The thermal energy can be converted to electric power immediately or stored to extend plant operation. The collected solar energy is most efficiently used as *receiver steam* to power a turbine-generator (Path A). If the energy is to be stored, receiver steam follows Path B and heats oil that is routed to and from the thermal storage tank (Path C) to generate steam, which is then sent to the turbine along Path D. The thermal storage system uses oil as both a thermal storage medium and as a heat-transport fluid.

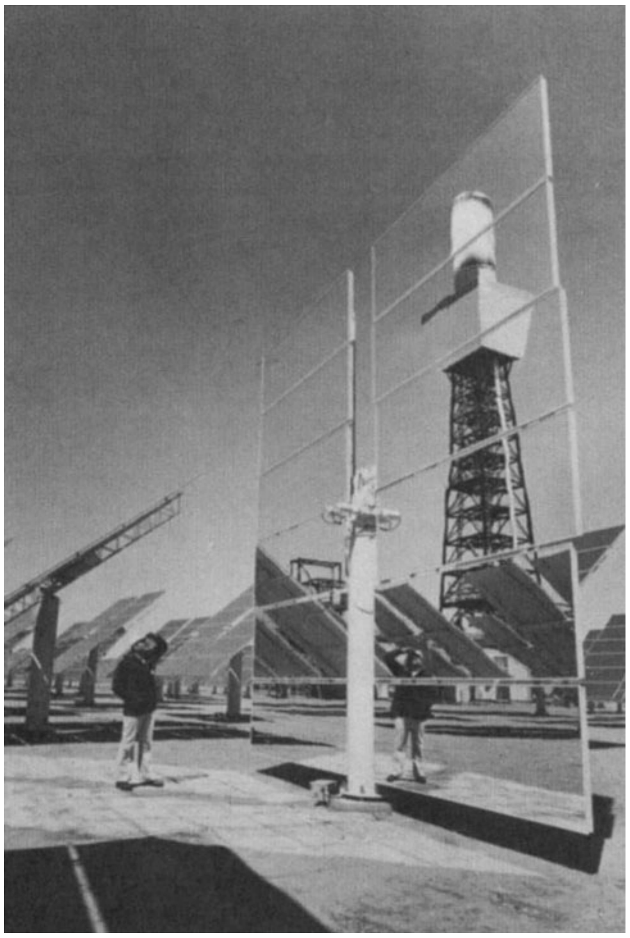


Fig. 18. Close-up of a heliostat rack assembly shown in vertical position for demonstration purposes. Note reflection of tower in mirrors. *Solar One* has a total of 1818 heliostats.

when all of them are focused on a receiver in one of the test bays, temperatures in excess of 2000°C (3632°F) can be generated. In practice, lower temperatures are used for receiver testing. The test facility has been used to try out innovative receivers that use gas, liquid sodium, or molten nitrate salts for thermal transport. In one system, molten salt has been used as the heat transport fluid and storage medium in an integrated central receiver system to produce an electrical power output of 750 kWe.

Solid-Particle Central Receiver. A new type of receiver has been under investigation. A novel concept for a central receiver uses sandsize refractory particles that free-fall in a cavity receiver. A conceptual design is shown in Fig. 20. Scientists observe that the advantages of a solid particle receiver over traditional fluid in-tube receivers are: (1) the particles can directly absorb solar radiation, and (2) the particles maintain their integrity at high temperatures. These advantages, coupled with the possibility that the particles can serve as the storage medium, could provide a cost-effective means of high-temperature solar energy utilization. High temperatures are attractive for fuels and chemical production, industrial process heat applications, or Brayton cycle electricity generation. The concept is in an early experimental stage.

Heat Engine Cycles for Solar Power

Heat engines for conversion of solar energy to electric power ideally should have the following attributes: (1) low cost per kilowatt output capacity; (2) long life and reliable operation with minimal maintenance; (3) safe and environmentally acceptable operation; (4) characteristics compatible with cycle top temperatures up to 1,000K; and (5) efficiency approaching Carnot values.

Heat engines that are potential candidates for coupling a solar heat source include thermoelectric, thermionic, thermochemical, magnetohydrodynamic, Rankine, Brayton (simple or recuperated), and cascaded cycles.

Rankine Cycle. The steam-Rankine cycle employing steam turbines has been the mainstay of utility thermal electric power generation for many years. The cycle, as developed over the years, is sophisticated and efficient. The equipment is dependable and readily available. A typical cycle (see Fig. 21) uses superheat, reheat, and regeneration. Heat exchange between flue gas and inlet air adds several percentage points to boiler efficiency in fossil-fueled plants. Modern steam Rankine systems operate at a cycle top temperature of about 800K with efficiencies of about 40%. All characteristics of this cycle are well suited to use in solar plants.

Brayton Cycle. In recent years, attention has been drawn to the Brayton cycle as a potential and practical alternative to the steam Rankine cycle for solar power and for high-temperature gas-cooled nuclear reactors. The Brayton cycle is most familiar in its open form as used in aircraft gas turbines. The open Brayton cycle cannot compete with steam-Rankine in efficiency. In a power-generation application, cycle efficiencies on the order of 20% would be expected. However, the Brayton cycle can achieve higher efficiency through recuperation, sometimes called regeneration. A representative cycle diagram is given in Fig. 22. The working fluid is an inert gas, typically helium. Inert gas mixtures, such as helium-xenon, have been studied and have potential advantages.

The recuperated Brayton cycle approaches Carnot efficiency in the ideal limit. As compressor and turbine work are reduced, the average temperatures for heat addition and rejection approach the cycle limit temperature. The limit is reached as compressor and turbine work (and cycle pressure ratio) approach zero and fluid mass flow per unit power output approaches infinity. It can be expected from this that practical recuperated Brayton cycles would operate at relatively low pressure ratios, but be very sensitive to pressure drop. With the assumption of con-

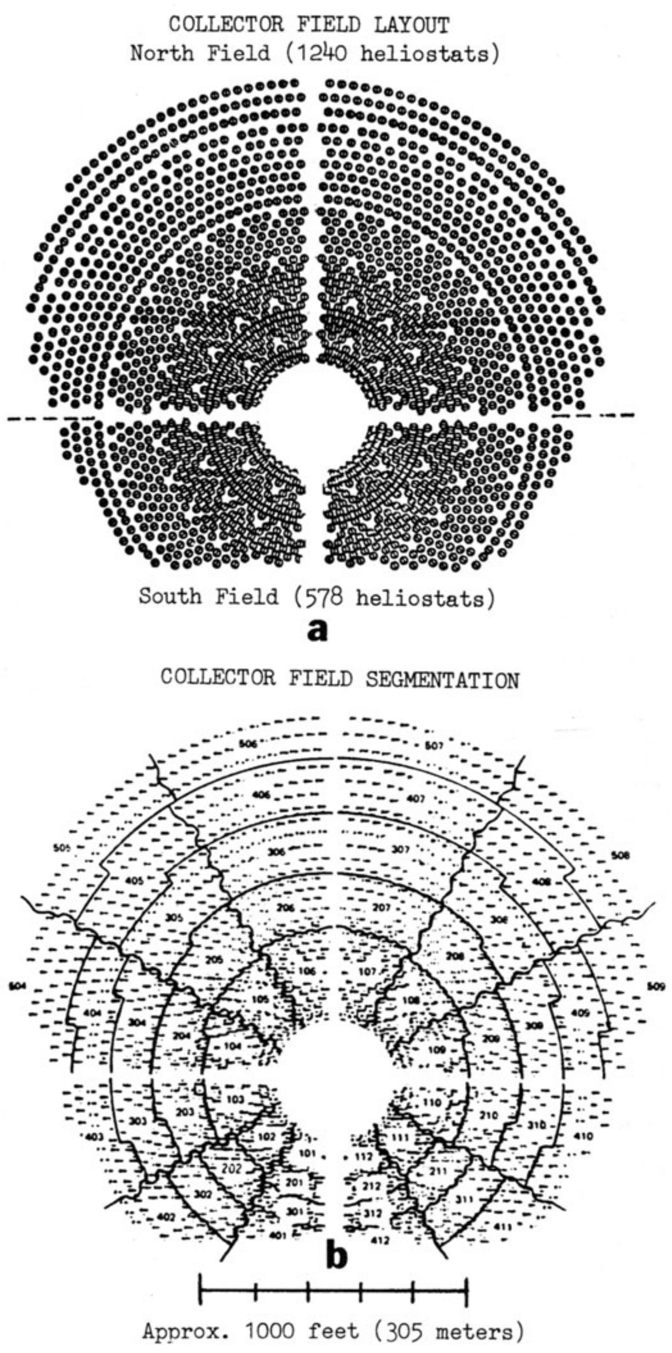

COLLECTOR FIELD LAYOUT
North Field (1240 heliostats)
South Field (578 heliostats)
a
COLLECTOR FIELD SEGMENTATION
b
Approx. 1000 feet (305 meters)

Fig. 19. (a) The heliostats are distributed in two fields—North and South; (b) for control purposes, the heliostats are segmented.

stant gas specific heat over the cycle temperature range, a good assumption for helium, the cycle efficiency of a recuperated Brayton cycle may be expressed:

$$\eta e = 1 - \left[\frac{\dfrac{r_{pc}\zeta - 1}{\eta_b} + \dfrac{\Delta T_r}{T_0}}{\dfrac{T_3}{T_0}\eta_T\left[1 - \left(\dfrac{G}{r_{pc}}\right)^{\zeta}\right] + \dfrac{\Delta T_r}{T_0}} \right]$$

where r_{pc} is compressor pressure ratio (> 1)
η_b is compressor efficiency
ΔT_r is temperature difference across recuperator ($T_4 - T_2$ in Fig. 21)
T_0 is cycle lower limit temperature
T_3 is cycle top temperature
η_T is turbine efficiency
ζ is specific heat factor, $(\gamma - 1)/\gamma = 0.4$ for $\gamma = 1.67$ as for helium
G is the product of the four pressure drop factors,

$$\left(\frac{P_1}{P_2}\right) \quad \left(\frac{P_2}{P_3}\right) \quad \left(\frac{P_4}{P_5}\right) \quad \left(\frac{P_5}{P_0}\right)$$

See also **Gas and Expansion Turbines.**

Solar Energy for Industrial/Metallurgical Processes

It is interesting to note that experiments on melting large masses of metals were conducted at the French Odeillo solar furnace as early as the mid-1970s. Since the mid-1980s, solar furnaces in other locations have researched what promises to be an important technology in the near future—namely, Solar Induced Surface Transformation of Materials (SISTM). In the past, numerous heat-treating methodologies have been practiced widely. Less traditional methods have been developed in recent years, including cladding/coating, self-propagating high-temperature synthesis, thin-film deposition, and ion-, laser-, and electron-beam processes.

Researchers have found that, for delivering large fluxes on target, the solar furnace is less capital intensive than competing methods that require an intermediate energy-conversion step. For example, a 1-square-meter (11 ft^2) solar dish can deliver about 1 kW of optical power to a target. To deliver the same amount of radiant energy, an arc lamp would have to be powered by the electrical output of an 11-square-meter (120 ft^2) dish, and a carbon dioxide laser would require the energy supplied by 26-square-meter (280 ft^2) disk.

The automotive industry, which turned to laser transformation hardening of engine and other drivetrain components, is now seriously investigating the "solar hardening" process.

Solar Furnaces No Longer Uncommon. The past decade has brought increased interest in using solar energy for industrial processes, not just to conserve energy or avoid the use of fossil fuels. Some of the important installations as of the early 1990s are listed in Table 1.

The time required to bring a part to treating temperature also is an important consideration. Solar furnaces have no competition on this point. See Table 2.

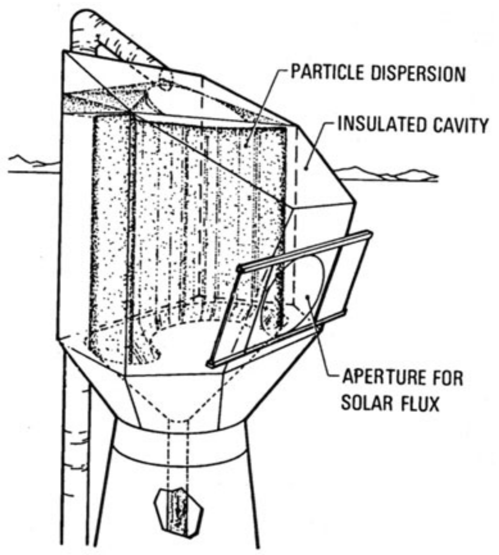
Chemicals Directly from Solar Energy. Whereas the early solar furnace located in Odeillo, France, was constructed for solar energy studies in general and later used for melting huge masses of metals and materials and the solar plant in the Mojave Desert in California is being used directly to generate electric power, scientists in the early 1990s are taking a somewhat different approach.

Researchers have recognized that the regions where ample sunshine is available for harvesting seldom coincide with the population and industrial centers of the world. Further, solar energy is of an intermittent nature because it depends upon clear skies with no cloud cover. I. Dostrovsky (Weizmann Institute of Science) observes that most of the limited amount of research on solar energy focuses on converting sunlight to electricity, mainly by photovoltaic and thermal methods, or, in the case of the Daggett, California, utility installation, to furnish heat to supply steam turbines.

An alternative approach is that of using solar energy to produce chemicals that can be stored and transported much as present fossil fuels. One installation along these lines is now undergoing demonstration operation in Saudi Arabia. This project, known as *HYSOLAR*, is a joint venture of Saudi Arabia and Germany. In essence, the facility consists of a plant that produces hydrogen. See also **Hydrogen (Fuel).**

One process involves the high-temperature decomposition of sulfuric acid, using recoverable iodine as an intermediate reactant. In a first stage, sulfuric acid yields water and sulfur dioxide. In a second stage, the sulfur dioxide plus water and iodine yield hydrogen iodide and sulfuric acid. In the third stage, the hydrogen iodide yields hydrogen (the desired product) and recoverable iodine.

In an electrolytic process, hydrogen is produced by electrolyzing a mixture of sulfur dioxide and water to produce sulfuric acid and hydrogen. In still another electrolytic process, bromine, sulfur dioxide, and water react to form hydrogen bromide and sulfuric acid. By applying 0.62V to the hydrogen bromide molecules, hydrogen is yielded and the original bromine is recovered.



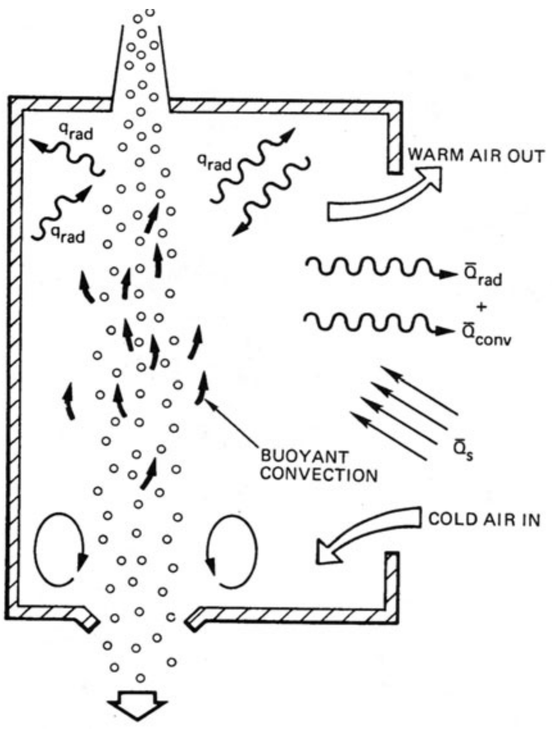


Fig. 20. Solid-particle central solar energy receiver: (a) conceptual design; (b) thermal phenomena in a solid particle receiver. (*Sandia National Laboratories*.)

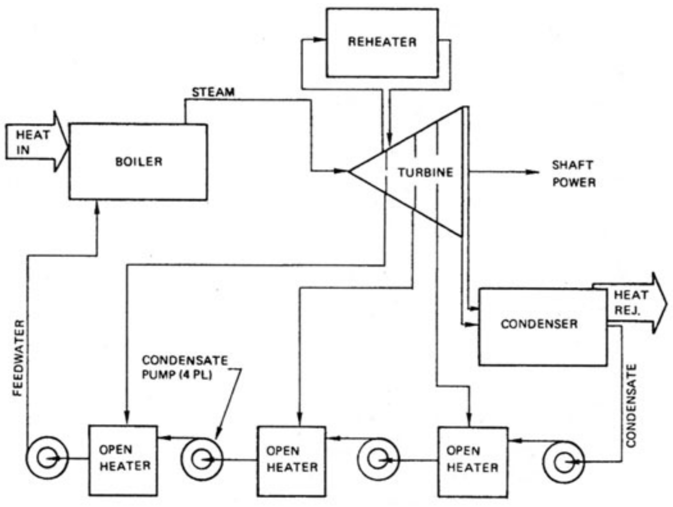


Fig. 21. Schematic diagram of regenerative-reheat steam Rankine cycle.

TABLE 1. COMPARISON OF HIGH-FLUX SOLAR FACILITIES

Location	Total Power kW	Peak Flux MW/m^2
Albuquerque, New Mexico		
Central receiver test facility	5000	2.4
Furnace	22	3.0
Atlanta, Georgia		
Furnace	1.3	9.5
Golden, Colorado	10.0	2.5
(measured using nonimaging secondary concentrator)		20.0
White Sands, New Mexico	30	3.6
Odeillo, France		
Horizontal furnace	1000	16.0
Vertical furnace	6.5	15.0
Rehovot, Israel		
Central receiver test facility	2900	
Furnace	16	11.0
Uzbek, Russia	1000	17.0

SOURCE: *Solar Energy Research Institute*.

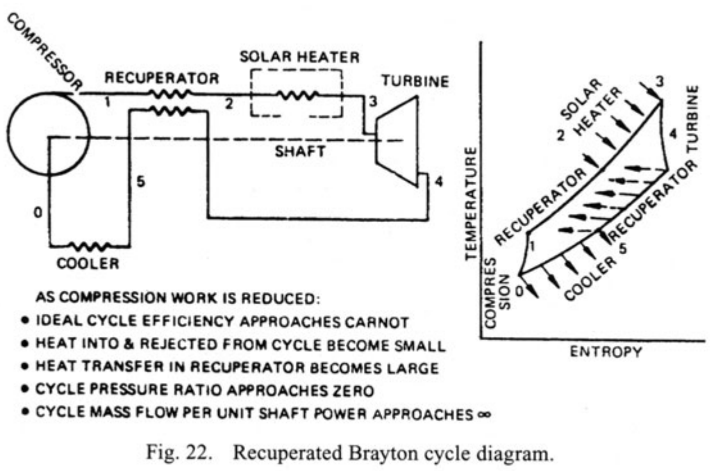

COMPRESSOR
RECUPERATOR
SOLAR HEATER
TURBINE
SHAFT
1
2
3
4
5
0
COOLER
AS COMPRESSION WORK IS REDUCED:
• IDEAL CYCLE EFFICIENCY APPROACHES CARNOT
• HEAT INTO & REJECTED FROM CYCLE BECOME SMALL
• HEAT TRANSFER IN RECUPERATOR BECOMES LARGE
• CYCLE PRESSURE RATIO APPROACHES ZERO
• CYCLE MASS FLOW PER UNIT SHAFT POWER APPROACHES ∞
TEMPERATURE
ENTROPY
SOLAR HEATER
RECUPERATOR
RECUPERATOR
TURBINE
COOLER
COMPRESSION

Fig. 22. Recuperated Brayton cycle diagram.

This type of approach is somewhat reminiscent of the chemistry of coal gasification. More detail is given in the Dostrovsky reference listed.

Solar Energy for Detoxifying Hazardous Chemicals. The Solar Energy Research Institute, Golden, Colorado, has developed a system for detoxifying hazardous chemicals in polluted groundwater. In essence, a photo-catalyst is added to the polluted water and then pumped through long, narrow glass tubes that are exposed to sunlight. High-energy photons activate the catalyst, which in turn breaks the pollutants down into nontoxic components. The tubes are mounted in reflecting glass troughs to improve the efficiency of the process. The system has proved particularly effective against trichloroethylene, once a common industrial cleaner. In this application, the polluted water is mixed with titanium dioxide catalyst. Hydroxyl radicals are created that break the offending solvent into water, carbon dioxide, and very dilute hydrochloric acid. The next step is that of determining how effective the proc-

TABLE 2. SOLAR FURNACE TIME TO REACH MELTING POINT OF MATERIALS (When exposed to absorbed solar flux of 20 MW/m^2)

Material	Melting point, T_m, °C	Time to reach T_m,[1] sec
Carbides		
TiC	3,200	9.14
NbC	3,500	0.86
ZrC	3,540	0.112
SiC	3,830	0.56
Metals		
Al	660	0.42
Cu	1,083	3.10
Ni	1,453	1.03
Steel	1,535	0.79
Ti	1,670	0.23
Cr	1,857	1.46
Mo	2,617	4.70
W	3,407	9.80
Nitrides		
Si_3N_4	1,900	0.059
AlN	2,200	1.320
BN	3,000	0.545
TiN	3,200	0.611
Oxides		
SiO_2	1,720	0.014
TiO_2	1,870	0.044
Al_2O_3	2,050	1.00
V_2O_3	2,410	0.107
CaO	2,580	1.66
HfO_2	2,780	0.188
MgO	2,800	2.59
ZrO_2	2,900	0.089

SOURCE: *Solar Energy Research Institute.*

ess may be in removing other chlorinated hydrocarbons, as well as such substances as benzene, various pesticides, and textile dyes.

Photovoltaic Conversion (Solar Cells)

Photovoltaic devices made of selenium have been known since the 19th Century. Pioneering research in semiconductors, which led to the invention of the transistor in 1947, formed the basis of the modern theory of photovoltaic performance. From this research, the silicon solar cell was the first known photovoltaic device that could convert a sufficient amount of the sun's energy to power complex electronic circuits. The conventional silicon cell is a solid-state device in which a junction is formed between single crystals of silicon separately doped with impurity atoms in order to create *n* (negative) regions and *p* (positive) regions which respectively are receptors to electrons and to "holes" (absence of electrons). See also **Semiconductors.** The first solar cell to be demonstrated occurred at Bell Laboratories (now AT&T Bell Laboratories) in Murray Hill, New Jersey in 1954.

In a photovoltaic device, the energy in light is transferred to electrons in the semiconductor when a photon collides with an atom in the material with enough energy to dislodge an electron from a fixed position in the material. A common technique for producing a voltage is by creating an abrupt discontinuity in the conductivity of the cell material (typically silicon) through the addition of dopants. A basic limit on the performance of these devices stems from the fact that light photons lacking the energy needed to lift electrons from the valence to the conduction bands ("band gap" energy) cannot contribute to photovoltaic current and from the fact that the energy given to electrons which exceeds the minimum excitation threshold cannot be recovered as useful electric current. Most of the photon energy not recovered as electricity is converted to thermal energy in the cell.

Photon energies in the visible light spectrum vary from 1.8 eV (deep red) to 3 eV (violet). In silicon, about 1.1 eV is needed to produce a photovoltaic electron; in gallium arsenide (GaAs), this is about 1.4 eV. Silicon is a comparatively poor absorber of light and consequently silicon cells must be from 100 to 200 micrometers thick to capture an acceptable fraction of the incident light. This places limitations on crystal grain size and thus, with present technology, single crystals must be used. Polycrystalline materials may alter this problem favorably and much research is being directed toward developing polycrystalline materials and, in general, for finding methods to minimize the impact of grain boundaries.

Thin films of gallium arsenide (GaAs) and cadmium sulfide/copper sulfide (CdS/Cu_2S) show potential because they are better absorbers of light and can be made thinner than crystalline silicon. Smaller crystal grains can be tolerated better than with crystalline silicon. See also **Thin Films.** These can be spray- or vapor-deposited, thus simplifying manufacturing. One possible drawback of the CdS and GaAs materials is their toxicity, particularly hazardous during manufacturing operations.

Where solar cells are used in *concentrated* sunlight, efficiency becomes of particular importance because of its effect upon total collector area needed, this being a major cost component of a solar energy system. A number of ingenious collector configurations have been developed. Further, there is the concept of the *thermophotovoltaic cell* which may be able to achieve efficiencies as high as 30–50% through shifting the spectrum of light reaching the cell to a range where most of the photons are close to the minimum excitation threshold for silicon cells. High efficiencies in intense radiation can be achieved, for example, with GaAs cells by covering them with a layer of $Ga_xAl_{1-x}As$, a material which reduces surface and contact losses. Clearly the interface between cell and solar radiation is of as great importance as development of new cell materials per se.

So-called *wet solar cells* show promise, particularly because of their relative ease of fabrication. In this type of photovoltaic cell, the junction is formed between a semiconductor and a liquid electrolyte. No doping is required because a junction forms spontaneously when a suitable semiconductor, such as GaAs, is contacted with a suitable electrolyte. Three knotty problems (accelerated oxidation of surface of semiconductor; exchange of ions between semiconductor and electrolyte forming a blocking layer; and deposition of ions of impurities on the surface of the semiconductor) all have been solved and thus the concept now appears technically viable.

Over a number of years, the photovoltaic cell developers received large financial incentives from the U.S. government. For example, the National Photovoltaics Act of 1978 was passed by the U.S. Congress, which authorized an expenditure of $1.5 billion for research, development, and demonstration of solar cell systems for converting sunlight into electric power. Also, in connection with the Federal Non-Nuclear Energy Research and Development Act of 1974, which established the concept of "net energy"—that is, the effect of new devices and systems on the overall energy balance. Projects were evaluated on the basis of their "potential for production of net energy." Although there have been some breakthroughs of particular significance to scientists and a gradually expanding market for photovoltaics in addition to use in space, particularly in various consumer products, the long awaited and ultimate application (generation of electric power in impressive amounts at competitive prices) has remained elusive. Scores of analyses have been made and forecasts range from very pessimistic to quite optimistic. The era of practically achieving this goal on the part of the photovoltaic cell community tends to be progressively shifted outward into the future. Forecasts usually are based upon numerous assumptions that are subject to periodic change and their reporting is best left to the periodicals and thus are not detailed here.

It is in order, however, to sum up the observations of the Electric Power Research Institute (EPRI): Photovoltaics need significant additional research to reduce cost and increase efficiency of the cells as well as their support systems (tracking and dc-to-ac power conversion) before they can be competitive with conventional electricity supply technologies. Current manufacturing costs for flat-panel arrays of interconnected, encapsulated cells are approximately $5000 per peak kW, and balance-of-plant costs double the effective system cost to $10,000 per peak kW. This compares roughly with $300 per kW for combustion turbines; $1400/kW for pulverized coal plants; and $2500/kW for nuclear plants. Two classes of photovoltaic converters that appear to show the most promise for producing large amounts of power are (1) the inexpensive, flat-plate, thin-film devices with target prices of less than $1500 per peak kW and efficiencies of 15%, compared with their current costs of $5000 per peak kW and efficiencies of about 10%; and (2) very high-efficiency, high-concentration devices with target prices less than \$1500 per peak kW and efficiencies of 25%, compared with their current costs of \$7000 per peak kW and efficiencies to utilities (largely subsidized or experimental programs).

Some authorities estimate that photovoltaic utility capacity could range from 0.6 to 16 GW by the year 2010, provided that needed technical performance is achieved.

Satellite Energy Collectors

Having proven their value in connection with relatively small space satellites, probes, etc., a huge satellite energy collector was first proposed in the late 1960s and, largely on the basis of national concerns with energy supplies precipitated by the oil embargo of the 1970s, considerable attention was given at the design level and in the literature to a solar power satellite (SPS). One proposal called for a space-based array requiring about 90 square kilometers (55 square miles)! That is about the size of Manhattan Island. The satellite would be in a geosynchronous orbit some 36,000 kilometers (22,000 miles) above Earth. Because nearly all authorities now consider such a project very "futuristic," no further details are reported here.

NOTE: To construct a perspective on the energy situation of the Early 1990s, some readers may wish to refer to the following articles in this encyclopedia: **Energy; Coal; Coal Conversion Processes; Electric Power Production and Distribution; Fuel Cells; Geothermal Energy; Hydroelectric Power; Hydrogen (Fuel); Natural Gas; Nuclear Power; Petroleum; Semiconductors; Tar Sands;** and **Tidal Energy.** See also alphabetical index.

Additional Reading

Becker, M., Editor: "Solar Thermal Central Receiver Systems," Springer-Verlag, New York, 1987.

Dostrovsky, I.: "Energy and the Missing Resource," Cambridge University Press, New York, 1988.

Dostrovsky, I.: "Chemical Fuels from the Sun," *Sci. Amer.*, 102 (December 1991).

Flood, D. J.: "Space Solar Cell Research," *Chem. Eng. Progress*, 62 (April 1989).

Gupta, B. P., Editor: "Solar Thermal Technology: Research and Development and Applications," *Proceedings of the Fourth International Symposium*, Albuquerque, New Mexico, 1990.

Holden, C.: "Sunlight Breaks Down Hazardous Chemicals," *Science*, 1215 (September 13, 1991).

Hubbard, H. M.: "Photovoltaics Today and Tomorrow," *Science*, 297 (April 21, 1989).

Stanley, J. T., Fields, C. L, and J. R. Pitts: "Surface Treating with Sunbeams," *Advanced Materials & Processes*, 16 (December 1990).

Waterbury, R. C.: "Solar Pump Delivers Remote Power," *InTech*, 74 (January 1990).

Wilson, H. G., MacCready, P. B., and C. R. Kyle: "Lessons of *Sunnyracer*," *Sci. Amer.*, 90 (March 1989).

Winter, Carl-Jochen, and J. Nitsch, Editors: "Hydrogen as an Energy Carrier: Technologies, Systems, Economy," Springer-Verlag, New York, 1988.

NOTE: For earlier references on solar energy, see prior edition of this encyclopedia.

SOLAR PARALLAX. The angle subtended by the equatorial radius of the earth at the distance of the sun is called the solar parallax. It is related directly to the astronomical unit which scales the solar system and is the baseline for the determination of stellar parallax.

One method for obtaining the solar parallax is to determine accurately the distance between the earth and another planet. Since their orbits are extremely well observed (on unit scale), knowing the separation at any given moment allows one to place an absolute value to the scale. The minor planet Eros has been used because of its close approach to Earth of only 26,000,000 kilometers and its point-like appearance on the photographic plate. Such observations have yielded a value for the solar parallax of

$$\pi = 8''.7984 \pm 0.0004$$

Radar observations of Venus accomplish the same purpose and yield a value for the astronomical unit of

$$\text{A.U.} = 149{,}598{,}640 \pm 250 \text{ kilometers}$$

which results in a solar parallax of

$$\pi = 8''.79414 \pm 0''.00002$$

and a lunar parallactic term in longitude of

$$P = 124''.987 \pm 0''.001$$

The radar observations are believed to be the better values by an order of magnitude. It is interesting to note that the results from lunar occultations lead to the value of

$$\pi = 8''.793 \pm 0''.003$$
$$P = 124''.97 \pm 0''.04$$

SOLENOIDAL. Applied to vector field having zero divergence, hence, one that may be expressed as the curl of another vector:

$$\mathbf{a} = \nabla \times \mathbf{b}$$

where **a** is the solenoidal vector and **b** is a vector field (sometimes called the vector potential of **a**), which can be determined from the differential equation. An equivalent definition of a solenoidal vector is one of which the integral over every reducible surface in its field is zero.

See also **Divergence (Mathematics).**

SOLENOID (Electrical). An electrically energized coil which may consist of one or more layers of windings. It is the basis of all forms of the electromagnet and is thus part of the operating mechanism of many electrically operated devices. One of the simplest forms and at the same time a widely used one is the plunger type solenoid. This is a coil wound on a non-magnetic form in which a magnetic plunger may move. Energizing the coil pulls the plunger up into the coil and thus operates the associated mechanism. The iron clad solenoid is similar except for an iron case surrounding the coil. This increases the magnetic pull on the plunger. Other types use a fixed core and various types of external armatures. Solenoids are widely used for operating circuit breakers, track switches, valves, and many other electromechanical devices.

SOLFATARA. A type of fumarole (volcanic), the gases of which are characteristically sulfurous. The *solfateric stage* is a late or decadent type of volcanic activity, characterized by sulfurous gases emitted from the vent. Also may refer to Etymol, the Solfatara volcano in Italy.

SOLID ANGLE. Consider a small cone with a base of area dS and a vertex at a fixed point P. This cone will cut out an area $d\sigma$ on a sphere of radius r with center at P. The solid angle subtended by dS at P is defined as $d\omega = d\sigma/r^2$. It is numerically equal to the area cut out by the same cone on a sphere of unit radius at the same point P. The unit used for measuring a solid angle is the steradian.

See also **Angle (Mathematics).**

SOLID (Geometry). A solid is a limited portion of space, bounded by a surface. A distinction should be made between the surface and the solid, thus a conical surface and a cone are not identical. A surface of revolution is a solid generated by the revolution of a plane area about a line, the axis of revolution. The study of solid properties is the main concern of solid geometry.

Among the figures which are considered are: anchor ring, cone, cylinder, ellipsoid, hyperboloid, paraboloid, parallelepiped, polyhedron, prism, prismatoid, pyramid, sphere.

See also **Conic Section;** and **Geometry.**

SOLID ROCKETS. See **Rocket Propellants.**

SOLID-STATE PHYSICS. The study of the physical properties (crystallographic, electrical and electronic, magnetic, acoustic, optical, thermal, mechanical, etc.) of substances in the solid phase.

In years past, much emphasis has been given to crystalline solids and this continues, but there has been a growing shift of interest to polymeric and amorphous substances as well. Much attention in the past has been given to metals and this also continues apace, but other substances are now under very serious investigation, including the ceramics, glasses, and organics. Interest in the solid state, of course, was given a tremendous boost by the discovery of semiconductors in the 1940s and, during the intervening years, this interest has been spurred by other electronic and electrical materials, including dielectrics, piezoelectrics, ferroelectrics, conductors and superconductors, electrodes, insulators, contacts, and polymers and macromolecular materials, notably those that are electroactive. Interest outside the electronics field, notably in the science of ceramics, glasses, and entirely new materials, such as composites, also has been adding to the body of knowledge of the solid state. However, because of the great need for solid materials with special properties for a host of applications, solid-state theory has tended to lag practice.

Nevertheless, solid-state theory has made excellent progress during the past decade. Just a few examples would include:

Excitonic Matter. The interaction of light with solid matter is a phenomenon of fundamental importance for exploring the quantum mechanics of materials. This field dates back to Einstein's finding that light energy is carried by quantized packets of radiation (photons). More recently, it has been found that a conduction electron can combine with a positively charged "hole" in a semiconductor to create an *exciton*, which, in turn, can form molecules and liquids. Some authorities consider the exciton as a new phase of matter. It was learned several years ago that the energy of incident photons can be converted inside a crystal into what might be termed short-lived neutral entities, i.e., excitons. As reported in an excellent paper by Wolfe and Mysyrowicz (1984), the exciton resembles the hydrogen atom. It consists of two oppositely charged carriers bound together by electrostatic attraction. In the hydrogen atom, the positive charge is a proton, which is surrounded by the negatively charged electron. In the exciton, the positive charge has a mass of an estimated $\frac{1}{1000}$th that of the proton. In the Wolfe/Mysyrowicz paper (details far beyond the scope of this encyclopedia), the investigators address several interesting questions. Can the exciton propagate freely through the crystal like a free hydrogen atom in a gas? Can two or more excitons combine to form a molecule? Can the excitonic "atoms" or the molecules made up of them form liquid or solid phases? Can more exotic phases of condensed excitonic matter come into being? How are excitons created by light in a crystal? Why does a crystal absorb light at all?

Electron Transport in Solids. It is well established (elucidated in several articles in this encyclopedia) that the production of integrated circuits (ICs) requires manufacturing techniques of extreme precision and sophistication. See **Microstructure Fabrication (Electronics).** The purity of materials used is also far higher than experienced by most other materials-processing industries. It has been observed by Howard, Jackel, Mankiewich, and Skocpol (AT&T Bell Laboratories) in a 1986 paper that a single-crystal silicon wafer 15 cm or more in diameter can be obtained with concentrations of undesired dopants at less than 1 part in 10 billion and with only about one defect per square centimeter. Accuracy in recent years is in terms of a few nanometers, and feature sizes in commercial circuits are down to 1 micrometer (micron) and getting smaller. Thus, it is no surprise that the silicon transistor can serve as a model for investigating numerous areas of the solid state. Using new patterning techniques, devices almost $\frac{1}{100}$th the size of commercial ICs can be made, making it possible to study transport physics in microstructures only a few hundred atoms across.

In 1985, two research institutions (IBM and AT&T Bell Laboratories) reported that electrons can travel through a semiconductor without being slowed by collisions (*ballistically*). The report was based upon experimental data showing a ballistic peak in the electron energy spectra of gallium arsenide (GaAs) test devices. This is reported in more detail in article on **Arsenic.**

Electroactive Polymers and Macromolecular Electronics. Electroactive polymers are of particular interest in connection with their use in fabricating improved electronic microstructures. Scientists at AT&T Bell Laboratories have been active in the investigation and development of electroactive polymers notably for electrodes. As reported by Chidsey and Murray (1986), electrodes can be coated with electrochemically reactive polymers in several microstructural formats called sandwich, array, bilayer, micro-, and ion-gate electrodes. These microstructures can be used to study the transport of electrons and ions through the polymers as a function of the polymer oxidation state, which is essential for understanding the conductivity properties of these new chemical materials. The microstructures also exhibit potentially useful electrical and optical responses, including current rectification, charge storage and amplification, electron-hole pair separation, and gates for ion flow. In their well illustrated paper, the investigators explore the three broad categories of electroactive polymers: (1) pi-conjugated, electronically conducting polymers; (2) polymers with covalently linked redox groups (redox polymers); and (3) ion-exchange polymers. In summary, the authors observe that although macromolecular electronics is still at a rudimentary level, the concepts involved are quite novel and with continued development may lead to practical applications. See also entry in this encyclopedia, **Molecular and Supermolecular Electronics.**

Quantized Hall Effect. In 1980, at the Max Planck Institute, Klaus von Klitzing discovered the quantized Hall effect, a phenomenon that occurs in certain semiconductor devices at low temperatures in very strong magnetic fields. As pointed out by Halperin (1986), the quantized Hall effect is observed in artificial structures known as two-dimensional electron systems. The conduction electrons in these systems are trapped in a very thin layer, such that the electronic motion perpendicular to the layer is frozen into its lowest quantum mechanical stage and thus plays no role in the conductivity of the device. In his experiment, Klitzing worked with a silicon field effect transistor (MOSFET). Electrons are trapped in what is called an inversion layer near the surface of a silicon crystal that is covered with a film of insulating silicon oxide, on top of which is deposited a metal gate electrode, used to control the density of conduction electrons in the inversion layer. This effect had been predicted as early as 1975 by Japanese investigators Ando, Matsumoto, and Uemura. Considerable detail pertaining to von Klitzing's experimental apparatus is given in the Halperin paper.

Surface Physics. Closely allied with solid state physics is the discipline of surface science. Investigations in this area have been quite intense during the past decade, notably in connection with catalysts. A catalyst is a species that changes the rate of a reaction and yet is regenerated by that reaction so that it seems to be unchanged in the net reaction. Although there are enzyme catalysts, for example, the majority of industrially interesting catalysts are found among the metals, the surfaces of which serve to catalyze reactions. The first catalytic phenomena were observed as early as 1835 by Berzelius and later better quantified by Ostwald in 1894. Aided by the great volume of catalysts used industrially ($ billions/year), the incentive for research is large. See articles on **Catalysis; Scanning Tunneling Microscope;** and **Silicon.**

Extremely High-Pressure Research. The invention and refinement of the modern diamond anvil cell (Carnegie Institution) occurred in the mid-1980s. This is a tool par excellence for optical, infrared and Raman spectroscopy and enables the researcher to study the changes in the electronic structure and chemical binding caused by the application of high pressure. Phase transitions, which involve changes in the atomic architecture can be determined with the diamond cell using the x-ray diffraction technique. Studies with the diamond anvil cell have been particularly valuable for obtaining geophysical information—for example, the state of silicate minerals and oxides in the mantle region right up to the core-mantle boundary to provide a view of the earth's interior, where high pressure and high temperature conditions exist. In solid-state physics, there is the fascinating challenge of making metallic hydrogen under ultrahigh pressure. This extraordinary change from a very good insulating to a metallic state in hydrogen is predicted to occur near 3 to 4 million atmospheres. See article on **Diamond Anvil High Pressure Cell.**

The foregoing examples are but a few to indicate the continuing vigorous research into the nature of the solid state. See also **Superconductivity.**

Concepts of Solids Simplified

The atoms which comprise a solid can be considered for many purposes to be hard balls which rest against each other in a regular repetitive pattern called the crystal structure. Most elements have relatively simple crystal structures of high symmetry, but many compounds have complex crystal structures of low symmetry. The determination of crystal structures, of atom location in the crystal, and of the dependence of many physical properties upon the inherent characteristics of the perfect solid is an absorbing study, one which has occupied the lives of numerous geologists, mineralogists, physicists, and other scientists for many years.

The rigid, hard-ball model is not adequate to explain many properties of solids. To begin with, solids can be deformed by finite forces, thus solids must not be completely rigid. Furthermore, atoms in a solid possess vibrational energy, so the atoms must not be precisely fixed to mathematically defined lattice points. This deformability of solids is built into the model by the assignment of deformable bonds (springs) between nearest atom neighbors. This ball-and-spring model has many successes; one important early use was that of Einstein to devise a reasonably successful theory of specific heats. Later incorporation by Debye of coupled motion of groups of atoms led to an even more successful theory.

Several measures exist of the strength of these bonds. One is the size of the elastic constants—for most solids, Young's modulus is about 10^{11} newtons per square meter. The other is the frequency of vibration of the atoms—values around 10^{13} to 10^{14} Hz are found.

The lack of perfection occasioned by elastic deformation of solids is but one of many kinds of crystalline imperfections. Defects are frequently found in crystals, produced in nature and in the laboratory. These defects may be characterized by three principal parameters—their geometry, size, and energy of formation.

All real crystals have atoms which occupy external surface sites and which do not possess the correct number of nearest neighbors as a consequence. Thus, a surface is a seat of energy and is characterized by surface tension. Furthermore, internal surfaces exist, grain boundaries and twin boundaries across which atoms are incorrectly positioned. In a crystal of reasonable size—say 1 cubic centimeter, these two-dimensional defects, called *surface defects*, contain only about 1 atom in 10^6, a rather small fraction. Even so, surfaces are important attributes of solids.

Some defects have extent in only one dimension—*line defects*. The most prominent of these, the dislocation, is a line in the crystal along which atoms have either an incorrect number of neighbors or neighbors which have not the correct distance or angle. In 1 cubic centimeter of a real crystal, one might find a wide variation of length of dislocations present—from near zero to perhaps 10^{11} centimeters.

Defects which have extent of only about an atomic diameter also exist in crystals—the *point defects*. Vacant lattice sites may occur—*vacancies*. Extra atoms—*interstitials*—may be inserted between regular crystal atoms. Atoms of the wrong chemical species—*impurities*—also may be present.

The properties of defects are intimately related to their energy of formation. A standard against which this energy can be compared is provided by the energy of sublimation—the energy necessary to separate the ions of a solid into neutral, noninteracting atoms. This energy is about 81,000 calories per mole for a typical metal, copper, at room temperature, about 3.5 eV per atom. Energies of surfaces, both free surfaces and grain boundaries, are about 1000 ergs per square centimeter, about 1 eV per surface atom. Dislocation energies are of similar size per atom length of dislocation, about 1 to 5 eV, so the energy of a dislocation is about $10-^4$ erg per centimeter of length. Point defects, too, possess an inherent energy of about 1 eV each. Vacancies in copper have an energy of about 1 eV; self-interstitials, 2 or 3 eV.

The energies per atom of these various defects, surface, line, and point, are all much larger than the average thermal energy per atom in a solid at reasonable temperatures. This thermal energy kT is only about $\frac{1}{40}$ eV at room temperature. Thus, defects can be produced only by conditions which exist during manufacturing (artificial and natural) by external means, such as plastic deformation or particle bombardment; or by large local fluctuations in thermal energy away from the average.

The total amount of energy which is bound up in ordinary concentrations of these defects is not large as compared to the total thermal energy of a solid at normal temperatures. All the vacancies in equilibrium in copper, even at the melting point, comprise less than 10 calories of energy per mole, much less than the enthalpy at 1357K (the melting point) of more than 7000 calories per mole. In a material with very heavy dislocation density, 10^{12} centimeters per cubic centimeter, the total dislocation energy is only a few calories per mole. And the total energy of a free surface of a compact block of 1 mole of copper is even less: about 10^{-3} calorie. Thus, the inherent energy of these defects is not large; even so they are immensely important in controlling many phenomena in crystals—as in the case of semiconductor devices.

Crystallographic defects need not remain stationary in the crystal; they may move about with time. Some of these movements may reduce the overall free energy of the solid; others (these are chiefly movement of the point defects) may simply be the wandering of random walk. Since these movements require larger than kT, the motion of defects depends upon rather large local fluctuations in energy. Consequently, their rate of motion depends upon temperature through a Boltzmann factor exp (— $\Delta H/RT$), where ΔH is the enthalpy increase necessary to move the defect from the lowest-energy site to the top of the barrier.

A convenient description of the crystalline structure of solids is thus seen to consist of successive stages of approximation. First, the mathematically perfect geometrical model is described; then departures from this perfect regularity are permitted. The deformability of solids is allowed for by letting the force constants between adjacent atoms be finite, not infinite. Then, misplacement of atoms is permitted and a variety of crystalline irregularities, called defects, is described. Some of these defects have intrinsic features which affect properties of the crystal; other affect the properties by their motion from site to site in the crystal. In spite of their relatively small number, defects are of immense importance.

Electronic Structure of Solids. In principle, the electronic structure of solids is determined by the electronic structure of the *free atoms* of which the solid is composed. Since the free atom structure is known rather well, especially for atoms of lower atomic number, the electronic structure of solids should be subject to determination by calculation. This is not the case. A wide variety of interactions occur between the electrons on adjacent atoms as they approach the equilibrium distance characteristic of solids. These interactions are of such complex nature that they tend to defy concise definition and involve such a host of charged particles, electrons, and ion cores, that only approximate calculations can usually be made. Nevertheless, the use of approximate models allows many general features of the electronic structure to be deduced, especially when close interplay between theory and experiment is established. As for the crystalline structure of solids, two stages are useful in understanding the electronic structure. First, the perfect electronic structure is defined. Then, irregularities in this structure, again termed defects, are described. Although both the geometry and energy of crystalline defects are defined, description of the geometry of the charge distribution of many of the electron defects is difficult and one must generally be content with description of the formation energy of the defect.

The nuclei of the atoms in a solid and the inner electrons form ion cores with energy levels little different from corresponding levels in free atoms. The characteristics of the valance electrons are modified greatly, however. The state functions of these outer electrons greatly overlap those of neighboring atoms and restrictions of the Pauli Exclusion Principle and the Uncertainty Principle force modification of the state functions and development of a set of split energy levels becomes a quasi-continuous band of levels of width which are several electron volts for most solids. Importantly, unoccupied levels of the atoms are also split into bands. The electronic characteristics of solids are determined by the relative position in energy of the occupied and unoccupied levels as well as by the characteristics of the electrons within a band.

Metals. The solid is called a *metal* if excitation of electrons from the highest filled levels to the lowest unoccupied levels can occur with infinitesimal expenditure of energy. Thus, excitation can occur by means of many external forces, such as electric fields, heat, light, radio waves. Metals are, therefore, good conductors of electricity and of heat; they are opaque to light and they reflect radio waves.

Insulators. Some solids have wide spacing between the occupied and the unoccupied energy states—2 eV or more. Such solids are called *insulators* since normal electric fields cannot cause extensive motion of the electrons. Examples are diamond, sodium chloride, sulfur, quartz, mica. They are poor conductors of electricity and heat and are usually transparent to light (when not filled with impurities or defects).

Semiconductors. Solids with conductivity properties intermediate between those of metals and insulators are called *semiconductors*. For them, the excitation energy lies in the range 0.1 to about 2eV. Thermal fluctuations are sufficient to excite a small, but significant, fraction of electrons from the occupied levels (the valence band) into the unoccupied levels (the conductance band). Both the excited electrons and the empty states in the valence band (aptly called *holes*) may move under the influence of an electric field, providing a means for conduction of current. Such electron-hole pairs may be produced not only by thermal energy, but also by incident light, providing photo-effects. The inverse process, emission of light by annihilation of electrons and holes in suitably prepared materials, provides a highly efficient light source (example, light-emitting diodes).

Crystallographic defects, in general, are also electronic defects. In metals, they provide scattering centers for electrons, increasing the resistance to charge flow. The resistance wire in many electric heaters consists of an ordinary metal, such as iron with additional alloying elements such as nickel or chromium providing scattering centers for electrons. In semiconductors and insulators, however, alloying elements and defects provide an even greater variety of effects, since they can change the electron-hole concentrations drastically in addition to providing scattering centers. This is the basis of semiconductory technology. See also **Semi-conductors.**

Interactions of Solids with Light. Solids are useful because of their interaction with external forces or stimuli, such as electric and magnetic fields, heat, and mechanical forces. Yet among these interactions, probably the most important is the interrelation between matter and light. This interaction, important to all photosynthetic phenomena and the production of food; to the artificial generation of light; to the use of phosphors in cathode-ray tubes—is also the basis of spectroscopy and its use in the study of solids. In this field, first came investigation of emission and absorption of radiation from free atoms. Later investigations included emission and absorption of radiation by atoms in solids—giving rise to maser and laser phenomena, Mossbauer spectroscopy, nuclear magnetic resonance, x-ray diffraction, infrared spectroscopy, fluorescence, the Raman effect, microwave emission and absorption, among many other useful effects.

Band Theory of Solids

The success of the simple free electron theory of metals was so striking that it was natural to ask how the same ideas could be applied to other types of solids, such as semiconductors and insulators. The basic assumption of the free electron theory is that the atoms may be stripped of their outer electrons, the resulting ions arranged in the crystalline lattice, and the electrons then poured into the space between.

The free electron model results from the neglect of the interaction of the various atoms and of the periodic variation of the potential in which the electrons move, i.e., as their distance from the nearest metallic ion changes. When the former is taken into account, it is found that each energy eigenstate of an isolated atom is split into N non-degenerate states, where N is the number of atoms in the crystal. The group of levels that result from a single atomic state form an *allowed band*. If we start from the free electron picture and consider the effect of the periodic variations of potential, the Bloch theorem leads to the conclusion that there will be discontinuities in the plot of energy vs. momentum whenever the wave vector $\mathbf{k}$ has magnitude and direction such that it satisfies the Bragg law for reflection, in which λ may be set equal to $\mathbf{1/k}$ to give $\mathbf{k} \cdot \mathbf{d} = n$. Here $\mathbf{k}$ is the wave vector, $\mathbf{d}$ is the vector separation of two atomic planes in the crystal and n is an integer equal to the scalar product. As with the atomic interaction model, the number of eigenstates between two energy breaks is equal to the number of atoms in the crystal. Thus either approach leads to the existence of a manifold of energy levels occurring in groups of N closely spaced levels, the groups being separated by energies that are often very large compared with the spacing of levels within a group, somewhat as shown in Fig. 1. Each group of levels is known as an *allowed band*; the energies between groups are said to be in a *forbidden band*. Because these levels depend on the properties of the body as a whole, the entire macroscopic crystal may be considered to be a single giant molecule. The electrical, mechanical, and thermal properties of the crystal are then largely determined by the electrons in the energy levels within the highest occupied bands.

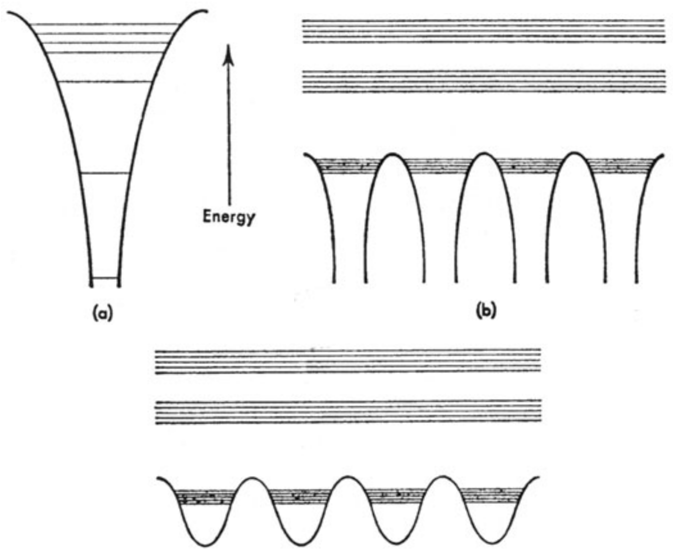


Fig. 1. Origin of the energy levels in a crystalline solid. The curves represent potential energy versus distance. At (a), the potential energy is that of an isolated ion; the energy levels, represented by the horizontal lines, are sharp. At (b), the overlap of the fields of the ions lowers the potential energy curve between the atomic positions and results in a splitting of each atomic level into a band of allowed levels. At (c), the model is derived from one in which the electrons are free, subject only to a periodic potential resulting from the ionic fields.

Because electrons obey the Pauli Exclusion Principle, not more than two of them (with oppositely directed spin) can exist in any single energy level. In thermal equilibrium at the zero of absolute temperature, than, all of the levels up to some particular energy, determined by the number of electrons present, will be occupied and all above this energy will be vacant. This highest level is known as the *Fermi level*. At higher temperatures there will not be a sharp discontinuity in occupancy—some of the levels below the Fermi level will be vacant and some above it will be occupied. The *Fermi level* is then defined as the energy of the state which has a 50% chance of being occupied.

The Fermi level is determined by the number of electrons present, and the properties of the material are therefore dependent on whether this energy falls near the bottom, top, or middle of an allowed band. If the number of electrons is such as to exactly fill certain bands, with a wide gap above them, the material will be an insulator (m). If the gap is very narrow, or if there are impurities present to create extra levels, the substance may be semiconducting (o). (See Fig. 2.) In these cases, it is difficult to supply sufficient thermal energy to an electron to promote it into the conduction band above the gap, where alone it is free to carry an electric current. In a metal (n), however, there is always a partially filled band, in which the electrons behave in many respects as if they were free. The existence of the partially filled band may be due either to the fact that each atom contains an odd number of electrons or to the overlapping of two allowed bands, each of which will be partly filled. Direct evidence for the existence of bands is provided by the soft x-ray emission spectra, but the importance of the theory is not so much its correctness in detail as the simplicity of the band scheme by which the energy relations between various phenomena may be shown on a single diagram.

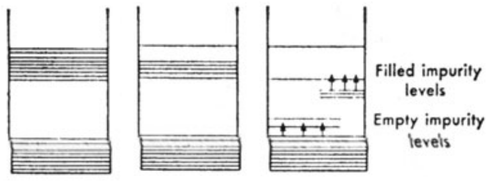


Fig. 2. Band diagrams of *m*, insulator; *n*, metal; and *o*, semiconductor.

Additional Reading

Bate, R. T.: "The Quantum-Effect Device: Tomorrow's Transistor?" *Sci. Amer.*, 96 (March 1988).

Blakely, J. M.: "Surfaces and Interfaces," in *Encyclopedia of Materials Science and Engineering* (M. B. Bever, Ed.), MIT Press, Cambridge, Massachusetts, 1986.

Bokor, J.: "Ultrafast Dynamics at Semiconductor and Metal Surfaces," *Science*, 1130 (December 1, 1989).

Brodsky, M. H.: "Progress in Gallium Arsenide Semiconductors," *Sci. Amer.*, 68 (February 1990).

Caruana, C. M.: "The Interdisciplinary Approach to Surface Science," *Chem. Eng. Progress*, 64 (July 1987).

Chidsey, C. E. D., and R. W. Murray: "Electroactive Polymers and Macromolecular Electronics," *Science*, **231**, 25–31 (1986).

Chin, G. Y.: "Magnetic Materials," in *Encyclopedia of Materials Science and Engineering* (M. B. Bever, Ed.), MIT Press, Cambridge, Massachusetts, 1986.

DeShazer, L. G.: "Optical Materials," in *Encyclopedia of Materials Science and Engineering* (M. B. Bever, Ed.), MIT Press, Cambridge, Massachusetts, 1986.

DiSalvo, F. J.: "Solid-State Chemistry: A Rediscovered Chemical Frontier," *Science*, 649 (February 9, 1990).

Fisk, Z., et al.: "Heavy-Electron Metals: New Highly Correlated States of Matter," *Science*, 33 (January 1, 1988).

Halperin, B. I.: "The 1985 Noble Prize in Physics (Quantized Hall Effect)," *Science*, **231**, 820–822 (1986).

Heiblum, M., and L. F. Eastman: "Ballistic Electrons in Semiconductors," *Sci. Amer.*, 102–111 (February 1987).

Howard, R. E., et al.: "Electrons in Silicon Microstructures," *Science*, **231**, 346–349 (1986).

Karasz, F. E., and T. S. Ellis: "Polymers: Structure, Properties, and Structure-Property Relations," in *Encyclopedia of Materials Science and Engineering* (M. B. Bever, Ed.), MIT Press, Cambridge, Massachusetts, 1986.

Landman, U., et al.: "Atomistic Mechanisms and Dynamics of Adhesion, Nanoindentation, and Fracture," *Science*, 454 (April 27, 1990).

LeComber, P. G.: "Amorphous Silicon—Electronics into the 21st Century," *University of Wales Review*, 31 (Spring 1988).

Lovinger, A. J.: "Ferroelectric Polymers," *Science*, **220**, 1116–1121 (1983).

Mott, N. F., and E. A. Davis: "Electronic Processes in Non-Crystalline Materials," Oxford Univer. Press, New York, 1979.

Pool R.: "Clusters: Strange Morsels of Matter," *Science*, 1184 (June 8, 1990).

Pool, R.: "A Transistor That Works Electron by Electron," *Science*, 629 (August 10, 1990).

Prinz, G. A.: "Hybrid Ferromagnetic Semiconductor Structures," *Science*, 1092 (November 23, 1990).

Williams, E. D., and N. C. Bartelt: "Thermodynamics of Surface Morphology," *Science*, 393 (January 25, 1991).

Wolfe, J. P.: "Thermodynamics of Excitons," *Physics Today*, **35**(12), 46–54 (March 1982).

Wolfe, J. P., and A. Mysyrowicz: "Excitonic Matter," *Sci. Amer.*, 98–107 (March 1984).

Yablonovitch, E.: "The Chemistry of Solid-State Electronics," *Science*, 347 (October 20, 1989).

SOLIDUS CURVE. A curve representing the equilibrium between the solid phase and the liquid phase in a condensed system of two components. The relationship is reduced to a two-dimensional curve by disregarding the influence of the vapor phase. The points on the solidus curve are obtained by plotting the temperature at which the last of the liquid phase solidifies, against the composition, usually in terms of the percentage composition of one of the two components.

SOLION. A small electrochemical oxidation-reduction cell consisting of a small cylinder containing a solution and divided into sections by platinum gauze, porous ceramics, or other materials. A type of solion for detecting sound waves consists of a potassium iodide-iodine solution in which the iodide ions are oxidized to triiodide ions at the anode, and the reverse process occurs at the cathode. The cell is constructed so that the sound waves cause agitation of the solution between the electrodes, and thus change the current. In addition to detection of sound, solions can be designed to detect changes in other conditions, such as temperature, pressure, and acceleration.

SOLSTICE. Either of the two points on the sun's apparent annual path, where it is displaced farthest, north or south, from the earth's equator, i.e., a point of greatest deviation of the ecliptic from the celestial equator.

The point north of the celestial equator is termed the summer solstice; the point south of the equator, the winter solstice, inasmuch as the sun is at these respective points at the commencement of summer and winter in the northern hemisphere.

SOLUBILITY. A property of a substance by virtue of which it forms mixtures with other substances which are chemically and physically homogeneous throughout. The degree of solubility is the concentration of a solute in a saturated solution at any given temperature. The degree of solubility of most substances increases with a rise in temperature, but there are cases (notably the organic salts of calcium) where a substance is more soluble in cold than in hot solvents.

SOLUBILITY CURVE. The graph showing the variation with temperature of the concentration by a substance in its saturated solution in a solvent.

SOLUBILITY PRODUCT. A numerical quantity dependent upon the temperature and the solvent, characteristic of electrolytes. It is the product of the concentrations of ions in a saturated solution and defines the degree of solubility of the substance. When the product of the ion concentrations exceeds the solubility product, precipitation commonly results. Strictly speaking, the product of the activities of the ions should be used to determine the solubility product, but in many cases the results obtained using concentrations, as suggested by Nernst, are correct.

SOLVENT. The term solvent generally denotes a liquid which dissolves another compound to form a homogeneous liquid mixture in one phase. More broadly, the term is used to mean that component of a liquid, gaseous, or solid mixture which is present in excess over all other components of the system. A *chemical solvent* is the term used for solvents in those instances where the process of solution is attended by a chemical reaction between the solvent and the solute. In contrast, a *physical solvent* is one that does not react with the solute. A *dissociating solvent* is one in which solutes that associate in many other solvents enter into solution as single molecules. For instance, various carboxylic acids associate and thus give abnormal elevations of the boiling point, abnormal depressions of the freezing point, etc., in many organic solvents; but in water, however, they do not associate. For this reason water is called a dissociating solvent for such solutes. A liquid that dissolves or extracts a substance from solution in another solvent without itself being very soluble in that other solvent is termed an *immiscible solvent*. A solvent whose constituent molecules do not possess permanent dipole moments and which do not form ionized solutions is termed a *nonpolar solvent*. *Polar solvents*, on the other hand, consist of polar molecules, that is, molecules that exert local electrical forces. In such solvents, acids, bases, and salts, that is, electrolytes, in general, dissociate into ions and form electrically conducting solutions. Water, ammonia, and sulfur dioxide are typical polar solvents. A *normal solvent* is one which does not undergo chemical association, namely, the formation of complexes between its molecules.

A *leveling solvent* is a solvent in which the acidity or basicity of a solute is limited (or leveled) by the acidity or basicity of the solvent itself. For example, the strongest acid which can exist in water is oxonium ion, H_3O^+. Consequently, even though HCl (for example) is intrinsically a much stronger acid than H_3O^+, its acidity in aqueous solution is "leveled" to that of H_3O^+ through the reaction $HCl + H_2O \rightleftarrows H_3O^+ + Cl^-$. Likewise the very strong base KNH_2 is leveled in water to the basicity of OH^-

$$KNH_2 + H_2O \rightleftarrows K^+ + OH^- + NH_3$$

The solvents which are leveling to both acids and bases are self-ionized solvents, e.g., water, ammonia, alcohols, carboxylic acids, nitric acid, etc. Basic non-protonic solvents are leveling to acids, but not to bases (i.e., they are differentiating toward bases), e.g., pyridine, ethers, ketones, etc., since the strongest acid attainable is the protonated solvent molecule (e.g., $C_5H_5N + HCl \rightleftarrows C_5H_5NH^+ + Cl^-$), whereas there is no corresponding basic species derived from the solvent. Though solvents leveling to bases but not to acids are in principle much more difficult to find, in practice, very strong acids like H_2SO_4 and $HClO_4$ are limiting to bases because the species HSO_4^- and ClO_4^-, which will be formed by almost any basic substance, are the strongest bases attainable in these solvents—$B^- + HClO_4 \rightleftarrows HB + ClO_4^-$—whereas practically no other acid is capable of producing the cations $H_3SO_4^+$ and $H_2ClO_4^+$ in these solvents (i.e., they are differentiating toward acids).

Differentiating solvents are solvents in which neither the acidity of acids nor the basicity of bases is limited by the nature of the solvent. These solvents are not self-ionized. The aliphatic hydrocarbons and the halogenated hydrocarbons are such solvents.

In industry it is generally understood that solvents are simple or complex, pure or impure, compounds or mixtures of compounds (either natural or synthetic) which dissolve many water-insoluble products like fats, waxes, resins, etc., forming homogeneous solutions; that such organic solvents dissolve these water-insoluble products in various proportions depending on the solvent power of the solvent, the degree of solubility of the solute, and the temperature; and that the solute can be recovered with its original properties by the removal of the solvent from the solution. It is also understood in industry that there is a much more limited number of solvents which do not have the properties given above but which nevertheless are of considerable importance; they are the inorganic solvents like water, liquid ammonia, liquid metals, and the like.

Solvents have been classified on various arbitrary bases: (1) boiling point, (2) evaporation rate, (3) polarity, (4) industrial applications, (5) chemical composition, (6) proton donor and proton acceptor relationships, and (7) behavior toward a dye, Magdala Red. Thus on the basis of industrial application one can classify solvents as those for (1) acetyl-cellulose, (2) pyroxylin, (3) resins and rubber, (4) cellulose ether, (5) chlorinated rubber, (6) synthetic resins, and (7) solvents and blending agents for cellulose ester lacquers. Solvents classified according to chemical composition are noted below.

The term solvent action is understood to mean any process of making substances water-soluble; but in a broader interpretation the term is understood to be the phenomenon of making a substance soluble in a solvent. Solvent power, diluting power, solvency and similar expressions indicate the property of solvents to disperse the molecules of a solute or vehicle thereby causing a decrease in viscosity.

The most common solvent is water. Water dissolves a great many gases, liquids, and solids, and is much used for this purpose. Other liquids similarly dissolve many substances without reacting chemically with them. Important considerations in connection with the choice of solvent for a given case are (1) vapor pressure and boiling point, (2) solvent power under stated conditions of temperature, (3) ease and completeness of recoverability by evaporation and condensation, and completeness of separation from dissolved material by evaporation, (4) heat of vaporization, (5) miscibility with water or other liquid, if present, (6) inertness to chemical reaction with the materials present, and with the apparatus, (7) inflammability and explosiveness, (8) odor and toxicity; (9) cost of solvent, loss in process, cost of recovering.

See also **Pollution (Air).**

Colligative Properties of Solutions. When solute is added to a pure solvent, thus forming a solution, properties of the solvent are altered, including (1) osmotic pressure; (2) vapor pressure (lowered); (3) melting point (lowered); and (4) boiling point (elevated). These properties bear a relationship to the number of solute molecules in solution and not to the nature of the molecules. These phenomena are explained by enhanced tension in the solvent. Complete explanation of these changes is beyond the scope of this book, but reference is suggested to H. T. Hammel's article on "Colligative Properties of a Solution" (*Science*, **192,** 748–756, 1976).

SOLVOLYSIS. A generalized conception of the relation between a solvent and a solute (i.e., a relation between two components of a single-phase homogeneous system) whereby new compounds are produced. In most instances, the solvent molecule donates a proton to, or accepts a proton from a molecule of solute, or both, forming one or more different molecules. A particular case of special interest occurs when water is used as solvent, in which case the interaction between solute and solvent is called *hydrolysis.*

SOMATOPLASM. The tissue of an organism exclusive of the reproductive cells (germ plasm). The somatoplasm of animals typically has the diploid chromosome number, and the reproductive cells have the haploid chromosome number.

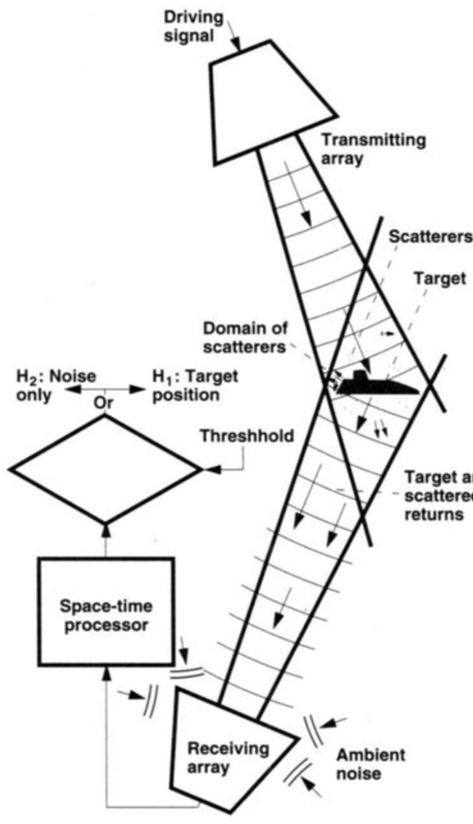
SOMITE. One of the longitudinal series of segments into which the bodies of many animals are divided. These segments are clearly shown in a simple form in the earthworms. In humans, they are made evident by the structure of the spinal column and the series of spinal nerves, but they are overshadowed externally by the high development of the appendages. The term is synonymous with metamere.

The segmental masses of mesoderm in the vertebrate embryo are also called somites. They are the primordia of the axial skeleton, voluntary muscles of the body and appendages, and the inner layer of the skin.

SONAR. A coined word derived from the phrase, "sound navigation and ranging." The term generally refers to the principles employed in the design and operation of systems that utilize acoustic energy transmitted in an ocean medium; while the systems themselves are referred to as sonar systems. Thus, sonar may be defined as a branch of applied acoustics concerned with the utilization of the ocean as the transmitting medium.

The problem of sonar is threefold: (a) understanding the transmission of acoustic energy through the transmitting medium, (b) developing sources which convert mechanical or electrical energy into acoustic energy, and (c) developing receivers which convert the acoustic energy back into mechanical or electrical energy. See accompanying illustration.

Whenever a body vibrates in a fluid, longitudinal waves are formed, which propagate outward from the vibrating body. The particles of the



General operating principle of sonar detection system. (*U.S. Naval Underwater Systems Center, Newport, Rhode Island.*)

fluid are set in motion, and temporary stresses are produced which increase and decrease during each vibration. The motion of the particles gives the fluid kinetic energy while the stresses induce potential energy. The sum of the two energies is called acoustic energy.

Traditionally, the starting point for a discussion of the transmission of acoustic energy in a fluid is to assume a point source radiating acoustic energy in an ideal homogeneous nonabsorptive medium of infinite extent. Under these assumptions, the energy from the source will radiate outwards with the wave front forming a spherical shell. As the radius of the shell increases, the sound intensity decreases. In practice it is customary to express the sound intensity by means of a logarithmic scale. The most generally used logarithmic scale is the decibel. The intensity level in decibels of a sound of intensity I is defined as $10 \log(I/I_0)$ where I_0 is a reference intensity. The intensity level can also be expressed as $20 \log(P/P_0)$ where P is the pressure and P_0 the reference pressure, usually 1 dyne/cm^2 in underwater acoustics. In this discussion the terms in all equations are expressed in decibels. The decrease in intensity as the shell increases in radius is called the spreading loss. The spreading loss from a unit range of R_0 to a range of R is 10 log [(intensity at R_0)/ (intensity at R)] $= 10 \log(R^2/1^2) = 20 \log R$. In most applications the use of such a simple model has been inadequate.

A more realistic model considers the following factors: the water-earth interface (bottom), the water-atmosphere interface (surface), the absorption of acoustic energy in the medium, the presence of foreign material in the medium, and the distribution of sound velocity. Considered as an acoustic medium, the waters of the ocean form a thin layer on the earth's surface. Some of the acoustic energy radiated into this layer by a source will reach either the surface or the bottom. At either of these surfaces abrupt discontinuities in acoustic properties occur. Because of these discontinuities part of the intercepted energy is reflected, part may be transmitted across the interface, and part may be scattered within the medium. Since the transmission of an acoustic wave in water is accompanied by a compression and expansion of the medium, friction will occur between water molecules. This friction results in the conversion of some of the acoustic energy into thermal energy. In addition to this frictional, or viscous, loss there is another loss of energy in seawater related to the salts which continuously undergo chemical changes because of pressure fluctuations. Energy losses associated with both of these phenomena are called absorption losses. Due to the presence of foreign bodies in the volume of water, reflection and scattering is not limited to the surface and bottom boundaries. Foreign matter and biological content vary widely in size and acoustic characteristics. All ocean waters contain such bodies which modify the direction in which the acoustic energy is transmitted. In sufficient number they may also modify and increase the total absorption loss. The effect of variations in sound velocity is to bend the wave front in the direction of the lower velocity. This bending of the wave front is referred to as refraction. Both refraction and reflection can result in the guiding of acoustic energy in certain directions.

The factors above affect the propagation of acoustic energy in seawater in two different ways. The first results in a spreading loss already mentioned, and the second results in a loss referred to as attenuation. Attenuation consists of both the scattering and absorption losses. The spreading and attenuation are related to the distance the acoustic energy travels in different ways. An important difference is that the spreading loss frequently is relatively independent of frequency while the attenuation is a function of frequency.

There are three basic types of sonar systems: direct listening systems, echo-ranging systems, and communication systems.

In direct listening the acoustic energy is radiated by the target, which is the primary source. The acoustic transmission is a one-way process. In their more elementary forms direct listening sonar systems may be nondirectional and only give a warning that a primary source is in the vicinity of the searching vehicle; or directional, and permit determination of the bearing of individual primary sources relative to the listening platform. They generally do not give range. Direct listening is limited by the magnitude of the signal when it reaches the receiving point and the magnitude of the interfering noise which tends to obscure its reception.

In echo ranging, the sonar system projects acoustic energy into the water with the expectation that this energy will strike a target and enough of the energy will be reflected back to the searching platform so that it can be recognized as a target echo. The primary source of acoustic energy is in the searching platform, with the target, upon reflection of the energy, becoming a secondary acoustic source. The transmission of the energy is a two-way process. Echo-ranging sonar systems permit a determination of the bearing of a silent target, and by timing the echo-signal transmission and by knowing the velocity of sound in seawater, a range may also be obtained. Echo ranging is limited by the relative magnitudes of the signal and of the locally generated interference. In some cases the sonar performance is limited by reverberation, which is the acoustic energy returning by reflectors other than the target of interest.

Acoustically, sonar communications systems are similar to direct listening systems in that they utilize a one-way transmission path. Instrumentally, they are similar to echo-ranging systems, one located at each of the two points between which communications is to be established. In these systems coded pulses or voice modulated signals are transmitted by one system and received by the other.

To hear a target by direct listening, it is necessary that the acoustic level of the target less the transmission loss along the acoustic path from the target to the listening equipment be equal to or greater than the level of the background noise. This may be expressed as $L - H \geq N$ where L is the source level of the target, H is the one-way transmission loss, and N is the noise level. The size of this inequality depends upon operator skill, signal processing, and method of presentation. It is called the signal excess, E. This inequality can be written as an equation where $E = L - H - N$. This equation is called the direct-listening sonar equation for an omnidirectional listening hydrophone. When using directional hydrophones a factor called the directivity index must be added to the right-hand member of the equation. The source level, L, is a measure of the amount of acoustic energy put into the water by the target vehicle and is equal to 10 log (sound intensity at unit distance from the source). The transmission loss, H, is the sum of the losses related to refraction, surface and bottom reflection, absorption, and scattering. The noise, N, results from unwanted acoustic energies arriving from many different sources and normally consists of thermal, ambient, and self noises.

To see a target by echo ranging it is necessary that the acoustic level of the primary source less twice the transmission loss along the acoustic path from source to target plus the target strength be equal to or greater than the noise. This may be expressed, for a nondirectional receiver against a noise background, as $L - 2H + T \geq N$ where T is the target strength, a function of the reflecting characteristics of the target. Against a reverberation background the inequality becomes $L - 2H + T \geq R$ where R is the reverberation level. As in the case of direct listening, the inequalities can be expressed in terms of the signal excess as $E_N = L - 2H + T - N$ or $E_R = L - 2H + T - R$ where E_N and E_R are the signal excesses for noise and reverberation. The noise, N, comes from own-ship's noise and target noise. Reverberation, R, is the energy that is returned from the outgoing acoustic energy to the receiving equipment after having been reflected from reflectors in the medium other than the target. Reverberation sources usually are backscattering from the surface, bottom, and foreign particles in the water.

Echolocating bats (*Eptesicus fuscus*) are capable of detecting changes as short as 500 nanoseconds in the arrival time of sonar echoes when these changes appear as jitter or alternations in arrival time from one echo to the next. As pointed out by Simmons (1979), the psychophysical function relating the bat's performance to the magnitude of the jitter corresponds to the half-wave rectified cross-correlation function between the emitted sonar signals and the echoes. The bat perceives the phase or period structure of the sounds, which cover the 25- to 100-kilohertz frequency range. The acoustic image of a sonar target is apparently derived from time-domain or periodicity information processing by the nervous system. The biological sonar of bats in the suborder Microchiroptera is of much interest because bat sonar represents a rather well-defined example of a biological communications system.

For use of sonar in medicine, see **Electrocardiography;** in oceanography, see **Ocean;** and **Ocean Research Vessels;** in petroleum exploration, **Petroleum;** in photography, **Photography and Imagery.** Also consult alphabetical index.

SORGHUM (*Andropogon Sorghum; Gramineae*). Sorghums are annual grasses of tropical origin. They have an extensive system of wiry roots and solid stems 3–15 feet (0.9—45 meters) tall. The leaves are smaller than those of corn, and capable of rolling up tightly during periods of drought, and quickly unrolling and starting to function when favorable moisture conditions return. Because of this habit sorghums are often grown in regions subject to frequent drought. The inflorescence is a panicle usually of very compact habit. Ordinarily the spikelets are paired, one of the pair being sessile or stemless, the other having a short stem or pedicel. The former is fertile, the latter staminate. The grains are enclosed in the glumes and vary considerably in shape in different varieties. There are two main groups of sorghums; the sweet or saccharine sorghums, the juicy pitch of which is a source of syrup; and the grain sorghums, which yield grain, stock food and ensilage. Kaffir is one of the latter group. In Asia, grain sorghums are employed in a countless variety of ways, as for fuel, brooms, mats, fences, windbreaks, roof thatch, and in making a fermented drink. In the United States, grain sorghums are finding increasing popularity in several industrial processes, as a source of starch, wax, and other products. See Figs. 1 and 2. See also **Broomcorn.**

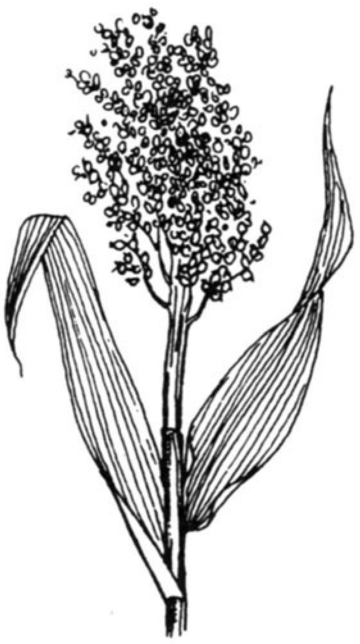


Fig. 1. Head of Schrock grain sorghum. (*USDA diagram.*)

Fig. 2. Sorghum residue being shredded for return to the soil to add organic matter and increase fertility. (*USDA photo.*)

SORPTION. A generalized term for the many phenomena commonly included under the terms adsorption and absorption when the nature of the phenomenon involved in a particular case is unknown or indefinite.

SOUTH ATLANTIC CENTRAL WATER. An oceanic water mass, extending at the surface roughly from southern Africa to Patagonia, where the Subantarctic Water is on the surface. This region of junction is south of the subtropical convergence in the South Atlantic. Temperature range 6–18°C (42.8–64.4°F), salinity range, 34.5%–36.0%.

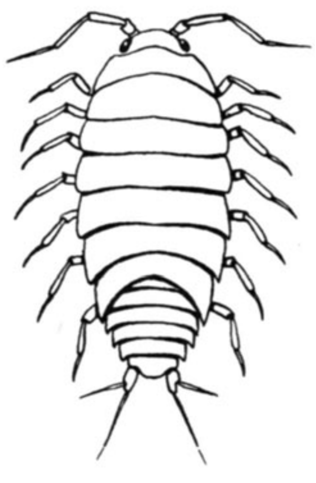
SOUTH ATLANTIC CURRENT. An eastward flowing current of the South Atlantic Ocean that is continuous with the northern edge of the Antarctic Circumpolar Water.

SOUTH EQUATORIAL CURRENT. In the Atlantic and Pacific Oceans, the westward counterclockwise drift of surface water in the Equatorial latitudes south of the Equator that is separated from the North Equatorial Current by the Equatorial countercurrent. Driven westward by the trade winds, these great currents are deflected by the land masses and also by the Coriolis effect. In the Southern hemisphere they are deflected to their left. Upon turning poleward, the current becomes a part of the west-wind drift, a product of the prevailing westerlies.

SOUTH PACIFIC CENTRAL WATER. Due to the great size of the Pacific Ocean, it contains more well-developed oceanic water masses than the Atlantic. Thus there are both an Eastern and Western South Pacific Central Water mass. Extending from the Pacific Equatorial Water on the north to the Subantarctic Water on the south, and covering the width of the Ocean, except for a transition zone on the eastern side, they differ from each other in properties to a greater degree than the Eastern and Western masses in the North Pacific because the Western South Central Water is closely similar to the Central Water of the Indian Ocean, having a temperature range of 8–15°C (46.4–59°F) and the salinity of 34.6–35.5%. The Eastern South Pacific Water has a temperature range of 10–18°C (50–64.4°F) and the salinity of 34.0–35.0%.

SOUTH PACIFIC CURRENT. An eastward flowing current of the South Pacific Ocean that is continuous with the northern edge of the Antarctic Circumpolar Water.

SOWBUG (*Crustacea, Isopoda*). Related to crayfish, the sowbug is a crustacean and not an insect. The sowbug is a common but not severely damaging pest on food crops. The pest is found most often in connection with glasshouse or home gardening operations. The sowbug does feed on the tender portions of some plants and also attacks mushrooms. The sowbug (*Porcellio leavis*, Koch) reproduces by way of eggs, but the eggs remain in a marsupium for approximately 2 months. The pest is slow in maturing, requiring nearly a year to become a fully-developed adult. The sowbug is a notable pest of watercress.


Sowbug. (*USDA*.)

The *pillbug (Armadillidium vulgate*, Latrielle), is closely related to the sowbug and is of similar habit. The pillbug is distinguished by its ability to curl up into nearly a perfect spheroidal form when disturbed.

SOYBEAN. Of the family *Leguminosae*, subfamily *Papilionaceae*, and the genus *Glycine max*, the soybean is a typical legume seed differing in color, size, and shape, depending upon variety. The common field varieties grown in the United States are nearly spherical and are yellow in color, (e.g., the Lincoln soybean). See accompanying figure. The food reserves of the seed are stored in the cotyledons (90% seed), the interior of which is filled with elongated palisade-like cells, themselves filled with proteins and oil. The bulk of the protein is stored in protein bodies varying between 2 μm (microns or micrometers) and 20 μm in diameter. The oil is located in smaller structures, the spherosomes, 0.2 μm to 0.5 μm, interspersed between the protein bodies. Isolated protein bodies may contain as much as 90% protein and together account for 60 to 70% total protein of defatted soy flour. Proximate analysis of the whole bean shows that protein (40%) and oil (21%) make up about 60% of the bean, the remaining third consisting of nonstarch carbohydrates, including polysaccharides, stachyose (3.8%), raffinose (1.1%), and sucrose. Nucleic acids are present only as minor constituents of the soybean, unlike some of the new single-cell protein sources, where nucleic acids can account for between 5 and 20% cell weight and may be included within the total nitrogen value of the isolate. Other minor constituents, such as the antitrypsin factor and so-called "beany flavor" components, have caused some adverse comment against the use of soybean products in the human dietary. Moreover, every legume (bean, pea, field bean) also has the flatus problem, but in processing this problem is not insurmountable.

Recently pulled soybean plant lying on table for examination. Root structure is at right. Mature pods drop from vines. (USDA photo.)

Importance of the Soybean. During the past couple of decades, a major source of protein, namely that derived from the soybean, has developed. In times when there is concern over the ability of the world to feed its growing population, soybean protein technology is one of several scientific tools for disproving the Malthusian prophecy.

In terms of present knowledge, soybeans are capable of producing the greatest amount of protein per unit of land of any major plant or animal source used as food by people today. See accompanying table.

COMPARISON OF PROTEIN-YIELDING COMMODITIES

Protein Source	Kilograms from 1 Hectare	Pounds from 1 Acre
Beef cattle	65	58
Wheat	202	180
Maize (corn)	362	323
Soybean	560	500

In considering the present importance and future potential of the soybean, there are at least two factors of merit: (1) Large numbers of people in the lesser developed areas of the world suffer from a protein deficiency in their diets, leading to serious health problems; and (2) in the more developed areas, many people are accustomed to meat, milk, and eggs and desire to expand consumption of these products at lower costs. Many countires have the capacity to expand livestock and poultry production and, in the process, to lower prices of these products. This potential accounts for much of the market growth opportunity for soybeans and other feedstuffs. But, even in areas where livestock and poultry development may come more slowly, there is the opportunity to expand protein consumption in other ways. Soy protein provides a means of both extending the animal proteins and of replacing them with high-quality protein that is relatively inexpensive to produce-thus, the opportunity to further utilize soybeans directly as human food.

A technical factor of large significance is the general excellence of soybean protein as a protein. Soybean protein has a high content of essential amino acids, particularly lysine, leucine, and isoleucine. Detailed analysis is given later. Only in the sulfur-containing amino acids, cystine and methionine, is the soybean low, indicating that methionine is the first limiting amino acid to be considered when using soy products in a diet. Full fat soy flour also offers a valuable nutritional contribution from the high proportion of essential fatty acids, linoleic (51%) and linolenic (9%), that its oil contains. Since the oil represents some 20% of the total flour, the use of soy flour as a supplement in high-protein breads also elevates the essential fatty acid content of the product.

The soybean originated in the northern provinces of China and was first described in about 2000 B.C. as one of the most important cultivated legumes and one of the five sacred grains essential to Chinese civilization. The soybean was used as a basic food and a source of medicinals from the Middle Ages until the early 1700s, when greater interest was shown in the legume. Two major developments occurred at that time: (1) The first extraction of oil from the soybean, which led to the development of a new industry in Manchuria; and (2) introduction of the soybean to Europe in 1712 and the United States in 1804.

SOYBEAN OIL. See **Vegetable Oils (Edible).**

SOY PROTEIN. See **Protein.**

SPACE CHARGE. The electric charge on a conductor is to be regarded as confined to an infinitely thin layer at the surface and thus, in a sense, as geometrically two-dimensional. Even when a current is flowing through the conductor, the quantities of positive and negative electricity within any element of volume at any instant are equal, thus neutralizing each other's electrostatic effect. For this reason, Laplace's equation applies to the interior of a current-bearing conductor as well as to a complete vacuum. This is true whether the conductor be of the metallic or the electrolytic type.

Quite different are the conditions in a vacuum tube, where streams of electrons or of positive ions (e.g., canal rays) may occupy sizable regions to the virtual exclusion of carriers of the opposite sign. (See **Ionized Gases.**) The electricity thus monopolizing such a region is called a "space charge," and may exert marked influence on the performance of the tube.

A typical instance of this occurs in thermionic rectifiers. If the cathode is emitting no electrons, the potential gradient across from cathode to anode is nearly uniform; that is, the potential increases by the same number of volts for each centimeter from the one electrode to the other. If now a small emission begins from the cathode (because of its being heated), the electron swarm is more dense near the cathode, due to their accelerated motion, much as the stream of water descending from a faucet is broader near the faucet than farther down. The potential gradient is no longer uniform, and the total increase of potential, that is, the "plate voltage," is less than before. Further increase in emission may result in an actual drop of potential near the cathode (followed by a rise farther on), with still further decrease of plate potential. Such a potential minimum is contrary to the condition expressed by Laplace's equation, which therefore does not hold in a region occupied by a space charge. (Another formula, called Poisson's equation, here applies instead. It is $\nabla^2 V = 4\pi\rho$, where V is the potential and ρ is the electric volume density at the point in question.)

In gas-filled discharge tubes the effect of electronic space charge is largely offset by the presence of great numbers of less mobile positive ions. As a result, such tubes not only have much lower resistance but have a more uniform potential gradient and give a better approximation to Ohm's law.

In atmospheric electricity, space charge refers to a preponderance of either negative or positive ions within any given portion of the atmosphere.

SPACE CHARGE LIMITATION OF CURRENTS. It has been shown by Child that the current between a plane cathode and a parallel plane anode at a distance d from it, when the anode potential is V, cannot exceed a certain maximum value, determined by the modification of the electric field near the cathode as a result of the space charge of electrons in that region. If the electrons leave the cathode with zero speed, the maximum current per unit area of the cathode is

$$i = \frac{4\epsilon_0}{9}\sqrt{\frac{2e}{m}}\frac{V^{3/2}}{d^2}$$

where e and m are the electronic charge and mass, respectively, and ϵ_0 is the electric permittivity constant in any self-consistent system of units. Langmuir has extended the equation to include the case of a cylindrical cathode of radius a, surrounded by a coaxial cylindrical anode of radius b. The maximum current per unit length is then

$$i = \frac{8\pi\epsilon_0}{9}\sqrt{\frac{2e}{m}}\frac{V^{3/2}}{b(\ln b/a)^2}$$

The dependence of the current on the $\frac{3}{2}$ power of the potential difference is general, and is the basis of the definition of perveance.

SPACE FRAME (or Space Structure). A three-dimensional framed structure. It may be composed of triangles, rectangles or a combination of these forms. Space frames are statically determinate or indeterminate depending on the number of members, support conditions and the rigidity of the joints. Bridges, framed domes, transmission towers, radio towers and building frames are all space structures.

If the space frame is statically determinate it may be analyzed by means of the equations for three-dimensional statics. These equations state that the summation of all forces parallel to three axes, usually taken as mutually perpendicular, must equal zero and the moments of all forces about these axes must also equal zero.

Indeterminate space frames may be analyzed by methods which are applicable to indeterminate planar structures. In recent years numerical methods have been developed for the analysis of rigid space frames. Bridges and building frames are highly indeterminate when considered as space frames because the joints have a certain amount of rigidity even though they may not have been designed to resist moment. Thus, all parts do participate to some extent in distributing the loads to the foundation. Due to this action some members may be subjected to torsional stress as well as direct and bending stresses.

SPACE (Minkowski). A flat space of four dimensions, of which three specify the position (x, y, z) of a point in space, and the fourth represents the time t at which an event occurs at that point. Usually, the coordinates in the space are denoted by x_μ ($\mu = 1, 2, 3, 4$) with $x_1 = x$, $x_2 = y$, $x_3 = z$, $x_4 = ict$, where $i^2 = -1$. It is also possible to write $x_4 = ct$, but then it is necessary to associate with the space the metric g^{ij}, where $g^{11} = g^{22} = g^{33} = 1$, $g^{44} = -1$.

SPACE-TIME. A space of four dimensions which specify the space and time coordinates of an event. In the absence of a gravitational field, space-time reduces to Minkowski space.

See also **Space (Minkowski); Gravitation;** and **Relativity and Relativity Theory.**

SPACE VEHICLE GUIDANCE AND CONTROL. The term guidance means the sum total of the orders and instructions which a space vehicle must be given to keep it on the chosen path to its objective. They may be transmitted from earth upon the basis of the course and position of the craft as signaled or observed, they may be given by instruments on board, or, in the case of manned space vehicles, they may originate, partly at least, from a human pilot. Then, on the basis of this information, the control system operates devices which change a course as necessary.

Obviously, the guidance program for a particular mission begins with the path through space which the space vehicle is to follow. That path may be suborbital, such as that taken by sounding rockets, which return to earth relatively promptly; it may be orbital, as in the case of satellites, which go into a closed orbit about the earth; or it may leave the earth's gravitational field to extend into space, as in the case of space probes and corresponding manned spacecraft. Of the paths followed by these various craft, the ideal, circular orbit is the most simple to calculate.

Consider a satellite launched from earth which, when it reaches the height of its projected orbit, is turned into a horizontal position and given a horizontal velocity (*injected*) into its orbit. What must this velocity be for a stable orbit, i.e., a satellite that is to continue circling the earth at that distance?

If we ignore certain complications, such as the fact that the earth is not perfectly spherical and its mass is not uniformly distributed, and also the effect of the drag upon the satellite of the atmosphere, corresponding to its density at that altitude, and so consider only the height and injection velocity of the satellite, the calculation becomes relatively simple. From Newton's First Law (see **Newton's Laws of Dynamics**) we know that if no force acted upon the satellite it would continue its horizontal motion (i.e., motion tangent to its orbit) and travel away into space. On the other hand, if it were not moving, it would fall straight to earth due to the acceleration of gravity. Since both effects are present, its velocity is the resultant of the injection (horizontal) velocity and that due to gravity, and there is obviously a value of the former at which the orbit is stable. This value is given by the expression

$$v_H = \sqrt{\frac{g_0 R_0^2}{R_0 + h}} \tag{1}$$

where g_0 is the acceleration of gravity at the earth's surface (32.16 ft-sec^{-2}), R_0 is the radius of the earth, and h is the height of the satellite above the surface of the earth. Since g_0 and R_0 are almost constant, this equation shows that the greater the height of a satellite above the surface of the earth, the smaller the injection velocity necessary to stabilize it there.

If the injection velocity does not have the value calculated for that height by Equation (1), then the orbit will not be circular. If the injection velocity is greater than the calculated value of v_H but not greater than $\sqrt{2}v_H$, the orbit will be an ellipse. If the injection velocity equals or exceeds $\sqrt{2}v_H$, then the satellite will cease to orbit about the earth, and move away into space. If the injection velocity is less than the calculated value of v_H, the satellite will fall toward earth along a curved path, the closer the velocity to v_H, the longer the path. In fact, if the injection point is at a sufficient height and the injection velocity is not too far below the value of v_H calculated by Equation (1), this path will not intersect the earth, but will take the form of an ellipse. It should also be noted that even if the injection velocity should have the exact value calculated by Equation (1), and if the injection is in a direction above or below the horizontal, the orbit will be an ellipse.

In view of these circumstances, it is not surprising that the satellites that have been launched have elliptical orbits. Further reasons are the facts that the mass distribution of the earth is not uniformly spherical and that the atmosphere exerts a drag upon satellites, especially those that approach the earth closely. Neither of these effects were taken into account in the derivation of Equation (1).

The path to be followed by a probe or manned spacecraft that is to leave the gravitational field of the earth to travel to such objects as the moon or one of the planets presents a more complicated problem than an ideal satellite. The difference is due partly to the fact that the destination is also a moving object, and its motion must be included in the calculations.

The first step, the escape velocity from the earth's gravitational field, as already noted in the discussion of satellites, is $\sqrt{2}v_H$, or by substituting into Equation (1),

$$v_E = \sqrt{2}\sqrt{\frac{g_0R_0^2}{R_0 + h}} = \sqrt{\frac{2g_0R_0^2}{R_0 + h}} \qquad (2)$$

If the starting point is at the earth's surface, the term h drops out of the equation, and the value so calculated is about 7 miles per second (neglecting the drag of the atmosphere). The first requirement for space journeys from earth is that the craft attain a velocity exceeding vE. Note that if the point of departure is at a height h above the earth's surface, such as would be provided by a space platform or by a spacecraft already in orbit, the value of v_E is reduced accordingly.

After the problem of escaping earth's effective gravitational field is solved, there are several ways of calculating the best path to another planet. We can calculate the path that minimizes the time, which obviously requires a greater expenditure of energy than a path which minimizes the energy. The latter calculation was made by W. Hohmann on the simplifying assumption of circular planetary orbits and the results, which are ellipses, are called Hohmann transfer ellipses. They show that for minimum energy, the spacecraft should leave the one planet tangentially to its orbit and land tangentially to the orbit of the second planet. Of course, this simple picture must be considerably modified in plotting actual interplanetary paths, and the result is usually a compromise between minimum-energy and minimum-time paths.

Having determined the path which the space vehicle is to follow, the next step is to determine the guidance necessary to keep it on that path. For convenience in planning, the guidance program is often divided into three stages: initial guidance (prior to orbital injection), mid-course guidance, and terminal guidance. Note that initial guidance applies not only to satellites, but also to probes and manned spacecraft that are to leave the earth's gravitational field, since they are commonly injected into an orbit from which they then leave for the moon or another planet.

Methods of guidance include: (1) *Command Guidance*, in which the space vehicle is tracked from earth, and commands are sent to it as necessary. Since radio (or radar) is used to send these instructions, as well as to receive positional data from the space vehicle, the method is also called *radio guidance*. (2) *Inertial guidance*, in which all guidance operations are carried out by instruments aboard the space vehicle. In a combination of (1) and (2), compensating adjustments to the inertial guidance system are sent from earth, the method being called *radio-inertial guidance*. (3) *Celestial guidance*, operated by celestial navigation methods applied automatically by instruments aboard the spacecraft.

As to the first method, it should be noted that, although it is called radio guidance, the frequencies actually used are those in the radar range. These signals are used not only for control systems, but also for determining the distance of the vehicle and its velocity component away from the observer. Then the trajectory of the vehicle is adjusted to change its orientation, which is done by controls that turn it about its three axes. The essential elements in these controls are gyroscopes, which are controlled by radar signals when used in a command guidance system.

In the second method, inertial guidance gyroscopes play an even more important part. They are used in the stable platform system which is set before the vehicle is launched and which provides a reference for the attitude of the vehicle throughout its flight. They are used as integrating devices to convert to velocity readings the acceleration measurements made by a number of devices, including spring-mass accelerometers, vibrating-element accelerometers, and pendulum accelerometers, which measure acceleration by its effect upon, respectively, a spring-held mass, an element in vibration, or the motion of the pendulum. Similarly, the velocity indications so obtained are integrated again to obtain the distance traveled. See also **Acceleration Measurement.**

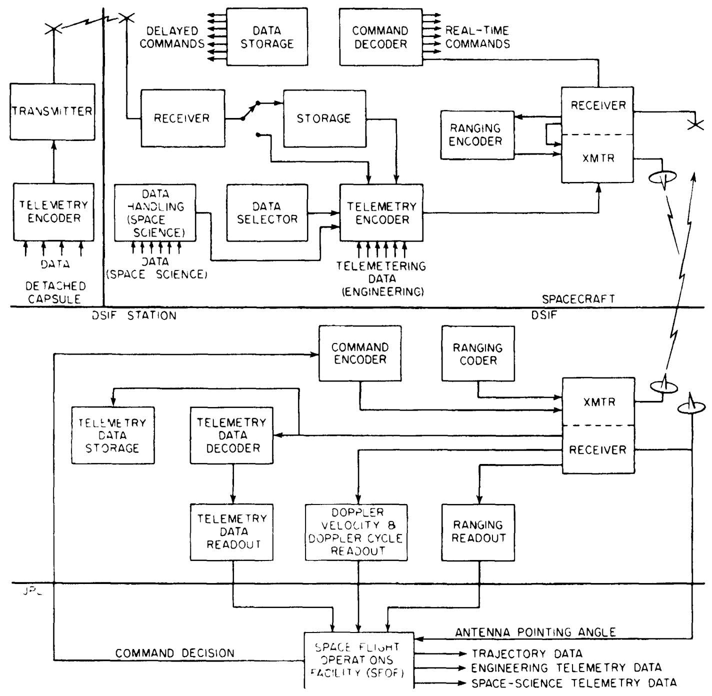
Since gyroscopes have a tendency to drift, the combined radioinertial guidance system sends command instructions to correct these errors. This is done by a radar communication hookup in which the space vehicle receives continuously its range data (as observed from earth) from an earth-based station. These observed range data signals are then fed into the vehicle's guidance system to correct instrumental errors that may have developed by such means as gyroscope drift.

Obviously some of the guidance functions described up to this point, such as changes in the altitude of the space vehicle, or even the close adjustment of gyroscopes, are characteristic of the early part of a mission, i.e., initial guidance. In fact, mid-course guidance is rarely used on satellites or other relatively short missions, but is important chiefly on missions to the moon or farther. Due to the far greater distances, the radio command method must be modified. If doppler shift measurements are used, a two-way system is necessary, in which the radar signals from earth actuate a transmitter on the spacecraft that amplifies and transmits the signals back.

In addition to doppler measurements, the time for signal transmission is determined and the phase difference between the earth signal and the spacecraft signal is found by interferometer methods using radio telescopes. With these various data, the distance, angular position, and radial velocity of the spacecraft can be calculated. In view of the amount of calculation necessary in mid-course guidance, it is not surprising that computers are a prominent feature of these operations. See accompanying diagram.

Astronomical observations from earth can be plotted to establish the path of the spacecraft. Whatever method of path measurement is used, corrections to it are made by first actuating gyroscopes to turn the craft to the correct attitude, and then signaling a rocket aboard it to fire.

Terminal guidance begins with the correction of the trajectory of the spacecraft, satellite, or vehicle to completion of its mission. The vehicle continues as necessary up to the end, whether that be a landing on return to earth, or a landing on or orbiting of a planet or the moon. For landing on earth, the path to be followed is easier to calculate in advance and to control, since the properties of the earth's atmosphere are known, and the attitude and position of the vehicle can be observed from earth. The velocity of the vehicle in its path depends upon its position and attitude at the commencement of the terminal stage, the acceleration of gravity, and the drag of the atmosphere, which also has a heating effect that must be held within limits. The function of guidance is to modify the effect of these variables by adjusting the attitude by means of gyroscopes and small rockets, as discussed earlier in this entry for initial guidance, and then to decrease the velocity by means of the larger retrorockets.



Fundamentals of space-probe communication subsystem. SFOF = space flight operations facility; DSIF = deep space instrumentation facility.

The space vehicle control system must act in concert with the guidance system. The chief controlled variables are attitude and velocity, the latter including both speed and direction of motion. The extent to which these variables are controlled varies with the kind of space vehicle and the stage of its mission, i.e., it differs for the initial, midcourse, and terminal stages.

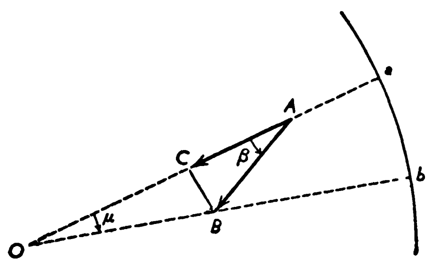
Attitude control can be effected during the initial (or launch) phase by changing the direction of the thrust of the rocket vehicle. The thrust is the force accelerating the vehicle in a forward direction, due to reaction to the opposite force exerted by the stream of gases leaving the rear of the vehicle. Under stable conditions of motion, the direction of the thrust is along the central axis of the vehicle. In one method of turning the vehicle, jet vanes, which are placed in the exhaust gas stream, can be moved so as to deflect the stream in various directions, thus exerting a turning moment which continues until the direction of the central axis of the vehicle corresponds to the new direction of thrust. These vanes may be moved by electric motors, or hydraulic or pneumatic systems which are actuated by the guidance system. Similar in its action to the jet vanes is the jetavator, which may be mounted in the exhaust gas stream, but consists of a collar-shaped member which is free to rotate when moved by an actuator. Since it rotates about only one axis, three or more of these devices may be used. Other methods turn the space vehicle by turning the nozzle itself, which is mounted on a swivel. In smaller vehicles, the entire motor may be attached to a hinge, so that it is free to turn. Such methods are not usually applicable to large rockets; they are usually turned by small (vernier) rocket motors. Since the control system must be capable of turning the space vehicle as much as is required about all three of the space axes, there may be six such motors mounted on swivels. Often these motors will operate on a medium different from that of the main rocket engine. Compressed gases or liquid monopropellants, such as hydrogen peroxide and hydrazine can be used. Another method of rocket attitude control is by means of a jet of noncombustible liquid that is discharged into the exhaust gas stream to change its direction.

Since velocity has direction as well as magnitude (speed), attitude control also controls the direction of the velocity. Control of the magnitude of velocity is of particular importance during the terminal stage of some missions, chiefly for landings, but also for rendezvous missions, or for missions which have orbits about a body as their destinations. Retrorockets may be used to effect slow-downs for such maneuvers. See also **Rocket Propellants.**

See also **Astronautics.**

SPACE VELOCITY (Stellar). The space velocity of a star is its actual velocity in space relative to the sun. The term *fixed star* is one which has been handed down to us from the era when philosophers believed that the stars were fixed on a sphere, rotating about the earth, which they referred to as the celestial sphere. At the present time we believe that practically all of the stars are actually in motion in space with velocities comparable in magnitude to the velocities with which the planets are moving about the sun in their orbits. In measuring any velocity it is necessary to have a definite reference point and the space velocity of a star is its velocity referred to the sun.

In the figure we have the space velocity of a star represented by the vector (directed straight line proportional to the velocity in direction and magnitude) *AB*. The angle β which this vector makes with the direction of the star from the sun at any particular instant gives the direction of the space velocity at that instant.



Vector representation of space velocity.

As observed from the solar system this space velocity may be resolved into two components: *AC* the radial velocity, and μ the proper motion. The radial velocity is determined directly in terms of linear velocity (i.e., in miles or kilometers per second), but the proper motion may only be determined in angular units, usually expressed in seconds of arc per year. In order that the space velocity may be known, the proper motion must be converted into a linear velocity, commonly known as the transverse velocity of the star. This may be accomplished only if the distance of the star is known. Expressing the distance in terms of the stellar parallax, π, and calling the transverse velocity *T* (the line *CB* in the figure), the following relations may be derived:

$$T = 4.74\ \mu/\pi \text{ kilometers/second}$$

or

$$T = 2.94\ \mu/\pi \text{ miles/second}$$

With both the transverse velocity, *T*, and the radial velocity, *R*, known in the same units, the problem of determining the space velocity *S* in the same units is merely that of solving the plane right-triangle *ACB*. This solution yields

$$S^2 = T^2 + R^2 \qquad \cos\beta = R/S \qquad \sin\beta = T/S$$

SPAR. In marine parlance, a spar is a round timber used to extend a sail. Used with this meaning, it could be a mast, a boom, or a yard.

A structural member used similarly to extend a surface to obtain an air reaction is found in the wing of an airplane. There, the spar is a principal structural member running the length of the wing. Usually there are two, parallel to one another, but wings have been built having only one spar. In such a design the spar is required to take torsion and bending as well as compression if the wing is externally braced. The spars support the ribs upon which the wing covering is stretched. The reaction of the air upon the latter is transmitted to the spars which in turn are attached to the fuselage. The air reaction may be carried by the spars entirely by bending as in the case of a truly cantilever wing, or it may be carried by spars which are braced outboard of the fuselage by wires or struts.

SPARK CHAMBER. A device used for the study of high-energy nuclear reactions and related particle physics. The device consists of a chamber filled with an inert gas, often helium or helium and neon, in which there is a stack of parallel conducting plates. These plates are connected alternately to the positive and negative terminals of a high-voltage source (10,000 volts upward). A chamber may have from 25 to 100 plates, each about 1 millimeter thick and 1 millimeter square, spaced about 0.6 centimeter apart. The chamber is operated in conjunction with a counter used as a detector for ionizing particles entering the chamber. The counter is used to connect the high-voltage circuit momentarily to the plates, and to trigger the shutter of a camera when a particle enters the chamber. As a particle traverses the chamber, it produces many ion-pairs along its path, so that the gas becomes conducting, and sparks occur between the plates in the region through which the ionizing particles pass. The light from them is focused by the lens system so that stereoscopic photographs may be obtained of the events. See also **Bubble Chamber; Cloud Chamber;** and **Particles (Subatomic).**

SPARROW (*Aves, Passeriformes*). A common form of small seedeating bird (Aves) of the family *Fringillidae*. The many species are for the greater part rather quietly colored in browns and grays with streaked and spotted plumage, but some bear conspicuous black or white marks and the browns of some species are very bright. Some of the sparrows have beautiful songs, although none rival such outstanding singers as the brown thrasher and the mockingbird.

The field sparrow, *Spizella pusilla*, chipping sparrow, *S. passerina,* white-throated, *Zonotrichia albicollis*, and white-crowned, *Z. leucophrys*, sparrows, song sparrow, *Melospiza melodia*, and in the west the lark sparrow, *Chondestes grammacus*, are among the well known North American species.

SPAWNING. Reproduction and the bringing forth of young in fishes. In the cases of many species of fishes, the reproductive process is quite simple in that the male sheds his sperm and the female sheds her eggs in adjacent water wherein the mingling of these two materials causes fertilization. But, in numerous species, the process is varied and often reasonably complex. Most fishes have particular preferences for what might be termed idealized spawning conditions for their particular species. These conditions include water temperature, water salinity (fresh versus marine waters), and geographic location. The relative stillness of the water also plays a role. Because temperature, in particular, varies with season, the spawning habits usually have a very marked seasonal influence. This factor alone can cause long migrations of fishes to their desired water conditions, but salinity essentially does not change with seasons and, therefore, migrations from marine waters to fresh waters (or vice versa) is a major factor. These water conditions not only affect spawning, but more importantly they affect the survival of the fertilized eggs and the early development of the young. The eggs, which usually float on the surface of the water with many species, are subject to all manner of danger. In compensation for the great degree of chance involved in survival of the early life-giving processes, most species produce eggs in astronomical numbers. The eggs sometimes number into the several millions. For this reason, in many species, the ovaries are extensive, for example comprising about one-fifth of the body weight of a female salmon. At a single spawning, the cod routinely deposits from 4 to 6 million eggs. Most fish eggs have a yolk, upon which the embryo feeds during incubation and often for a period after hatching. Then the very tiny mouth of the tiny fish commences to consume plankton. In the very early stages of life, the tiny fish frequently is referred to as a *larva* or *fry*. It is interesting to note that the larva forms of many species appear quite different in form from the adult fish, although a larval cod, for example, leaves no question but that it is a very small cod.

Another protective means occurs in the species of some fishes by way of the phenomenon of changes in sex and hermaphroditism. Some species are always self-fertilizing.

The oviparous (egg-laying) species often exhibit unusual ways of depositing and protecting their eggs: burrowing in a sandy bottom; attaching eggs by way of sticky substances to vegetation; depositing eggs by drilling into the bodies or shells of other creatures where the eggs are left for safekeeping and development. An unusual adaptation is found in the male Australian kurtus where eggs are incubated in a pouch on the male's forehead. Some species deposit their eggs in moist places on land, where they may remain for a few minutes or few days. The grunion is very precise about time of year, phase of moon, and tide, depositing its eggs in wet sand whereupon they are immediately fertilized by the male, then within a few minutes are hatched and carried back to sea by the waves. See **Silversides (Osteichthyes).**

The spawning habits of several species are described briefly in this volume. Refer to list of entries under **Fishes.**

There are numerous species where the young are born alive. These are termed ovoviviparous, the incubation and hatching of the eggs occurring within the body of the female. The young in these instances usually require careful attention and monitoring of one or both parents. In the discus, for example, the young "nurse" upon fluids and mucus present on the sides of both male and female parent. The parents take turns in this procedure.

SPECIFIC ACTIVITY. Three common uses of this term are: 1. The activity of a radio element per unit weight of element present in the sample. 2. The activity per unit mass of a pure radionuclide. 3. The activity per unit weight of any sample of radioactive material. Specific activity is commonly given in a wide variety of units (e.g., millicuries per gram, disintegrations per second per milligram, counts per minute per milligram, etc.). See **Radioactivity.**

SPECIFIC GRAVITY. For a given liquid, the specific gravity may be defined as the ratio of the density of the liquid to the density of water. Because the density of water varies, particularly with changes in temperature, the temperature of the water to which a specific gravity measurement is referred should be stated. In exacting, scientific observations, the reference may be to pure (double-distilled) water at 4°C (39.2°F). In engineering practice, the reference frequently is to pure water at 15.6°C (60°F). A value of unity is established for water. Thus, liquids with a specific gravity less than 1 are lighter than water; those with a specific gravity greater than 1 are heavier than water. From a practical standpoint, it usually is more meaningful to express the specific gravity of gases with reference to pure air rather than to pure water. Thus, for a given gas, the specific gravity may be defined as the ratio of the density of the gas to the density of air. Since the density of air varies markedly with both temperature and pressure, exacting observations should reflect both conditions. Common reference conditions are 0°C and 1 atmosphere pressure (760 torr; 760 millimeters Hg; 29.92 inches Hg).

Specific Gravity Scales. Arising essentially from a lack of communication between various scientific and industrial communities, a number of different specific gravity scales were formulated in earlier times and, because so much data and experience have been accumulated in terms of these scales, several methods of expressing specific gravity persist in common use. The most important of these scales are defined here.

API Scale—This scale was selected in 1921 by the American Petroleum Institute, the U.S. Bureau of Mines, and the National Bureau of Standards (Washington, D.C.) as the standard for petroleum products in the United States.

$$\text{Degrees hydrometer scale (at 15.6°C; 60°F)} = \frac{141.5}{\text{sp gr}} - 131.5$$

Balling Scale—This scale is used mainly in the brewing industry to estimate percent wort but also is used to indicate percent by weight of either dissolved solids or sugar liquors. Hydrometers are graduated in percent weight at 60°F or 17.5°C.

Barkometer Scale—This scale is used essentially in the tanning and tanning-extract industry. Water equals zero. Each scale degree equals a change of 0.001 in specific gravity. The following formula applies:

$$\text{Sp gr} = 1.000 \pm 0.001 \times (\text{degrees Barkometer})$$

Baumé Scale—This scale is used widely in connection with the measurement of acids and light and heavy liquids, such as syrups. The scale originally was proposed by Antoine Baumé, a French chemist, in 1768. The scale has been widely accepted because of the simplicity of the numbers which represent liquid specific gravity. Two scales are in use:

$$\text{For light liquids, °Bé} = \frac{140}{\text{sp gr}} - 130$$

$$\text{For heavy liquids, °Bé} = 145 - \frac{145}{\text{sp gr}}$$

The standard temperature for these formulas is 15.6°C (60°F).

To calibrate his instrument for heavy liquids, Baumé prepared a solution of 15 parts by weight of sodium chloride in water. On his hydrometer, Baumé marked zero at the point to which the float submerged in pure water; and he marked the scale 15 at the point to which the float submerged in the salt solution. He then divided the distance between the two marks into 15 equal spaces (or degrees as he termed them). In connection with liquids lighter than water, Baumé prepared a 10% sodium chloride solution. In this case, he marked the scale zero at the point to which the float submerged in the salt solution; and he marked the scale 10 at the point to which the float submerged in pure water. Thus, he created a scale which provided increasing numbers with decreases in density.

Users of the Baumé method found that the scale generally read 66 when the float was submerged in oil of vitriol. Thus, early manufacturers of hydrometers calibrated the instruments by this method. There were variations in the Baumé scale, however, because of lack of standardization in hydrometer calibration. Consequently, in 1904, the National Bureau of Standards made a careful survey and finally adopted the scales previously given for light and for heavy liquids.

Brix Scale—This scale is used almost exclusively by the sugar industry. Degrees on the scale represent percent pure sucrose by weight at 17.5°C (63.5°F).

Quevenne Scale—This scale is used for milk testing and essentially represents an abbreviation of specific gravity. For example, 20° Quevenne indicates a specific gravity of 1.020; 40° Quevenne, a specific gravity of 1.040, and so on. One lactometer unit approximates 0.29° Quevenne.

Richter, Sikes, and Tralles Scales—These are alcoholometer scales which indicate directly in percent ethyl alcohol by weight in water.

Twaddle Scale—This scale is the result of attempting to simplify the measurement of industrial liquids heavier than water. The range of specific gravity from 1.000 to 2.000 is divided into 200 equal parts. Thus, 1° Twaddle equals 0.005 sp gr.

An abridged compilation of specific gravity conversions is given in Table 1. The specific gravities of numerous materials are given throughout this volume.

Determination of Specific Gravity. The principal means for measuring specific gravity (and density) of liquids and gases are listed in Table 2.

Hand Hydrometer. This instrument consists essentially of a long, slender glass float weighted at the lower end and provided with a scale so graduated that the depth to which the instrument sinks in the liquid indicates the specific gravity by direct reading of the scale. See Fig. 1. The numbering of the scale increases from the top downward. The instrument sometimes is proportioned so that the numbering begins with unity at the top, being applicable only to liquids heavier than water; in others, it increases from the top to unity at the lower end and is for use with liquids lighter than water; in still other designs, unity is marked at the middle of the scale and thus the instrument may be used for both light and heavy liquids. To be sensitive, the stem carrying the scale must be slender. It may be observed that the scale intervals corresponding to equal increments of density cannot be equal where the stem is of uniform diameter. These intervals, in fact, are inversely proportional to the square of the density, being much smaller at the lower than at the upper end of the scale. To avoid this, some hydrometers are graduated with an arbitrary scale having uniform spacings, as in the case of the Baumé scale, the readings of which may be converted into density by reference to tables. Nicholson devised a hydrometer for measuring the densities of small solids, the specimen being placed on the hydrometer, first above and then below the surface of the water in which the instrument floats, and its volume deduced from the resulting alteration in buoyant force. See Fig. 2. Hand hydrometers are used extensively where automatic, continuous, and remote readings of specific gravity or density

TABLE 1. SPECIFIC GRAVITY SCALE EQUIVALENTS

Specific Gravity 60°/60°F	°Baume	°API	Specific Gravity 60°/60°F	°Baume	°API
0.600	103.33	104.33	0.800	45.00	45.38
0.620	95.81	96.73	0.820	40.73	41.06
0.640	88.75	89.59	0.840	36.67	36.95
0.660	82.12	82.89	0.860	32.79	33.03
0.680	75.88	76.59	0.880	29.09	29.30
0.700	70.00	70.64	0.900	25.56	25.72
0.720	64.44	65.03	0.920	22.17	22.30
0.740	59.19	59.72	0.940	18.94	19.03
0.760	54.21	54.68	0.960	15.83	15.90
0.780	49.49	49.91	0.980	12.86	12.89
			1.000	10.00	10.00

Specific Gravity 60°/60°F	°Baume	°Twaddle	Specific Gravity 60°/60°F	°Baume	°Twaddle
1.020	2.84	4	1.500	48.33	100
1.040	5.58	8	1.520	49.61	104
1.060	8.21	12	1.540	50.84	108
1.080	10.74	16	1.560	52.05	112
1.100	13.18	20	1.580	53.23	116
1.120	15.54	24	1.600	54.38	120
1.140	17.81	28	1.620	55.49	124
1.160	20.00	32	1.640	56.59	128
1.180	22.12	36	1.660	57.65	132
1.200	24.17	40	1.680	58.69	136
1.220	26.14	44	1.700	59.71	140
1.240	28.06	48	1.720	60.70	144
1.260	29.92	52	1.740	61.67	148
1.280	31.72	56	1.760	62.61	152
1.300	33.46	60	1.780	63.54	156
1.320	35.15	64	1.800	64.66	160
1.340	36.79	68	1.820	65.33	164
1.360	38.38	72	1.840	66.20	168
1.380	39.93	76	1.860	67.04	172
1.400	41.43	80	1.880	67.87	176
1.420	42.89	84	1.900	68.68	180
1.440	44.31	88	1.920	64.98	184
1.460	45.68	92	1.940	70.26	188
1.480	47.03	96	1.960	71.02	192
			1.980	71.77	196
			2.000	72.50	200

NOTE: 60°F = 15.6°C

TABLE 2. SPECIFIC GRAVITY AND DENSITY INSTRUMENTATION

Instrument	Liquids	Gases	Solids
Hydrometers			
Nicholson's hydrometer			x
Hand type	x		
Photoelectric type	x		
Inductance-bridge type	x		
Specific-gravity balance		x	
Fixed-volume methods			
Balanced-flow vessel	x		
Displacement meter	x		
Chain-balanced plummet	x		
Buoyancy gas balance		x	
Differential-pressure methods			
Liquid-purge systems	x		
Air-bubbler systems	x		
Viscous-drag method		x	
Boiling-point rise system	x		
Radiation gages	x		
Pycnometer[a]	x		x

[a]See separate editorial entry under **Pycnometer.**

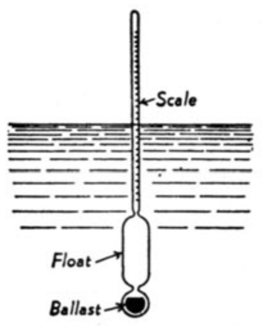


Fig. 1. Hydrometer for liquids.

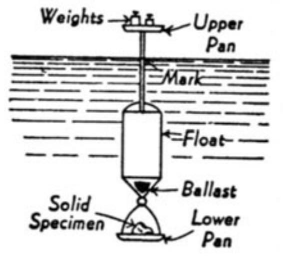


Fig. 2. Nicholson's hydrometer for small solids.

are not required. However, a simple hydrometer can be used in a standpipe, equipped with an overflow at reading level, thus allowing for visual observations of a continuously flowing liquid.

Automated Hydrometers. One form of hydrometer utilizes an opaque stem which, as the stem rises and falls, effects the amount of light which passes through a slit to a photocell. In this way, the photocell output is proportional to specific gravity and may be recorded by an electric instrument. In another industrial version, designed for remote transmission, the hydrometer is contained within a metal cylinder. A rod connects the bottom portion of the float to an armature which moves vertically between inductance coils. Changes in inductance are transmitted by cable to a central instrument panel receiving instrument.

Chain-balanced Float. In this device, a submerged plummet, which is self-centering and operates essentially without friction, moves up or down with changes in specific gravity. A section of chain is fastened to the bottom of the plummet to provide a counter-buoyancy force. The effective chain weight varies as the plummet moves up and down. For each value of density or specific gravity, the plummet assumes a specific point of equilibrium. By means of a differential transformer, readings may be transmitted to a receiving instrument. A resistance thermometer bridge may be used where compensation for density changes with temperature is required. See Fig. 3.

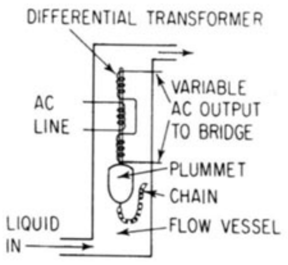


Fig. 3. Chain-balanced float or plummet-type specific gravity or density meter.

Balanced-flow Vessel. In this method of specific gravity measurement, a vessel with a fixed-volume is weighed automatically by means of a scale, spring, or use of a pneumatic force-balance system. The liquid being measured flows continuously into and out of the vessel by way of flexible connections. Accuracy of the system is very good.

Displacement-type Meter. In this instrument, the liquid being measured flows continuously through a chamber. A displacer element, usually a hollow metal sphere or cylinder containing air, is submerged fully in the liquid. The buoyant force on the displacer is measured, often by a pneumatic force-balance system, and any variations in this force reflect changes in the specific gravity or density of the liquid. The system can be compensated for changes in liquid temperature by thermostatically heating the chamber.

Differential-pressure Method. As shown in Fig. 4, two bubbler tubes, the exit of one being lower than the other, are installed in a vessel containing the liquid being measured. Air is bubbled into the liquid through these tubes. The difference in pressure required by each tube represents the weight of a constant-height column of the liquid equal to the difference in level of the ends of the two tubes. Thus, the instrument can be calibrated directly in terms of specific gravity or density. The accuracy ranges from 0.3 to 1% and provides a good approach for liquids that do not tend to crystallize in the measuring pipes. The system is used extensively in the pulp and paper industry for the measurement of white liquor, light black liquor, and bleach. There are several variations of the air-bubbler method, including the use of a reference column and a system with range suppression.

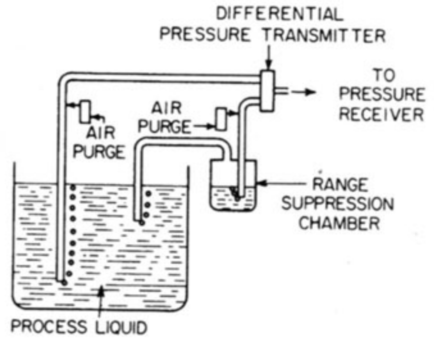


Fig. 4. Differential-pressure method for specific gravity measurement.

Boiling-point Rise System. For certain liquids, the temperature of a boiling solution of the unknown may be compared with that of boiling water at the same pressure. For a given solution, the boiling-point elevation may be calibrated in terms of specific gravity at standard temperature. Usually two resistance thermometers are used. The system finds use in the control of evaporators to determine the endpoint of evaporation. Good accuracy is achieved in the determination of one dissolved component, or of mixtures of fixed composition.

Nuclear-type Density Meters. As illustrated in Fig. 5, a radioisotope source is placed on one side of a pipeline while a detector is placed on

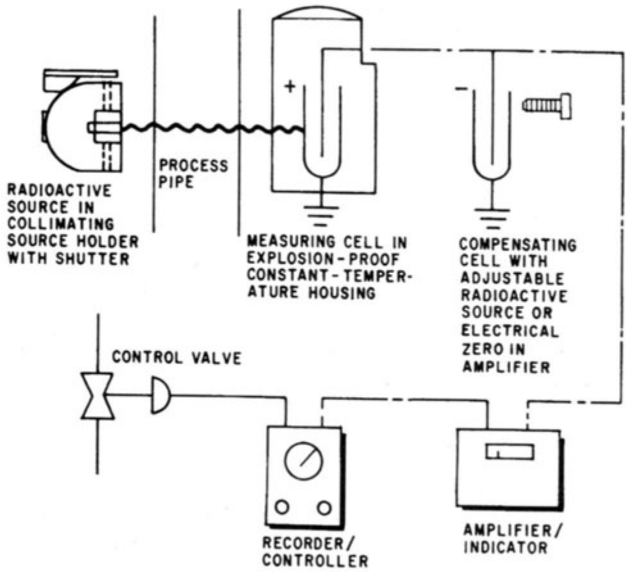

PROCESS PIPE
RADIOACTIVE SOURCE IN COLLIMATING SOURCE HOLDER WITH SHUTTER
MEASURING CELL IN EXPLOSION-PROOF CONSTANT-TEMPERATURE HOUSING
COMPENSATING CELL WITH ADJUSTABLE RADIOACTIVE SOURCE OR ELECTRICAL ZERO IN AMPLIFIER
CONTROL VALVE
RECORDER/ CONTROLLER
AMPLIFIER/ INDICATOR

Fig. 5. Nuclear radiation-type density meter.

the opposite side. Transmitted radiation is in proportion to the density of the material within the pipeline. Standard radiation detectors and associated electrical instrumentation can be used. An ac type amplifier is used for accurate determinations when the measurement spans are narrow. For wider spans, dc amplifiers may be used. A compensating cell may be used for zero suppression. This essentially eliminates zero drift rate that results from radioisotope source decay. The method is particularly attractive for the density and specific gravity determination of slurries. See also **Radioactivity.**

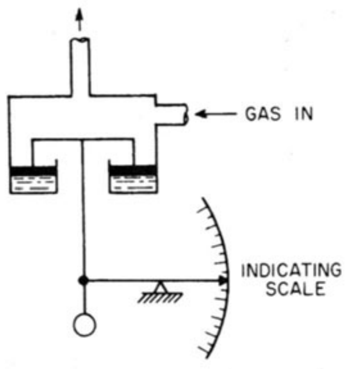


Fig. 6. Gas-specific gravity balance.

Gas Density Measurements. In the gas specific gravity balance shown in Fig. 6, a tall column of gas is weighed by the floating bottom of the vessel. This weight measurement is converted to the motion of an indicating pointer or recording pen over a graduated scale, calibrated in units of density or specific gravity. In a viscous-drag type of instrument, one set of impellers is driven in a chamber containing a standard reference gas, and the power required to achieve rotation is measured. A matching set of nonrotating impellers is located in a chamber containing the gas under test. The sets of impellers are connected by a linkage and measure the relative drag shown by the tendency of the impellers to rotate in the test gas. The balance point is a function of relative gas density. In the buoyancy gas balance, a displacer is mounted on a balance beam in the vessel containing the test gas. A reading of the air pressure required to maintain the displacer at a perfectly horizontal position (observed through a window in the chamber) is read from a manometer. Then, the air is displaced in the chamber by the gas under test. The foregoing procedure is repeated. The ratio of pressure with air to pressure with the gas provides a measure of the density of the gas relative to air. This is primarily a laboratory type measurement.

SPECIFIC HEAT. Sometimes called specific heat capacity. The quantity of heat required to raise the temperature of unit mass of a substance by one degree of temperature. The units commonly used for its expression are the unit mass of one gram, the unit quantity of heat in terms of the calorie. See also **Heat.**

Specific Heat at Constant Pressure. The amount of heat required to raise unit mass of a substance through one degree of temperature without change of pressure. Usually denoted by C_p, when the mole is the unit of mass, and c_p when the gram is the unit of mass.

Specific Heat at Constant Volume. The amount of heat required to raise unit mass of a substance through one degree of temperature without change of volume. Usually denoted by C_v, when the mole is the unit of mass, and c_v when the gram is the unit of mass.

SPECIFIC HEAT (Electronic). In the original formulation of the Drude free electron theory of metals, it was assumed that the electrons formed a classical gas whose specific heat is just $3Nk$ (N being the number of particles, k, Boltzmann's constant). No such specific heat was observed, but it was pointed out by Sommerfedd that the electrons should be treated as a Fermi-Dirac gas, for which the heat capacity at constant volume per mole of electrons is given by

$$C_v = \tfrac{1}{2}\pi^2 NkT/T_F$$

where T_F is the Fermi temperature. This formula suggests that only the fraction $(\pi^2/6)(T/T_F)$ of the electrons can actually contribute to the specific heat. Since this fraction is of the order of 10^{-3} at room temperatures, the electronic specific heat is negligible compared with the lattice specific heat except at temperatures of a few degrees absolute. In the band theory of solids, C_v is roughly proportional to the effective mass of the electrons.

SPECIFIC HEAT (Humphries Equation). An expression for the ratio of specific heats of moist air, useful in the calculation of the velocity of sound in the atmosphere:

$$C_p/C_v = \gamma = 1.40 - 0.1e/p$$

where γ is the ratio of specific heats, p is the total atmospheric pressure, e is the water vapor pressure.

SPECIFIC SURFACE. The surface, or area, of a substance or entity per unit volume; obtained by dividing the area by the volume, and expressed in reciprocal units of length.

SPECIFIC VOLUME. The volume of a substance or entity per unit mass, obtained by dividing the volume by the mass; and expressed in units of length to the third power and reciprocal units of mass. Reciprocal of the density.

See also **Density.**

SPECTRAL ANALYSIS. A method of analyzing stationary time-series into a series of harmonic terms. In effect, the series is correlated with trigonometrical functions of type cos $(\omega t + \beta)$ for a range of values of ω. A high correlation indicates the possibility of a harmonic of corresponding period in the original series; and in any case the pattern of variation of the correlation over a range of ω is typical of the constitution of the series. The process is analogous to the splitting of a ray of polychromatic light into a spectrum exhibiting its monochromatic constituents.

The correlation, or some simple function of it called the intensity, is usually graphed as ordinate either against the period $2\pi/\omega$, giving the *periodogram*, or against the frequency ω, giving the *power spectrum*.

The word periodogram is also used for other methods which attempt to detect periodicities in the series.

Similar methods can be applied to examine the relationship between two or more series. For two series this leads to two components, real and imaginary, of the spectral density. The real component is called the co-spectrum and the imaginary component the quadrature spectrum. The two are amalgamated to form a measure known as the coherence.

The spectrum can be regarded as a Fourier transform of the autocorrelation fraction. (See **Spurious Correlation.**) A more sophisticated generalization concerns the Fourier transform of third-order moments and is known as the bi-spectrum, not a term to be recommended.

Sir Maurice Kendall, International Statistical Institute, London.

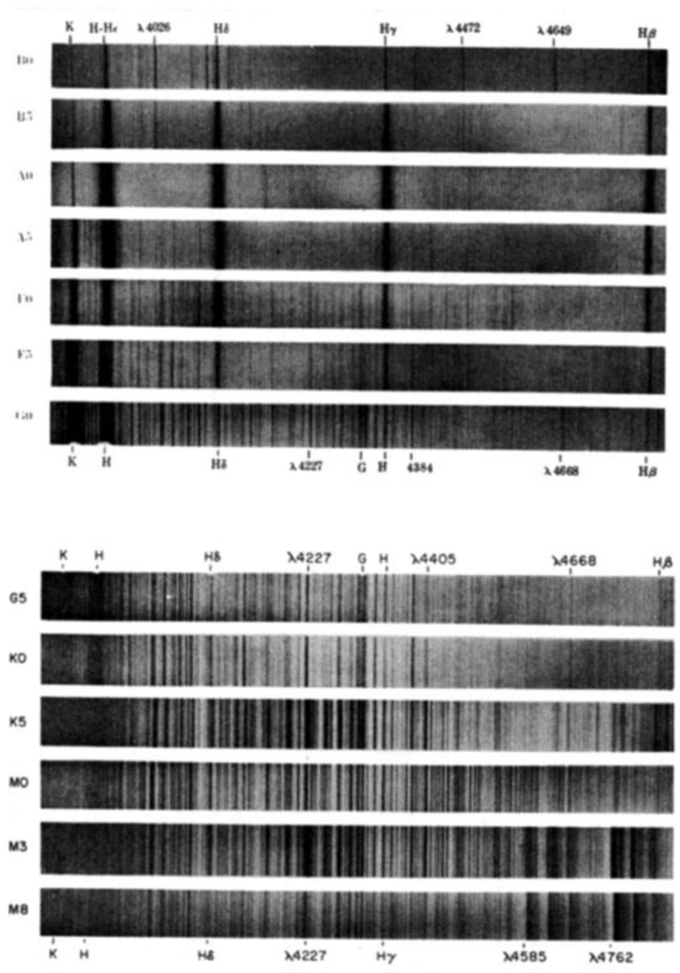
SPECTRAL CENTROID. An average wavelength, computed especially for light filters and other light-transmitting devices, by taking a weighted average, for each wavelength, of the spectral energy distribution, of the incident light, the transmittance of the device, and the luminosity data of the eye.

SPECTRAL CHARACTERISTIC. A relation, usually shown by a graph, between wavelength and some other variable. 1. In the case of a luminescent screen, the spectral characteristic is the relation between wavelength and emitted radiant power per unit wavelength interval. 2. In a photoelectric device, it is the relation between wavelength and sensitivity per unit wavelength interval.

SPECTRAL CLASS. A casual examination of the stars with the eye or small telescope reveals that stars have different colors. If stars radiate approximately as black bodies, then it will be expected that they can be graded according to their temperature, i.e., their color. The first attempt to grade stars into groups or classes was at Harvard in the famous Henry Draper catalogue. This system has been refined into the present powerful system called the Yerkes or MKK classification and is contained in the Yerkes Atlas of Stellar Spectra.

In the modern classification, the Harvard classes are still retained; thus, there is the sequence of stars W, O, B, A, F, G, K, M, R, N, and S. The Yerkes classification is essentially one of ratios of various ionized and excited levels of atoms. For example, a BO-type star is determined primarily by the fact that the ratio of the 4552 Si III line to that of the 4089 Si IV line is less than one; whereas, in a B1-type star, this ratio is greater than one. For the later types of stars (in particular, the M, R, N, and S stars), the intensity of the TiO bands is read for classification. The classes from B through M are subdivided into ten units; thus, there is B0, B1, B2, etc., to B9 and G0, G1, G2, etc., to G9.

At the time when this spectral classification was developed, it was apparent that the stars were of different luminosity; i.e., that there were giant and supergiant stars. A criteria was developed for increasing luminosity based essentially on ratios of certain lines. The ratios will be slightly different due to the sharpening of lines in a giant atmosphere. This is due to the reduced pressure in these atmospheres, and is exhibited in the accompanying figure, a series of typical spectra.



Stellar spectra.

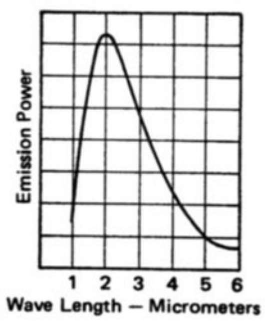
The luminosity types are Ia, the most luminous supergiants; lb, the less luminous supergiants; II, the luminous giants; III, the giants; IV, the subgiants; V, the main sequence stars or dwarfs; and VI, the subdwarfs. It is apparent, then, that there is a two-dimensional classification involving luminosity and spectral type. This is called the luminosity spectral diagram or the Hertzsprung-Russell diagram, and is exhibited in the article on the **Giant and Dwarf Stars.**

In order to use the MKK classifications, one merely obtains spectrograms at the dispersion available to his spectrograph and compares them with a series of MKK standards. To do this, he selects an unknown star, takes its spectrum at the same dispersion, and compares it with his catalogue of MKK standards. In this way, he can easily classify within one subdivision of a spectral type many, many stars in any given period of time.

A second and perhaps much more powerful method of classification has been developed by Barbier, Chalonge, and Divan. The spectrum of a star is taken at very low dispersion, traced out by means of a microphotometer, and the Balmer decrement is measured. The three coordinates are obtained—ϕ, D, and λ. Thus there is a three-dimensional classification, which differentiates very neatly between Population I and Population II. However, the resulting three-dimensional surface applies only to spectral types O through approximately F6 and in the various luminosity classes, and so has the drawback that the fainter or redder end of the spectral sequence is not considered.

SPECTRAL COLORS. 1. Colors present in the spectrum of white light. 2. Colors that are represented by points on the chromaticity diagram that lie on straight lines between the achromatic point and the spectrum locus.

SPECTRAL ENERGY DISTRIBUTION. When radiation exhibiting a continuous spectrum, as that from a hot stove or the light from an incandescent lamp, is quantitatively analyzed, it is found that quite different amounts of power are represented by the radiation within equal ranges of wavelength or of frequency having different limits. The proportion in any such range depends upon the character of the source. Thus, in the radiation from a candle, the ratio of the energy output between 6,500 and 6,600 angstroms (red) to that between 4,500 and 4,600 angstroms (blue-violet) is greater than the corresponding ratio for the radiation from an arc lamp. If we divide the spectrum into small intervals of wavelength, say 10 angstroms, and plot the power output for each range as ordinate, with the mean wavelength of the interval as abscissa, the result is a curve showing the distribution of power through the spectrum. When the radiation is due to high temperature, as in the above examples, there is always a wavelength interval having maximum power, that is, the curve has a "peak," from which the ordinates fall off in both directions. Wien pointed out that the higher the temperature of the source, the farther toward the short-wavelength end of the spectrum does this peak lie.



Spectral energy distribution for black body at 1170° (1443° absolute) with peak at 2000 angstroms (0.2 micrometer).

An instrument utilizing a prism for dispersing the radiation, together with a thermocouple or similar device for measuring its flux density in different ranges, may be used to analyze infrared thermal radiation, and is called a spectroradiometer. The spectrophotometer performs a similar service for visible light, except that the results in this case are usually tabulated in terms of the visibility rather than the actual power of the emission. See also **Wien Laws.**

SPECTRAL FUNCTION. A necessary and sufficient condition for $\rho(\tau)$, where $\tau = 0, 1, 2, \ldots$, to be an autocorrelation function of a discrete stationary stochastic process is that it is expressible in the form

$$\rho(\tau) = \frac{1}{\pi} \int_0^{\pi} \cos \tau\omega \, dF(\omega)$$

where $F(\omega)$ is a non-decreasing function with $F(0) = 0$, $F(\pi) = \pi$. For a continuous process the corresponding condition is that:

$$\rho(\tau) = \frac{1}{\pi}\int_0^\infty \cos \tau\omega \, dF\,(\omega)$$

with $F(0) = 0$, $F(\infty) = 1$. Conversely, we have

$$F(\omega) = \omega + 2\sum_{j=1}^{\infty} \frac{\rho_j}{j} \sin j\,\omega,\ 0 \le \omega \le \pi$$

for the discontinuous process, and

$$F(\omega) = \frac{2}{\pi}\int_0^\infty \rho(x)\,\frac{\sin x\omega}{x}\,dx,\ 0 \le \omega \le \infty$$

for the continuous case. The function $F(\omega)$ is variously called the spectral function, *integrated function*, power spectrum or integrated power spectrum; the first appearing to be the simplest usage. Similarly, $dF(\omega)/d\omega$ is called the *spectral density*.

Both spectral function and spectral density can be defined directly without invoking the concept of autocorrelation.

SPECTRAL SENSITIVITY. Three uses of this term are: 1. The sensitivity of a detector measured for narrow spectral bands throughout the spectrum. 2. The emitted radiant-power wavelength distribution of a luminescent screen under a given condition of excitation. 3. The sensitivity of a photoelectric device in relation to the wavelength of the incident radiant energy.

Spectral sensitivity is usually displayed on a spectral characteristic.

SPECTROCHEMICAL ANALYSIS (Visible). Chemical systems which exhibit a selective light absorptive capacity are colored. Hence, the terms *colorimetric analysis* and *colorimetry* often are used to designate the measurement of such systems when the objective is to determine the concentration of the constituent responsible for the color. The use of the term *colorimetry* in this respect is not to be confused with the use of the same term in physics where the term refers strictly to the measurement of color.

Visible spectrometry may be used to determine a constituent constituting the major part of a sample, but it also may be applied to the determination of trace quantities. Some methods are applicable to amounts of a few parts per million, with some tests sensitive to 0.01 ppm or less.

Like other areas of spectroscopy, visible spectrometry has a wide range of applications. Included are most of the elements, many anions, functional groups, and innumerable compounds.

The main practical problems in the methodology of visible spectrometry are (1) to prepare a suitable colored solution; and (2) to measure the light absorptive capacity of this solution, or to compare it with that of a colored solution of known concentration.

Visual spectrochemical analysis has largely been displaced by other, more automated instrumental methods.

SPECTRO INSTRUMENTS. *Spectro* is used as a prefix for a wide assortment of analytical instruments. *Spectro* is derived from *spectrum*, which originally referred to the component colors that make up visible light, the so-called rainbow colors of violet, indigo, blue, green, yellow, orange, and red. A very simple device made up of a glass prism to break up sunlight into color bands is referred to as a spectro*scope*. Much more sophisticated instruments are available for manual manipulation and observation, which still rely on this basic, simple principle; these are termed visual spectroscopes, and the field is called visual spectroscopy.

Over the years, the term *spectrum* has increased in application and meaning and now embraces the total electromagnetic radiation span—no longer being confined to the visible portion. Concurrently, the term *spectroscope* has broadened in meaning so that mention of spectroscope no longer signifies an instrument which operates in the visible region. This situation gave rise to the need for modifying words for use with the term *spectroscope* to signify the portion of the electromagnetic spectrum with which the instrument is concerned. Thus, there are infrared spectroscopes, x-ray spectroscopes, ultraviolet spectroscopes, microwave spectroscopes, and so on. Further, the term *spectrum* has widened to include practically any orderly array of phenomena. For example, a mass spectrometer sorts atoms or radicals by their atomic weight—over a spectrum of values. Other spectrometers analyze the decay "spectra" of radioactive isotopes.

In terms of what is measured or observed, there are (1) portions of the electromagnetic spectrum: gamma-ray, cosmic ray, x-ray, ultraviolet, infrared, far-infrared, microwave, and radiowave instruments; (2) regions pertaining to the energies of particles: beta ray (electrons), protons, neutrons, and mass associated instruments; and (3) instruments dealing with other "spectra" such as radioactive decay and Mossbauer effects.

The suffix *scope* was ample when spectro instruments were essentially manual and confined to visual observations with the unaided eye. As the functions of these instruments increased, new suffixes were required to completely describe the instruments. Thus, if the instrument provides a record of the measurement, either by means of photographic film or by pen recording on a chart, the suffix *graph* is used. Where a meter is utilized to display the information detected by a device such as a bolometer, thermocouple, or thermistor, and so on, the suffix *meter* may be used. In cases where an instrument is designed to measure the intensity of various portions of spectra (again not confined to visible light), the suffix *photometer* may be used. Thus, there are spectrographs (spectrography), spectrometers (spectrometry), and spectrophotometers (spectrophotometry).

But, with increasing complications of instrument design and flexibility of use, the foregoing terms still are not sufficient to fully describe many instruments. Although some instruments display a continuous spectrum of what is being measured, other instruments filter out, for example, certain incident radiation. In other words, certain radiation may be absorbed, giving rise to the terms *absorption spectrometer* or *absorption spectrograph*. In some cases, the energy level of the specimen under analysis must be raised, as by means of flame heating or a spark discharge. Instruments in this category are of the emission type—and hence such terms as *flame photometer* and *optical emission spectrometer*. The names of spectro instruments can be complicated further where some special function may be incorporated in the title. A few examples are:

Spectroheliograph: a spectrograph designed for use with a telescope and for making photographs of the sun, in which the radiation of a particular element in the radiation of the sun may be recorded.

Ultraviolet-visible Spectropolarimeter: an instrument for measuring the rotation of the plane of polarization in accordance with wavelength and light intensity.

Spectrophotofluorometer: an instrument which provides means for both controlling the exciting wavelength and for identifying and measuring light output of a fluorescing sample. If the term *fluorometer* alone were used, it could indicate an instrument with a filtered light input that will measure the fluorescent light from a sample, usually by wavelength.

SPECTROSCOPE. Several types of instruments for producing and viewing spectra are included under this term. Variations in form are due, not only to differences in principle, but also to the type of radiation or phenomena to be examined. Terminology employed in this field of instrumentation is described under **Spectro Instruments.**

The earlier spectroscopes, developed by Fraunhofer, Ångström, and others in the early part of the last century, adapted Newton's discovery of the dispersion of light by a prism. The essential features shown in Fig. 1 are a slit, S, a collimator, C, for rendering the light from the slit parallel before entering the prism, one or more dispersing prisms, P, and a telescope, T (or a camera), for forming images of the slit in the various wavelengths and thus providing a method for viewing or photographing the spectrum. The light passes through these in the order named, being deviated by the prism through various angles according to the wavelength. When a spectroscope is provided with a graduated circle for measuring deviations, it is called a "spectrometer." The"direct-vision" or non-deviation spectroscope, employing an Amici prism, is a compact instrument for qualitative purposes. The photographic spectroscope,

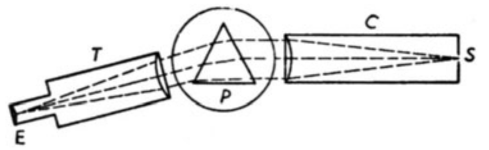


Fig. 1. Simple prism spectroscope: Collimator, C; slit, S; prism, P; telescope, T; and eyepiece for viewing spectrum, E.

known as a "spectrograph," is now almost universally used in spectral research.

Many modern spectroscopes employ the diffraction grating instead of the prism. In the concave grating spectroscope, developed by Rowland, the collimating lens and telescope or camera objective are unnecessary because of the focusing effect of the grating itself. See Fig. 2.

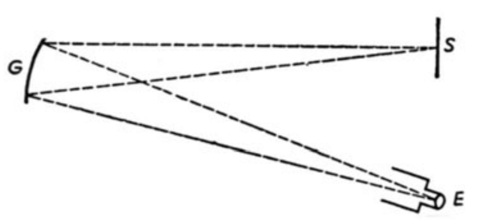


Fig. 2. Concave grating spectroscope: Slit, S; grating, G, eyepiece or plateholder, E.

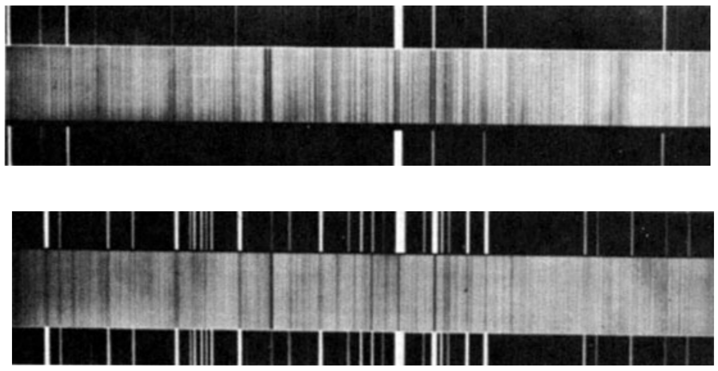
The growth of the importance of infrared spectrography and spectrophotometry in determining the structure of compounds and the composition of substances has led to the development of many infrared spectroscopes and other instruments. Most infrared spectroscopes and spectrophotometers employ front-surface mirrors instead of lenses. This eliminates the necessity for energy to pass through glass, quartz, or similar material. Furthermore, it would be difficult or impossible to make lenses which would bring rays of the widely varying wavelengths in the infrared region to focus at one point. Occasionally a rock salt lens will be found in an infrared spectrophotometer, but such lenses are usually not essential optical components. Parabolic mirrors bring energy of all wavelengths to focus at one point. Reflection from most metallic surfaces is generally very efficient in the infrared region. Both gratings and prisms can be used for dispersing the energy, but prisms seem to be more common, perhaps because energy in the infrared region is at a premium and none can be wasted in higher order spectra. The materials which are found most suitable in the infrared region are quartz, calcium fluoride, sodium chloride, and sodium bromide, all in the form of single crystals.

The energy source for an infrared spectroscope or spectrophotometer is a Nernst glower or a Globar. The receiving element for the infrared radiation must be a bolometer, Golay cell or thermistor since photocells are not sensitive in this region of the spectrum. Indeed, the thermoelectric element chosen must be extremely sensitive, since the average energy in the dispersed beam is very small.

Most available infrared instruments use the Littrow mount for the prism, the beam being reflected from a plane mirror behind the prism and thus returning it through the prism a second time. This doubles the dispersion produced. Actually, a double-pass system is also used so that the beam goes through the prism four times. Other design modifications include those with single beam and double monochromator, double beam and double monochromator, and related combinations. See also **Infrared Radiation.**

Check alphabetical index for a wide range of spectroscopic instruments that are based upon numerous materials-energy relations. Also see **Analysis (Chemical)**.

SPECTROSCOPIC BINARIES. Discovery of the brighter component of Mizar in 1889 introduced a class of binary stars which are not double stars in the ordinary sense of the term. The components are too close together for them to be observed separately even with optical telescopes of high resolving power. In such binaries the period is usually short and the orbital velocities high. Unless the orbit plane happens to be perpendicular to the line of sight, the orbital velocities will have components in the line of sight and the observed radial velocity of the system will vary periodically. Since radial velocity is measured with the spectroscope, employing the Doppler-Fizeau principle, the binaries so observed are known as spectroscopic binaries. In some spectroscopic binaries the spectra of both stars are visible and the lines are alternately double and single. Such stars are known as double-line binaries. In others the spectrum of only one component is seen and the lines in this spectrum move periodically from violet toward the red and back again.


Spectrogram of *Mizar*. The lines of the two components are separated in the upper photo and superposed in the lower. This was a classical observation made several decades ago at Yerkes observatory.

The determination of the orbit of a spectroscopic binary is made from a long series of observations of the radial velocity of one or more components of the system. The observations are first plotted against time and from the resulting curve the period may be obtained. With this period determined observations are then reduced to a single epoch and the best possible curve drawn through the points, obtaining what is known as the velocity curve of the system. If the orbit is circular the velocity curve will be a sine curve; if elliptical, the shape of the curve will depend upon the eccentricity of the ellipse and the orientation of the major axis with reference to the line of sight. From the shape of the velocity curve the orbit of the system in space may be determined. In the solution of the spectroscopic orbit it is impossible to determine individually the semimajor axis, a, and the inclination of the orbit plane, i. However, the product of the semimajor axis by the sine of the inclination (i.e., $a \sin i$) may be determined directly in linear units (i.e., in either miles or kilometers). If either a or i can be obtained from other types of observations, as in the case of eclipsing binaries, a complete solution for the orbit can be made.

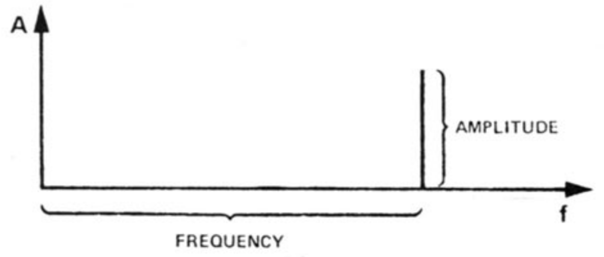
See also **Binary Stars; Eclipsing Binary;** and **Visual Binaries.**

SPECTROSCOPIC PARALLAX. The term *spectroscopic parallax* of a star is applied to a determination of the distance of a star in which the stellar parallax is determined from observations of spectral peculiarities of the star together with determinations of the apparent brightness of the object.

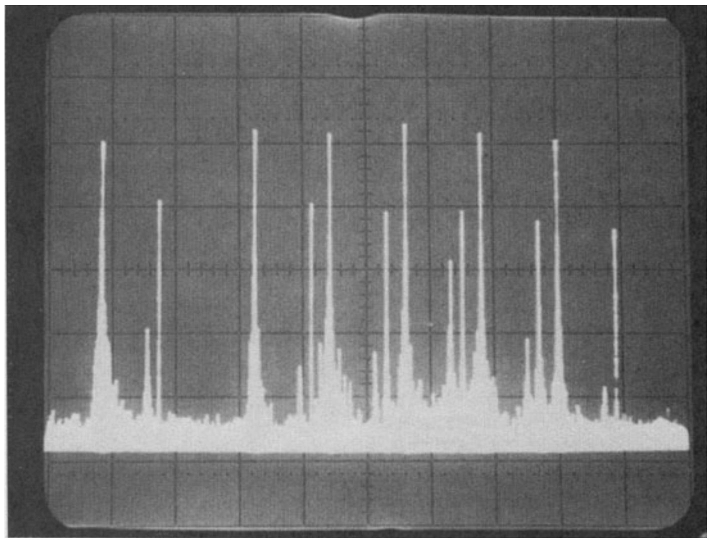
The apparent brightness of a star depends upon two fundamental factors: the intrinsic brightness of the star and its distance from the observer. Expressed on the stellar magnitude scale, we find the apparent magnitude, m, the absolute magnitude, M, and the stellar parallax, π, to be connected by the analytical expressions: $M = m + 5 + 5 \log_{10} \pi$. The apparent magnitude of a star may be determined by a variety of methods of stellar photometry, and if a method is available for the determination of the absolute magnitude the value of the stellar parallax may be determined.

The relative intensities of certain spectral lines are different in giant and dwarf stars of the same spectral type. The relative intensities of selected pairs of lines may be compared in stars of the same spectral type and known absolute magnitudes and a "calibration curve" obtained. The relative intensities of the same pairs may then be found in stars of unknown absolute magnitudes and the calibration curves used to determine the absolute magnitude. Thus, the parallax may be determined from a study of spectra. The accuracy of the determinations of spectroscopic parallax compares very favorably with parallaxes obtained from the relative trigonometric methods.

SPECTRUM ANALYSIS. An electronic measurement technique for quantifying the level of electrical signals over a specific frequency range. The spectrum analyzer, an electronic instrument for making spectrum analysis measurements, conveys its signal information usually by displaying a rectangular plot of signal amplitudes (level) versus signal frequency (spectrum) on a cathode ray tube (CRT). Figure 1 shows a diagramatical plot of a frequency spectrum and the comparable CRT photo of a spectrum analyzer output.



(a)



(b)

Fig. 1. (a) Information content of spectrum analyzer display; (b) Six-channel spectrum analyzer CRT display of six television broadcasts.

Signal analysis. An electrical *signal* transfers energy from one point to another. A signal may be strictly for power transfer, such as in high voltage power distribution, or it may be for information transfer, such as a telephone voice channel. It is necessary to quantify electronic signals for the effective design and maintenance of all types of equipment and machinery.

An electrical signal is composed of one or more frequency components channeled into a single transmission path, such as a cable, waveguide, or antenna. (A dc signal is no exception, it has one signal component at zero frequency.) An instrument that is capable of identifying and quantifying each frequency component is a *signal analyzer*. A spectrum analyzer is a type of signal analyzer.

The number of different types of signal analyzers are differentiated mainly by the electronic measurement techniques used. Each has a significant contribution to make to spectrum analysis measurements. These techniques are: (1) Fourier transformation of an input time domain signal by mathematical computation (Fourier analyzer); (2) filtering an input signal through a number of piecewise tuned bandpass filters (real-time spectrum analyzer); (3) scanning an input signal with a tunable filter (spectrum viewer); and (4) heterodyne receiver (spectrum analyzer, tuned voltmeter, and wave analyzer).

Fourier analyzer. Any nonsinusoidal periodic signal has sinusoidal components predicted by the mathematical Fourier series equations. The Fourier analyzer processes a given input voltage signal with a digital computer, transforming the signal into the frequency domain with both phase and amplitude information. See Fig. 2.

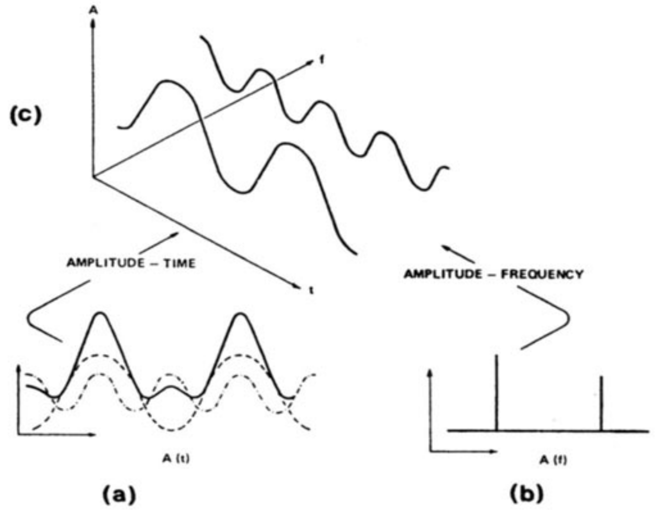


Fig. 2. (a) A time-varying input signal; (b) the Fourier analyzer transforms the signal into the frequency domain (with phase information also available, but not shown; (c) composite 3-dimensional plot showing relationship of time and frequency graphs.

The Fourier analyzer is also capable of translating nonperiodic voltage signals by calculating the Fourier transform in a form convenient for the digital computer sampling processes:

$$S_x''(t) = \Delta t \sum_{n=-\infty}^{n=+\infty} x(n\Delta t)e^{-2\pi fn\,\Delta t}$$

where $x(n\Delta t)$ are the analyzer-measured values of the input voltage signal.

Any time-varying voltage can be transformed. When a signal is not periodic, the total elapsed input time, T, is taken as the period. The analyzer develops a frequency spectrum (with phase information) of the signal as if it were being repeated at a $1/T$ rate. See Fig. 3. The CRT display is capable of showing either the real or imaginary frequency components on a rectangular grid or the magnitude/phase diagram on a polar plot.

The fundamental parameters that govern the Fourier analyzer's performance capability can be summarized by these equations:

$$T = n \times \Delta t \text{ (for the time domain)}$$

where $T =$ total time of the sample, in seconds
$n =$ number of samples taken, usually a power of 2, reflecting the amount of computer storage available
$\Delta t =$ time between samples, in seconds

$$F_{\max} = n/2 \times \Delta f \text{ (for the frequency domain)}$$

where $F_{\max} =$ maximum frequency of display, bandwidth, in Hz
$n/2 =$ the number of frequency points, $\frac{1}{2}$ of n points are dedicated to20either the real or imaginary frequency planes
$\Delta f =$ the number of Hz between displayed frequency points, frequency resolution

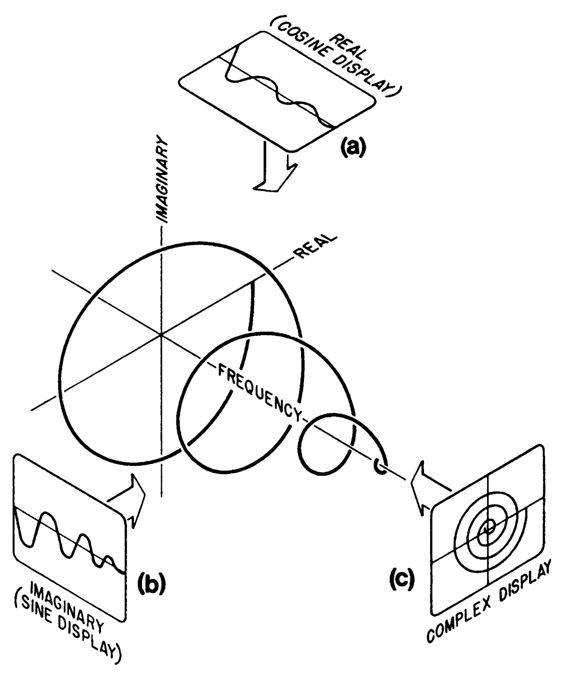


Fig. 3. Fourier analyzer processing a nonperiodic input voltage signal with the digital Fourier transform. (a) The real frequency component envelope (individual frequency spikes are not shown); (b) imaginary frequency component envelope; and (c) complex amplitude and phase display.

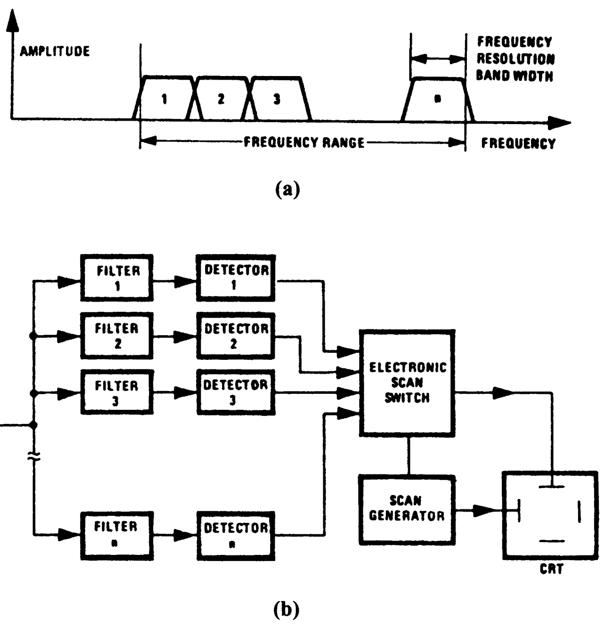


Fig. 4. (a) The real-time spectrum analyzer consists of a number of passband filters aligned such that their passbands continuously cover the spectrum of interest; (b) the electronic scan switch samples the filters fast enough to display instantaneous response of the input signals.

Solving these to eliminate n:

$$\Delta t = \tfrac{1}{2}F_{\max} = \text{samplespacing in time}$$

$$\Delta f = 1/T = \text{frequency resolution}$$

$F_{\max}$ is typically 100 kHz or below. For a given n, either $F_{\max}$ or Δf can be optimized only at the expense of the other. However, the selection of sample times and bandwidths can be tailored to the individual application requirement.

Because of the Fourier analyzer's measurement and computational strength, it can be used to simulate the transfer function for processes that have measurable input/output voltage signals. The analyzer can then be used to simulate the processed frequency spectrum response to stimuli too difficult, dangerous, or costly to input to the real process.

Real-time spectrum analyzer. This instrument, shown in Fig. 4, displays the frequency spectral components on a CRT as they occur. The display response is directly proportional to the input signal at each instant of time. (The Fourier analyzer presents a computed display that simulates this real-time response, and thus it is sometimes referred to as a real-time analyzer.)

A real-time spectrum analyzer is able to respond to input stimulus essentially instantaneously because it passes the input directly to the display through a set of filters and detectors which quantify the spectral energy. See Fig. 5.

The real-time spectrum analyzer shows accurate stimulus-response characteristics up to 10 kHz. With as many as 500 high-resolution bandpass filters, the analyzer can display signals close together in frequency even when they differ in amplitude. Filter rolloff is typically a 15:1 voltage decrease per octave.

Spectrum viewer. The *spectrum viewer* uses the same input filter-detector as the real-time spectrum analyzer except the viewer uses only a single filter which is swept-tuned through the frequency band of interest. See Fig. 6.

The advantage is cost effective octave or greater band spectral coverage from 2 GHz to over 18 GHz made possible by a tunable yttrium iron garnet (YIG) filter. As with all swept-tuned analyzers, the spectrum viewer is not a real-time analyzer. Since it views only one small segment of frequency spectrum in any one instant, the filter must charge and discharge as it is tuned past various spectral signals. (The parallel filters of the real-time analyzer always remain responsive to input spectra, even though they are sampled for display by sweeping process.)

The spectrum viewer is the least expensive technique for displaying

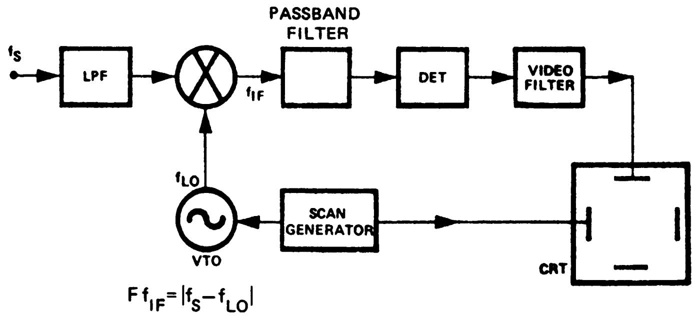


Fig. 5. Application of superheterodyne principle in real-time spectrum analyzer.

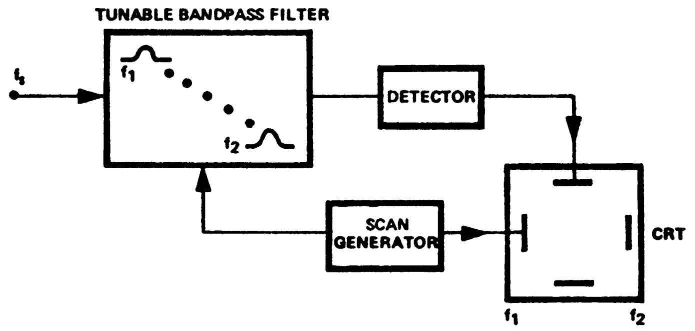


Fig. 6. Spectrum viewer operation. Synchronizing the tunable filter center frequency with the cathode ray tube horizontal deflection results in a spectral display for each sweep.

frequencies above 1 GHz; however, it generally does not have absolute amplitude calibration, high frequency resolution, or high sensitivity.

Superheterodyne receiver spectrum analyzer. The final category of signal analyzers utilizes the same technique used in almost every commercial radio receiver, the *superheterodyne* principle. The technique uses a bandpass filter as in the other analyzers, but in the heterodyne process the filter is not centered at the input frequency. Rather the filter is at a fixed frequency called the intermediate frequency (IF). Since most of the instrument's signal processes are done at the single IF, more sophisticated filtering and scaling is possible. This results in accuracy and high performance.

The local oscillator (LO) frequency is varied from a control on the instrument's front panel until the following condition is met:

$$f_{\mathrm{LO}} - f_{\mathrm{S}} = f_{\mathrm{IF}}$$

where f_{LO} = local oscillator frequency
f_{S} = input signal frequency
f_{IF} = intermediate frequency

The resulting IF signal is processed with the familiar bandpass filter and displayed. This technique is operable from dc to over 40 GHz.

Tuning the superheterodyne signal analyzer is done either manually, reading only one point in the spectrum at a time, or automatically, sweeping and reading over a frequency range called a *frequency* span. Traditionally only the swept-tuned analyzer is called a *spectrum analyzer*. See Fig. 7. The manually tuned analyzer, or *wave analyzer*, is capable of spectrum analyzer measurements. Many of the operational characteristics are common to both the wave and spectrum analyzers.

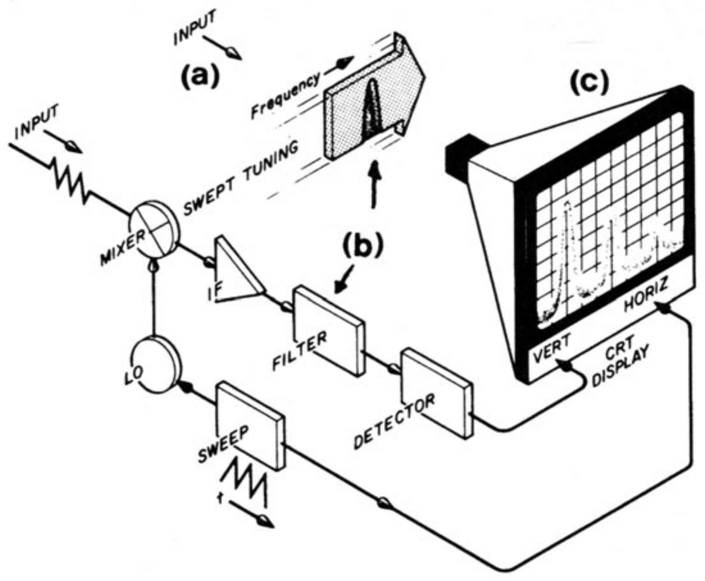


Fig. 7. Swept-tuned spectrum analyzer. Input spectrum (a) is mixed with sweeping local oscillator to produce three successive responses at f_{IF}. These are ushered through the IF filter (b) to the display (c). Note the displayed signal takes on the filter shape.

The output for the wave analyzer is a meter or digital display, calibrated in units convenient for the industry to be served. Most spectrum analyzer detection methods are proportional to power; thus, units of power and voltage referred to a transmission characteristic impedance (Z_0) are the most commonly used.

Here are their definitions:

V = volts (rms) into Z_0
mV = millivolts into Z_0
μV = microvolts into Z_0
dBm = decibel above a milliwatt
$= 10 \log \dfrac{P}{10^{-3}}$
dBmV = decibel above a milliwatt referred to Z_0
$= 20 \log \dfrac{v}{10^{-3}}$
dBμV = decibel above a milliwatt referred to Z_0
$= 20 \log \dfrac{v}{10^{-6}}$

where P = power in watts
v = voltage in rms volts
Z_0 = transmission line characteristic impedance, such as 50 Ω, 75 Ω, 600 Ω

Absolute amplitude calibration means each point on the meter or CRT display represents a specific voltage or power level. *Dynamic range* is the amplitude ratio of the largest to the smallest signal that can be displayed simultaneously with no analyzer distortion products present. (Distortion products are harmonic and intermodulation signals generated inside the spectrum analyzer which are displayed as legitimate signals.)

The lowest level signal that can be detected by the analyzer is specified by *sensitivity*. A spectrum or wave analyzer's sensitivity is limited by its inherent average noise. An unknown signal can be detected when the signal power is equal to the inherent average noise power. On the CRT this signal will appear 3 dB above the noise level. This relationship is given by

$$\frac{P+N}{N} = 2$$

where P = signal power, in watts
N = inherent analyzer noise power, in watts.

The ability to distinguish between two input signals which are close together in frequency is *resolution*. In spectrum analyzers this capability is primarily dependent upon IF resolution bandwidth, the 3 dB down points (half power) on the filter response curve (Fig. 8). The narrower the resolution bandwidth, the better the resolution. However, since the IF filter must respond fully to the input signal ($f_{\mathrm{IF}} = f_{\mathrm{s}} - f_{\mathrm{LO}}$) on each sweep, the sweep speed (Hz/second) must be slow enough to accommodate the filter. This relationship is shown by

$$T \propto \frac{S}{(BW)^2}$$

or

$$BW \cong \frac{S}{T} \times K$$

for Gaussian shaped filter,

where BW = 3 dB bandwidth, or resolution bandwidth of filter, in Hz
S = frequency span, in Hz
T = time to sweep S, in seconds
K = 0.94 (Hz sec)$^{1/2}$

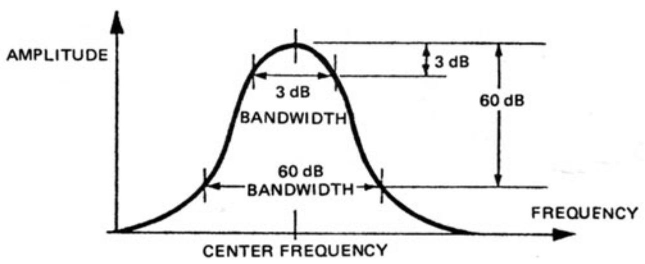


Fig. 8. Typical Gaussian filter. Signals of equal amplitude can just be resolved when they are separated by the 3 dB bandwidth. Unequal signals can be resolved if they are separated by greater than half the bandwidth at the amplitude difference between them. Shape factor is defined as the ratio of the 60 dB bandwidth to the 3 dB bandwidth.

Applications. Spectral response from electrical signals provides insight in a number of scientific fields. With the wide selection of transducers available, almost every industry and technology has areas

APPLICATIONS OF SPECTRUM ANALYSIS

Type of Instrument	Frequency Range	Application	Examples
Fourier analyzer	dc-100 kHz	Sound and vibration	Acoustic imprints for voice, noise pollution, and sonar; structural analysis for rotating machinery; stimulus-response testing
		Time-to-frequency transformation	Mechanical structure models; real-time transfer function determination; improving signal-to-noise ratios (noise rejection)
Real-time analyzer	dc-10 kHz	Sound and vibration	Real-time acoustic and structural analysis
		Electroic circuit analysis (100 kHz)	Filter design
Swept-tuned analyzer (Spectrum analyzer)	10 Hz-40 GHz	Electronic circuit analysis	General circuit design measurements; modulation, noise, distortion, power, frequency, and stability

where spectrum analysis can make a useful contribution. Some of these are summarized in the accompanying table.

Jeffrey L. Thomas, Hewlett-Packard, Santa Rosa, California.

SPEECH CLIPPING. The clipping of peak speech signals (peak clipping) or the reduction of weaker speech signals to zero (center clipping) in intelligibility tests.

SPEED. The magnitude of the vector velocity. Speed is a scalar quantity and is expressed in units of length divided by time. See also **Velocity.**

SPERRYLITE. A mineral diarsenide of platinum, $PtAs_2$. Crystallizes in the isometric system. Hardness, 6–7; specific gravity, 10.58; color, white; opaque. Named after Francis L. Sperry, Sudbury, Ontario.

SPHALERITE BLENDE. Also known as zinc blende, this mineral is zinc sulfide, (Zn, Fe)S, practically always containing some iron, crystallizing in the isometric system frequently as tetrahedrons, sometimes as cubes or dodecahedrons, but usually massive with easy cleavage, which is dodecahedral. It is a brittle mineral with a conchoidal fracture; hardness, 2.5–4; specific gravity, 3.9–4.1; luster, adamantine to resinous, commonly the latter. It is usually some shade of yellow brown or brownish-black, less often red, green, whitish, or colorless; streak, yellowish or brownish, sometimes white; transparent to translucent. Certain varieties are phosphorescent or fluorescent. Sphalerite is the commonest of the zinc-bearing minerals, and is found associated with galena, chalcopyrite, tetrahedrite, barite, and fluorite, as a result of contact metamorphism, and as replacements and vein deposits.

There are very many European localities, including Saxony; Bohemia; Switzerland; Cornwall, in England; Spain; Sweden; Japan; and elsewhere. In the United States, sphalerite is found in Arkansas, Iowa, Wisconsin, Illinois, Colorado, New Jersey, Pennsylvania, Ohio, and especially in the area which includes parts of Kansas, Missouri, and Oklahoma. The word sphalerite is derived from the Greek, meaning treacherous, and its older name, blende, meaning blind or deceiving, refers to the fact that it was often mistaken for lead ore.

SPHENE. This mineral occurs as a yellow, green, gray, or brown calcium titanosilicate, corresponding to formula $CaTiSiO_5$, crystallizing in the monoclinic system. Fracture conchoidal to uneven; brittle; habit usually wedge-shaped and flattened crystals, also massive and lamellar; luster, resinous to adamantine; transparent to opaque; hardness, 5–5.5; specific gravity, 3.45–3.55.

Sphene is an accessory mineral of widespread occurrence in igneous rocks, and calcium-rich schists and gneisses of metamorphic origin, and very common in nepheline-syenites. In the United States sphene is found in Arkansas, California, New Jersey, New York; in Ontario and Quebec in Canada; and from Greenland, Brazil, Norway, France, Austria, Finland, Russia, Madagascar and New Zealand, as well as many other world localities.

SPHENISCIFORMES (*Aves, Spheniscidae*). A group of flightless seabirds with very distinguishing characteristics. Their relationship to other orders of birds is uncertain, and therefore some ornithologists separate the penguins as a superorder or even a subclass which is distinct from all other living birds. Their closest relatives appear to be the Tube-Nosed Swimmers (*Procellariiformes*).

The morphological characteristics which distinguish these birds from other birds are a result of their adaptation to life in the water. The length is from 40 to 115 centimeters (16 to 45 inches), and the weight ranges from 1 to 30 kilograms (2 to 66 pounds). The long, spindle-shaped body has legs which are inserted far back, so they are most effective as oars and steering organs. The tail, as a steering rudder, is streamlined and triangular; the wings are transformed into flippers, but contain all the bones of a wing suitable for flying. The bones, however, are shortened, flattened, and tightly connected by ligaments, thus forming a rigid surface. The breast muscles (wing muscles) are large, taking up the whole front from the neck on down to the lower abdomen. The trachea is, as in *Procellariiformes*, divided lengthwise. The body is uniformly covered with feathers except for the brood patch; they have thick subcutaneous fat-pads. There are 6 genera with 18 species, which are confined to the Southern Hemisphere. See also **Penguin.**

SPHERE. A solid bounded by a spherical surface. It may be generated by revolving a semicircle about its diameter as an axis. A *radius* of a sphere is a straight line from the center to the surface of the solid. A *diameter* is twice as long as a radius for it passes through the center of the sphere and ends on the surface. All radii of a sphere are equal and all diameters are equal. A great circle of a sphere is determined by a plane passing through the center; if the plane does not pass through the center, the section is a small circle. An *axis* of a sphere is any diameter and its ends are called poles, often a north or a south pole.

Two arcs of great circles intersecting on a sphere determine a spherical angle. A spherical polygon is a part of the surface of a sphere bounded by three or more arcs of great circles. When it has three sides it is a spherical triangle, which may be right, obtuse, acute, equilateral, isosceles, or scalene (see **Angle**).

The volume of a sphere, $V = 4\pi r^3/3$, where r is its radius and its area, $A = 4\pi r^2$, which also equals the lateral area of a cylinder circumscribed about it. If the sphere is hollow, that is a closed spherical surface, so that its radius to the inside wall is r_1 and to the outside wall r_2, then $V = 4\pi(r_2^3 - r_1^3)/3$.

For area and volume of torus, see **Anchor Ring (or Torus).**

For other properties of a sphere, see **Spherical Surface.**

SPHERE OF INFLUENCE (Planet). The surface in space about a planet where the ratio of the force with which the sun perturbs the motion of a particle about the planet, to the force of attraction of the planet equals the ratio of the force with which the planet perturbs the motion of a particle about the sun, to the force of attraction of the sun on the particle. The volume inside this surface defines the region where the attracting body exerts the primary influence on a particle. See also **Planets and the Solar System.**

SPHERICAL ABERRATION. If the surfaces of a lens or the reflecting surface of a mirror are spherical, the rays refracted through or reflected from the outer portions will be brought to a focus in a different plane than those from the center, thus producing a blurring of the resultant image known as spherical aberration. This effect is more pronounced in short-focus lenses or mirrors than in long-focus instruments, for the curvature of the surfaces of the short-focus instruments is greater.

The decrease of spherical aberration with increase in focal length was discovered very early in the history of optical instruments, and during the seventeenth century we find telescope builders increasing the focal lengths of their instruments to tremendous proportions. Telescopes with focal lengths between 100 and 200 feet were not uncommon during this period, and the problem of supporting the long thin tubes so that they could be used for astronomical purposes and remain straight was one calling for great ingenuity. Descartes in 1637 published the theory of spherical aberration and showed that theoretically it could be corrected by grinding the surfaces of lenses and mirrors in curves other than spheres. The difficulties of grinding the required lens curves were apparently greater than those of operating the long-focus telescopes. About the middle of the eighteenth century John Dolland published the fact that spherical aberration could be corrected by the same method employed for the treatment of chromatic aberration, i.e., by using two lenses, one convergent and the other divergent. All modern telescopic lenses of good quality now employ the double object glass with the figures of the lenses and the separation between the components depending upon the ideas of the makers. For wide-angle, short-focus lenses, such as are used in modern hand cameras and in astrographic cameras, the simple pair of lenses does not provide sufficient correction for all of the aberrations, and three or more lenses are used in combination. Perhaps the most common is the so-called doublet, in which two pairs of lenses are used with a considerable separation between the pairs.

Spherical aberration may be corrected in the case of a mirror by grinding the concave surface in the form of a paraboloid of revolution instead of in the form of a sphere. This provides almost complete correction for rays entering the mirror parallel to the axis of revolution of the paraboloid, i.e., along the principal axis of the mirror; but for rays entering at a moderately large angle, spherical and various other aberrations make their appearance in the resultant image. Hence, while the reflecting type of telescope can be more easily corrected for the aberrations along the axis, nevertheless it cannot be used to obtain photographs of large areas with good definition throughout.

SPHERICAL HARMONICS. In analogy to harmonic functions in the plane, the solutions of the Laplace equation in spherical coordinates. Spherical surface harmonics are special sets taken over the surface of a sphere; therefore, the harmonic components are restricted to an integral number of waves over the sphere. Spherical harmonics have been applied in the study of the large-scale oscillations of the atmosphere.

SPHERICAL POLAR COORDINATES. A curvilinear coordinate system. Its parameters are: (1) the radius vector r from an origin or pole to the point; (2) the colatitude θ, an angle made by r and a fixed axis, the polar axis; (3) the longitude ϕ made by the plane of θ with a fixed plane through the polar axis, called the meridian plane. The coordinate surfaces are: (1) concentric spherical surfaces about the origin, $r =$ const.; (2) right circular conical surfaces with apex at the origin and axis along the Z-axis, $\theta =$ const.; (3) planes from the Z-axis, $\phi =$ const. The range of the variables is $0 \leq r \leq \infty$; $0 \leq \theta \leq \pi$; $0 \leq \phi \leq \pi$. In terms of a right-handed rectangular system with the same origin:

$$\begin{aligned} x &= r \sin \theta \cos \phi; & r^2 &= x^2 + y^2 + z^2 \\ y &= r \sin \theta \cos \phi; & \theta &= \cot^{-1} z/\sqrt{x^2 + y^2} \\ z &= r \cos \theta; & \phi &= \tan^{-1} y/k \end{aligned}$$

Synonymous terms are spherical coordinates or polar coordinates in space. If $\theta = \pi/2$, the point lies in the XY-plane and if the longitude ϕ is then called θ, as is customary, the system becomes that of polar coordinates in a plane.

See also **Conical Surface; Coordinate System** and **Polar Coordinates.**

SPHERICAL SURFACE. A surface all points of which are at a fixed distance, the *radius*, from a fixed point, the *center*. The term sphere is frequently used for this surface but it more properly means a solid bounded by a spherical surface.

In rectangular coordinates its general equation is

$$x^2 + y^2 + z^2 + Gx + Hy + Kz + L = 0$$

but if the center is taken at the origin of the coordinate system, the equation becomes $x^2 + y^2 + z^2 = r^2$, where r is the radius of the spherical surface.

The surface is measured in terms of the following parts: (1) *Zone*, a portion of the surface included between parallel planes. The bases of the zones are circumferences made by the planes but if one of the bounding planes is tangent to the surface, it is a zone of one base. The distance between the planes is the *altitude* of the zone. (2) *Lune*, a part of the surface bounded by the circumferences of two great circles. (3) Spherical pyramid, part of a sphere bounded by a spherical polygon and the planes of its sides. (4) Spherical *sector*, part of a surface generated by the revolution of a circular sector about any diameter of which the sector is a part. (5) Spherical *segment*, a portion of a spherical surface between two parallel planes. (6) Spherical *wedge*, a portion of the surface bounded by a line and two great semicircles.

If A is surface area, the various cases give for it: sphere, $4\pi r^2$; zone, $2\pi rh$; lune, $\pi r^2 a/90$; triangle or polygon, $\pi r^2 E/180$, where r is the radius of the surface, h is the altitude of the zone, a is the number of degrees in the angle, and E is the spherical excess, defined by $E = T - 180(n - 2)$, where T is the sum of the angles and there are n sides to the polygon.

See also **Sphere.**

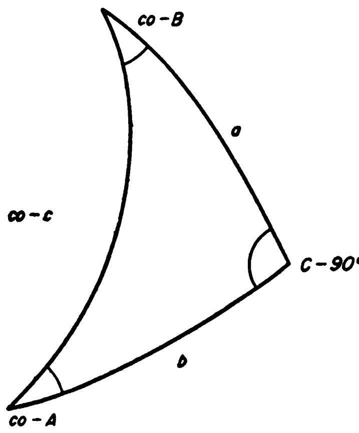
SPHERICAL TRIGONOMETRY. A generalization of plane trigonometry, spherical trigonometry is primarily concerned with the solution of spherical triangles. It is used in many navigation and astronomical problems, as well as in the construction of certain kinds of maps.

Solution of a spherical triangle means the finding of unknown sides and angles from given values for other sides and angles. The right spherical triangle is the simplest case, but it, unlike the plane right triangle, can have two or even three right angles. If, however, it has two right angles, the sides opposite them are quadrants and the third angle has the same measure as its opposite side. If all three angles are right angles, the measure of each side is 90°. These cases are all relatively simple; hence, we will consider a triangle that has only one right angle, its opposite side in general not being 90°.

Let a, b, c be the sides of a right spherical triangle, measured by the angle subtended at the center of a sphere, and with opposite angles A, B, C, where $A = 90°$. Then 10 relations exist, as follows:

$$\begin{aligned} \sin a &= \sin A \sin c & \sin b &= \sin B \sin c \\ \sin a &= \tan b \cot B & \sin b &= \tan a \cot A \\ \cos A &= \cos a \sin B & \cos B &= \cos b \sin A \\ \cos A &= \tan b \cot c & \sin B &= \tan a \cot c \\ \cos c &= \cot A \cot B & \cos c &= \cos a \cos b \end{aligned}$$

Since these equations are rather awkward. Napier's rules are convenient in actual use. They can be stated as follows: let co-A, co-c, co-B mean the complements of the angles (co-$A = 90° - A$) and arrange the parts as shown in the diagram. Then, calling any angle a middle part, it will have two parts adjacent to it and two parts opposite to it. Napier's rules are then: (i) the sine of a middle part equals the product of the tangents of the adjacent parts; (ii) the sine of part equals the product of the cosines of the opposite parts.



Designation of angles in Napier's rules.

In the case of an oblique spherical triangle, the following relations may be obtained. (1) Law of sines:

$$\frac{\sin A}{\sin a} = \frac{\sin B}{\sin b} = \frac{\sin C}{\sin c}$$

(2) Law of cosines: $\cos a = \cos b \cos c + \sin b \sin c \cos A$ and $\cos A = -\cos B \cos C + \sin B \sin C \cos a$. (3) Haversine law: have $a = \text{hav}(b - c) + \sin b \sin c \text{ hav } A$. (4) Half-angle formulas:

$$\sin A/2 = \sqrt{\frac{\sin(s-b)\sin(s-c)}{\sin b \sin c}}$$

and

$$\tan A/2 = \frac{r}{\sin(s-a)},$$

where $2s = (a + b + c)$ and $r^2 \sin s = \sin(s - a)\sin(s - b)\sin(s - c)$. (5) Napier's analogies:

$$\frac{\tan(a+b)/2}{\tan c/2} = \frac{\cos(A-B)/2}{\cos(A+B)/2}.$$

(6) Gauss formulas:

$$\sin(A-B)/2 = \frac{\sin(a-b)/2 \cos C/2}{\sin c/2}.$$

(7) Rule of quadrants: in any spherical triangle, one-half the sum of two angles is in the same quadrant as one-half the sum of the sides opposite. (8) Spherical excess: $\tan^2 E/4 = \tan s/2 \tan(s - a)/2 \tan(s - b)/2 \tan(s - c)/2$. There are additional relations in several of these cases (for example, 3, 4, 5), obtainable by cyclic permutation of A, B, C and a, b, c.

In solving a spherical triangle there are six possible cases, depending on the parts given, as follows: (I) a, b, C; (II) A, B, c; (III) b, c, C; (IV) A, B, a; (V) a, b, c; (VI) A, B, C. Case (III) is an ambiguous case, since there may be two solutions or only one. Each of the six cases may be solved by several combinations of the relations given in the preceding paragraph.

See also **Trigonometry.**

SPHEROID. An ellipsoid. Also called ellipsoid of revolution, from the fact that it can be formed by revolving an ellipse about one of its axes. If the shorter axis is used as the axis of revolution, an *oblate spheroid* results, and if the longer axis is used, a *prolate spheroid* results. The earth is approximately an oblate spheroid.

See also **Ellipsoid** and **Surface (Of Revolution).**

SPHEROIDAL COORDINATE. A degenerate system of curvilinear coordinates obtained from ellipsoidal coordinates when two axes of the quadric surface are equal in length. There are two special cases: *oblate* and *prolate* spheroidal coordinates. In the oblate case, the coordinate surfaces are families of oblate ellipsoids of revolution around the Z-axis with semi-axes $a = c\sqrt{1 + \xi^2}$, $b = c\xi(\xi = \text{const.})$; hyperboloids of revolution of one sheet with $a = c\sqrt{1 + \eta^2}$, $b = c\eta$ (η = const.) and planes from the Z-axis (ϕ = const.). The following additional relations hold:

$$0 \le \xi \le \infty; \quad -1 \le \eta \le 1; \quad 0 \le \phi \le 2\pi$$
$$x = c\sqrt{(1+\xi^2)(1-\eta^2)}\cos\phi$$
$$y = c\sqrt{(1+\xi^2)(1-\eta^2)}\sin\phi$$
$$z = c\xi\eta$$

Alternative variables often used are $\xi = \sinh u$; $\eta = \cos v$; $0 \le u \le \infty$; $0 \le v \le \pi$.

For prolate spheroidal coordinates, the coordinate surfaces are families of prolate ellipsoids of revolution around the Z-axis with semi-axes $a = c\xi$, $b = v\sqrt{\xi^2 - 1}$, ξ = const.; hyperboloids of revolution of two sheets with $a = c\eta$, $b = c\sqrt{1 - \eta^2}$, η = const. and planes from the Z-axis, ϕ = const. The coordinates are limited in range: $-1 \le \eta \le 1$; $1 \le \xi \le \infty$; $0 \le \phi \le 2\pi$ and in terms of rectangular coordinates

$$x = c\sqrt{(1-\eta^2)(\xi^2-1)}\cos\phi$$

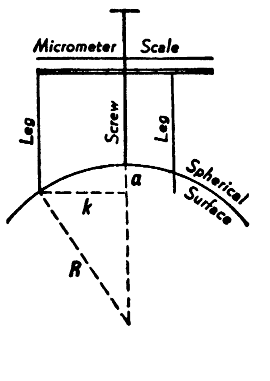
$y = c\sqrt{(1 - \eta^2)(\xi^2 - 1)} \sin \phi$; $z = c\xi\eta$. The variables $\xi = \cosh u$; $\eta = \cos v$; $0 \le u \le \infty$; $0 \le v \le \pi$ are often used.

This system is particularly useful in quantum mechanical two-center problems, for if the centers are taken at the foci of the system, the focal radii to a point where the surfaces of revolution intersect satisfy the relations $(r_1 + r_2) = 2c\xi$; $(r_1 - r_2) = 2c\eta$.

See also **Coordinate System; Curvilinear Orthogonal Coordinates; Ellipsoidal Coordinate** and **Hyperboloid.**

SPHEROMETER. An instrument for measuring the curvature of solid spherical surfaces, either convex or concave, such as those of lenses; a measurement in which high precision is not easily attained. The most familiar mechanical device for this purpose is a form of micrometer. It resembles a small three-legged stool, the sharp steel points of whose legs form an equilateral triangle. The micrometer screw, also with a sharp point, is mounted at the center of this supporting trivet, and is adjusted to read zero when all four points are in one plane, as determined by standing the instrument on a flat plate of glass and screwing the micrometer point down to it. The distance from each of the legs to the central axis must be accurately known. It is called k, and if the elevation (or depression) of the micrometer point to fit a given spherical surface is a, then the radius of curvature of the surface is readily calculated as

$$R = \frac{k^2 + a^2}{2a}$$



Diagrammatic section of spherometer.

The chief source of error are in determining just when contact takes place between micrometer point and surface, and in measuring *k*.

See also **Circular Curves** and **Sphere.**

SPHYGMOMANOMETER. An apparatus for measuring the blood pressure. It consists of a rubber-bag cuff which is wrapped around the upper arm. This is inflated by a hand bulb. The cuff is connected by rubber tubing to a measuring device which is either a sealed column of mercury or a spring scale. Sufficient pressure is pumped into the rubber cuff to compress the brachial artery in the upper arm. A stethoscope is applied over the artery below the cuff and air is gradually allowed to escape from the cuff until the pulse can be heard. The reading on the scale or column of mercury at this point indicates the systolic pressure or the highest pressure in the arteries during contraction of the heart. The deflation of the cuff is continued, and that point on the scale when the last sound of the disappearing pulse is heard is the distolic pressure, or lowest pressure in the artery during diastole, or relaxation of the heart muscle between beats. The normal systolic reading of an adult varies from 110 to 130 or 140 millimeters of mercury. Normal diastolic readings vary from 60 to 90 millimeters of mercury. See also **Hypertension (High Blood Pressure).**

SPICA (α Virginis). Ranking fifteenth in apparent brightness among the stars. Spica has a true brightness value of 2,800 as compared with unity for the sun. Spica is a blue-white, spectral type B star and is located in the constellation Virgo, a zodiacal constellation. Estimated distance from the earth is 260 light years.

Spica is particularly interesting in that it is believed to be the star that provided Hipparchus with the data which enabled him to discover the precession of the equinoxes. The temple at Thebes was oriented with reference to Spica in about 3200 B.C. Later temples which were oriented to this same star indicated the motion of the star due to precession and provided the necessary data.

See also **Constellations.**

SPIDER (*Arachnida, Araneida*). Arthropods of almost exclusively terrestrial habits, known commonly for their ability to spin silken webs. They differ from other arthropods in one or more of the following characters: The body is divided into cephalothorax and abdomen, and the latter is unsegmented. The head bears a group of simple eyes. Four pairs of legs are present. The jaws are perforated by the ducts of poison glands. The ventral surface of the abdomen bears the openings of the lung books, respiratory organs of peculiar form, and a group of spinnerettes through which the ducts of the silk glands open.

Spiders are among the most interesting of all animals, and with few exceptions, e.g., the black widow spider, are harmless. Even though they secrete poison most of them are too small to bite a human being unless on a very thin fold of tissue, and most seem to have no inclination to bite. Even the large hairy species commonly called banana spiders or tarantulas are mild-mannered creatures.

To what extent the bad reputation of the black widow, *Latrodectus mactans*, is deserved seems difficult to establish. Apparently its bite is severely poisonous and occasionally fatal, and apparently it is vicious in its habits. There is contradictory evidence, however, so the case is not wholly settled.

Black widow spider. (*A. M. Winchester.*)

Spiders vary greatly in habits. Some spin funnel-like webs in which they hide to await their prey, others form irregular webs, and still others make the orb webs so beautifully demonstrated in our gardens on dewy mornings. Other forms spin no web but capture their prey by pouncing on it from concealment or by open chase. Among these forms are the crab spiders, named from their short broad form, which lie in wait on plants and are sometimes almost perfectly hidden in flowers by their concealing coloration. The wolf spiders are stout hairy species, often black in color. They hunt like the predators for which they are named.

The spiders that capture prey without the use of webs have other uses for silk, such as the formation of cocoons or egg-sacs in which the eggs are deposited, and the construction of a smooth lining for their hiding places. The most remarkable example of the latter use is the nest of the trapdoor spiders of warm regions. These nests consist of a silk-lined burrow with a beveled margin at the surface of the ground. A lid hinged with tough silk fits perfectly into this beveled depression and can be held shut by the spider, which provides in its lining two depressions to be gripped by the claws.

The mating habits of spiders are also remarkable. The male, in many species much smaller than the female, goes through a courting procedure as complex as that of the birds, and is often killed by his consort. Reproductive adaptations are also peculiar in the males, in that the palpi are modified to convey the seminal fluid to the genital passages of the female. When sexually mature the male spins a web in which the contents of the reproductive organs are discharged, to be taken up into the cavities in the palpi. When the individual is successful in securing a mate he thrusts the palpi one at a time into her genital aperture.

See also **Arachnida.**

Additional Reading

Eisner, T., and S. Nowicki: "Spider Web Protection through Visual Advertisement: Role of the Stabilimentum," *Science*, **219**, 185–187 (1983).
Foelix, R. F.: "Biology of Spiders," Harvard University Press, Cambridge, Massachusetts, 1982.
Hadley, N. F.: "The Arthropod Cuticle," *Sci. Amer.*, 104–114 (July 1986).
Jackson, R. R.: "A Web-Building Jumping Spider," *Sci. Amer.*, 102–115 (September 1985).
Lubin, Y. D., and M. H. Robinson: "Dispersal by Swarming in a Social Spider," *Science*, **216**, 319–321 (1982).
Masters, W. M., and H. Markl: "Vibration Signal Transmission in Spider Orb Webs," *Science*, **213**, 363–365 (1981).
Witt, P. N., and J. S. Rovner, Eds.: "Spider-Communication," Princeton University Press, Princeton, New Jersey, 1982.

SPIDER MONKEY. See **Monkeys and Baboons.**

SPILLWAY. One of the important adjuncts of a dam of the overflow type is a spillway, which is simply an opening through or over which excess water may flow when the reservoir is full. The spillway may be a certain overflow section of the dam, or it may be located at one side of the dam.

SPINACH. Of the family *Chenopodiacea* (goosefoot family), spinach (*Spinacia oleracea*) is a pot herb, cooked and eaten much as other greens, such as chard, turnip greens, and mustard greens. All greens, of course, are not members of the same botanical family. Many are brassicas, members of the *Cruciferae* (mustard) family. Spinach is a hardy cool-weather plant that withstands winter conditions in the southern United States. In most of the northern states, spinach is primarily an early-spring or late-fall crop, but in some areas, where summer temperatures are mild, spinach may be grown continuously from early spring until late fall. Winter culture of spinach is possible only where moderate temperatures, as found in California, prevail.

It is believed that the cultivation of spinach commenced in Persia (Iran) about A.D. 300 or 400. There is no mention of spinach in very ancient records. One writing in China, dating back to A.D. 647, refers to spinach as the "herb of Persia." The vegetable was introduced into Spain by the Moors in about A.D. 1100. By the late 1200s, spinach was known and consumed throughout most of Europe. An English writer referred to the vegetable as "spynoches" in 1390. There are no records, but it is assumed that spinach was brought to the United States by the early colonists. By 1806, three varieties were listed in seed catalogs. The first savoyed-leaf variety was introduced in 1828.

A healthy spinach plant of *Long Standing Bloomsdale variety. (Spinacia oleracea). (Ferry Morse Seed Co.)*

There are two principal variations of spinach: (1) the *savoy* or *wrinkled-leaf* kind; and (2) the *semisavoy* or *flat-leafed* kind. For the fresh market, the savoy type is usually preferred. The semisavoy kind is used for processing. Some botanists also classify spinach in terms of whether a variety is *prickly seeded* or *smooth seeded*. Modern growers greatly prefer the smooth-seeded varieties for ease and precision of planting. A savoy-type plant is shown in the accompanying illustration. One further classification distinguishes the *long-standing* varieties. Long-standing refers to the fact that the plant is slow in bolting—that is, it does not go to seed early, a desirable characteristic for growers.

SPINACH LEAF MINER (*Insecta, Diptera*). One of several leaf-mining flies and maggots that attack a variety of plants. The spinach leaf miner (*Pegomya hyoscyami*, Panzer) attacks beet, chard, mango, spinach, and sugarbeet, as well as a number of weeds, such as chickweed and lamb's quarters. The insect is of much greater economic importance in Europe than in North America. It is believed that the insect was introduced into North America in the early 1880s. The small maggots produce blisterlike blotches on the leaves of the aforementioned plants, rendering them unfit for market. Even where the leaves of such plants may not be used, the roots and seeds do not develop properly if the plant is vigorously attacked by this insect. The adult is a slender, grayish-black, two-winged fly, about $\frac{1}{4}$ inch (6 millimeters) long. The adult emerges in mid-spring and mating and egg-laying occur almost immediately. The eggs are laid on the underside of leaves. The very small maggots immediately commence feeding on leaves when hatched. They migrate from one leaf to the next and thus destruction is not localized. There are from 3 to 4 generations of the insect per year. Effective control can be achieved by destroying nearby host weeds. Screening of the plant beds with cheese-cloth also can be an effective control measure. Spinach planted in early spring or late fall usually is not affected. Where cultural methods are not sufficient, chemical controls, such as spraying or dusting parathion or diazinon in the affected areas will bring about control. Care must be exercised to avoid using chemicals near the time of harvest because of the possible presence of poisonous residuals on the leaves.

SPIN-DEPENDENT FORCE. The force between two particles which depends on their relative spin orientations and possibly on their spin directions relative to the line joining the particles. Physical basis could be the interaction between the magnetic moments of the particles, or in the case of nuclear forces, to the exchange of π-mesons between the nucleons.

SPINE. The vertebral column or backbone, in humans composed of 33 vertebrae which, grouped according to regions, are: 7 cervical; 12 thoracic; 5 lumbar, 5 sacral, and 4 coccygeal vertebrae. See also **Skeletal System.**

SPINEL. The mineral spinel is one of a group of minerals which crystallize in the isometric system with an octahedral habit, and whose chemical compositions are analogous. These minerals are combinations of bivalent and trivalent oxides of magnesium, zinc, iron, manganese, aluminum, and chromium, the general formula being represented as $R''O \cdot R_2'''O_3$. The bivalent oxides may be MgO, ZnO, FeO, and MnO, and the trivalent oxides Al_2O_3, Fe_2O_3, Mn_2O_3, and Cr_2O_3. The more important members of the spinel group are spinel, $MgAl_2O_4$; gahnite, zinc spinel, $ZnAl_2O_4$, franklinite $(Zn,Mn^{2+},Fe^{2+})(Fe^{3+},Mn^{3+})_2O_4$, and chromite, $Fe\,Cr_2O_4$. True spinel has long been found in the gem-bearing gravels of Sri Lanka and in limestones of Burma and Thailand.

Spinel usually occurs in isometric crystals, octahedrons, often twinned. It has an imperfect octahedral cleavage; conchoidal fracture; is brittle; hardness, 7.5–8; specific gravity, 3.58; luster, vitreous to dull; transparent to opaque; streak white; may be colorless, rarely through various shades of red, blue, green, yellow, brown, or black. These colors are doubtless due to small amounts of impurities. The clear red spinels are called spinel-rubies or balas-rubies and were often confused with genuine rubies in times past. Rubicelle is a yellow spinel. A violet-colored manganese-bearing spinel is called almandine spinel.

Spinel is found as a metamorphic mineral, also as a primary mineral in basic rocks, because in such magmas the absence of alkalies prevents the formation of feldspars, and any aluminum oxide present will form corundum or combine with magnesia to form spinel. This fact accounts for the finding of both ruby and spinel together. In addition to the localities mentioned above which yield beautiful specimens, spinel is found in Italy and Sweden and in Madagascar. Also in the United States in Orange County, New York, and in Sussex County, New Jersey, are many well-known spinel localities. Spinel is found also in Macon County, North Carolina, and in Canada in Quebec and Ontario.

The name spinel is derived from the Greek, meaning a spark, in reference to the fire-red color of the sort much used for gems. Balas ruby is derived from Balascia, the ancient name for Badakhshan, a region of central Asia situated in the upper valley of the Kokcha River, one of the principal tributaries of the Oxus.

Elmer B. Rowley, F.M.S.A., formerly Mineral Curator,
Department of Civil Engineering, Union College,
Schenectady, New York.

SPINNERET. A spinning organ of the spiders. The spinnerets are located on the ventral surface of the abdomen, near or at its tip, and vary from one to three pairs. They are conical to cylindrical in form. Each has a membranous terminal portion called the spinning field, through which run many minute spinning tubes from the silk glands. The nature of the spinning tubes varies, different tubes producing different kinds of silk.

In the act of spinning the liquid silk is forced through the spinning tubes, to harden on exposure to the air as silk. The spinnerets bear spines, some of them apparently tactile, and are moved by muscles, so that the spider is able to form and place its threads with precision.

Glass fibers and various synthetic fibers, such as polyesters, are also spun in a process known as *melt spinning*. In modern spinning plants, the polymer is heated and conveyed to the spinning head by means of a melt extruder. If the polymer is initially in the form of quenched chips, it must be thoroughly dried before melting or the molten resin will degrade by ester hydrolysis during the spinning process. The molten polymer in its passage from extruder to jet has a viscosity of the order of 1,000 poises. Pressures in the area range from 1,000 to 5,000 psi (6.9 to 34.5 mPa). The spinneret is made of stainless steel with a number of holes ranging from 14 to several hundred, depending upon the filament count of the yarn to be spun. The holes have diameters ranging from 0.008 to about 0.025 inch (0.2 to 0.6 millimeter), in accordance with the denier being spun. The depth of the holes is 1.3 to 3 times the diameter. See also **Fibers.**

SPIN (Nuclear). The total angular momentum of an atomic nucleus, when it is considered as a single particle.

SPIN STATE. A system is said to be in a definite spin state when the quantum mechanical wave function describing the system is an eigenfunction of the various spin operators corresponding to the square of the total spin angular momentum and the component(s) of the spin angular momentum being used to designate the system. For example, the spin state of a single particle might be designated by specifying the particle's spin projection on its direction of motion. The operator corresponding to this dynamical variable is $\mathbf{s} \cdot \mathbf{p}$ where $\mathbf{s}$ is the vector operator corresponding to the spin angular momentum of the particle and $\mathbf{p}$ is the vector operator corresponding to the linear momentum. For a particle to be in a definite spin state according to this method of designation, it must be described by a wave function which is an eigenfunction of $\mathbf{s} \cdot \mathbf{p}$. As another example, the spin state of a two particle system might be specified in one of two ways—either by specifying S^2 and S_z, the square of the total spin angular momentum and its z-component respectively, or by specifying S_{1z} and S_{2z}, the z-components of the spin angular momentum of each particle. If the case for the two particle system are considered in detail, using non-relativistic quantum mechanics and the Pauli spin operators, one has

$$\begin{aligned} \mathbf{S}_1^2 &= \tfrac{1}{4}\hbar^2\,\boldsymbol{\sigma}_1^2 \\ \mathbf{S}_2^2 &= \tfrac{1}{4}\hbar^2\,\boldsymbol{\sigma}_2^2 \\ \mathbf{S}^2 &= (\mathbf{S}_1 + \mathbf{S}_2)^2 = \mathbf{S}_1^2 + \mathbf{S}_2^2 + \mathbf{S}_1 \cdot \mathbf{S}_2 \\ &= \tfrac{1}{2}\hbar^2\,(\boldsymbol{\sigma}_1^2 + \boldsymbol{\sigma}_2^2 + \boldsymbol{\sigma}_1 \cdot \boldsymbol{\sigma}_2) \end{aligned}$$

and

$$S_z = S_{1z} + S_{2z}$$

with

$$S_{1z} = \tfrac{1}{2}\sigma_{1z} \text{ and } S_{2z} = \tfrac{1}{2}\sigma_{2z}$$

Then letting α represent the column matrix $\binom{1}{0}$ and β represent the column matrix $\binom{0}{1}$, one has

$$\begin{aligned} S_1^2\alpha_1 &= \tfrac{1}{2}(\tfrac{1}{2} + 1)\hbar^2\,\alpha_1 \\ S_{1z}\alpha_1 &= \tfrac{1}{2}\hbar\,\alpha^1 \\ S_1^2\beta_1 &= \tfrac{1}{2}(\tfrac{1}{2} + 1)\hbar^2\,\beta_1 \\ S_{1z}\beta_1 &= -\tfrac{1}{2}\hbar\,\beta_1 \end{aligned}$$

Thus $\alpha_1\alpha_2$, $\alpha_1\beta_2$, $\beta_1\alpha_2$, and $\beta_1\beta_2$ represent spin states with, respectively,

$$\begin{aligned} S_{1z} &= \tfrac{1}{2}\hbar, & S_{2z} &= \tfrac{1}{2}\hbar \\ S_{1z} &= \tfrac{1}{2}\hbar, & S_{2z} &= -\tfrac{1}{2}\hbar \\ S_{1z} &= -\tfrac{1}{2}\hbar, & S_{2z} &= \tfrac{1}{2}\hbar \\ S_{1z} &= --\tfrac{1}{2}\hbar, & S_{2z} &= -\tfrac{1}{2}\hbar \end{aligned}$$

Also $\alpha_1\alpha_2$; $\tfrac{1}{2}(\alpha_1\beta_2 + \beta_1\alpha_2)$; and $\beta_1\beta_2$ all represent spin states with $S^2 = 2^{\breve{}2}$ but $S_z = \breve{}$; $S_z = 0$, and $S_z = -\breve{}$ respectively. The state $\tfrac{1}{2}(\alpha_1\beta_2 - \beta_1\alpha_2)$ is one with $S^2 = 0$ and $S_z = 0$.

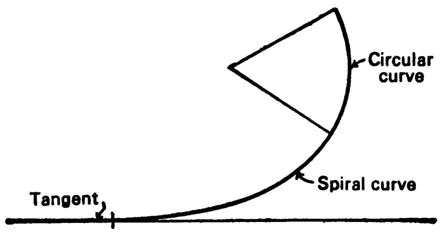
SPINY ANTEATER. See **Monotremata.**

SPIRACLE. 1. A small opening on the surface of the insect body, leading into an air tube of the respiratory system. Insect spiracles are simple openings in some species, but usually they are guarded by some structure which prevents the entrance of foreign particles. This guard varies from a fringe of hairs to an elaborate sievelike plate. Some insects depend on a closing apparatus of the associated trachea for the exclusion of particles.

In the primitive condition, spiracles apparently were paired and metameric, one opening on each side of each segment. This arrangement is only slightly modified in some species, which are said to be peripneustic. In other cases the number of openings is greatly restricted, only one or two pairs providing the sole entrance to the respiratory apparatus. Insects with a single pair at the anterior end of the body are called propneustic, those with a pair at the posterior end metapneustic, and those with a pair at each end amphipneustic. The spiracles may also be associated with tubes, either extending the tracheae beyond the general surface of the body or forming external conduits for air, by which the insect may reach the air from within the water or decaying matter in which it lives. 2. The greatly reduced external opening representing the first of the series of gill slits in elasmobranch fishes.

SPIRAL CURVE. In railway or highway alignments a spiral curve, sometimes called an easement or transition curve, is one which provides a gradual change of curvature when passing from a tangent (straight line) to a circular curve. The cubic parabola and cubic spiral are curves well suited for this purpose. The former is used when spirals are to be laid out by offsets from the tangent, since the offset to any point on the curve varies as the cube of the distance along the tangent from the point where the curve begins. When the curve is to be laid out by deflection angles the cubic spiral is used. In this curve the offset from the tangent is proportional to the cube of the distance along the curve measured from the point of tangency. Spiral curves are particularly well adapted for the use of superelevation since they furnish a means for gradually increasing this quantity from zero to the amount required on the circular portion of the curve. Railroads use a combination of spiral and circular curves to connect the tangents on their main lines. The value of the spiral is now recognized in modern highway design for high-speed traffic.

See also **Superelevation.**



Spiral curve.

SPIRAL (Mathematics). A transcendental plane curve, for which the equation in many cases can be written in polar coordinates and in general form as $r = a_0\theta^n + a_1\theta^{n-1} + \cdots + a_n$. A spiral can also be defined as the locus of a point which moves about a fixed pole, while its radius vector $\mathbf{r}$ and its vectorial angle θ continually increases or decreases according to some law. The best-known cases have special names as follows: Archimedes spiral, Cornu spiral, Sici spiral, hyperbolic spiral, lituus, logarithmic spiral, parabolic spiral.

See also **Curve (Plane).**

SPIRAL VALVE. A thin ridge projecting from the lining of the intestine of elasmobranch fishes to the center of the cavity and following a spiral course through the length of the tube. The structure provides greatly increased surface for the digestion and absorption of food. It is so arranged that the food must follow a spiral course, but beyond this regulating effect is not a valve in the usual sense. It merely slows the passage of food.

SPITTLE BUG (*Insecta, Homoptera*). Of the family *Cercopidae*, the spittle insects, also commonly called frog hoppers because of their remote resemblance to frogs, are quite damaging to certain crops. The immature spittle bug sucks the sap and juices of plants and secretes a protective frothy mass about itself. Control chemicals must be applied at just the right time—about one week after the first appearance of the nymphs. There are no apparent natural enemies to these insects. When infestations are severe, reaching the proportion of several hundred nymphs per plant, as much as 30–40% of a leguminous hay crop and nearly all of an alfalfa seed crop can be lost. Infested strawberry plants also show great reductions in yield as the result of the spittle bug.

The insect winters over as an egg, only about 1 millimeter long. Masses of up to 30 eggs will be deposited in grain stubble several inches above groundlevel. The eggs hatch from early spring to early summer and the first instar nymphs migrate to host plants and attempt to find locations where the humidity will be relatively high. Immediately, the nymph covers itself with the aforementioned secretion, a procedure that is highly effective against desiccation. There are 5 nymphal instars in the cycle, which ranges from 1 to 3 months, largely depending upon temperature.

SPLEEN. An organ of vertebrates derived from mesenchyme and lying in the mesentery. It is closely associated with the circulatory system. The organ consists of masses of tissue of granular appearance, known as lymphoid tissue, located around fine terminal branches of veins and arteries. According to one interpretation, these vessels are connected through the spleen pulp by modified capillaries called splenic sinuses. The pulp is supported by a reticular tissue foundation and contains blood cells of all kinds in addition to the characteristic mesenchymal cells. The functions of the organ are the formation of blood cells, the destruction of old red cells, the removal of other debris from the blood stream, and as a reservoir for blood. It is estimated that the spleen can store from one-fifth to one-third of the total volume of blood. By undergoing periodic contractions during severe exercises, lowering of barometric pressure, and in cases of carbon monoxide poisoning, asphyxia, and hemorrhage, the organ can accommodate emergency calls for blood to be added to the circulatory system.

In humans, the spleen is situated in the left upper part of the abdomen, behind the stomach, just below the diaphragm. The organ, in an adult human, measures about 5 × 3 × 2 inches (12.5 × 7.5 × 5 centimeters) in size. In some diseases, it enlarges (*splenomegaly*) and may even fill a large portion of the left side of the abdomen. Enlargement occurs in certain pathological conditions, such as Banti's disease, Gaucher's disease, and certain anemias. *Splenectomy* (surgical removal of the spleen) frequently gives favorable results in these cases. Enlargement also occurs in malaria, bacterial endocarditis, leukemia, and Hodgkin's disease, among others. In these latter cases, removal of the spleen is medically contraindicated. In persons who have undergone splenectomu, or who are functionally asplenic (such as in sickle cell disease), there are attendant risks of overwhelming bacterimias with certain microorganisms, such as *Streptococcus pneumoniae* and *Haemophilas influenzae* Type B. This is particularly true of children.

The spleen is classified as a ductless gland and is also regarded as one of the centers of activity of the reticuloendothelial system. Its presence is not necessary for life. It may be removed surgically and often is, following abdominal injuries with rupture and hemorrhage from the spleen, or in the treatment of certain blood diseases (hemorrhagic purpura, familial jaundice, etc.), or for the removal of splenic tumors or cysts. Congenital anomalies such as accessory spleens occur, and rarely has the spleen been found to be completely absent.

SPODUMENE. The mineral spodumene is a lithium aluminum silicate corresponding to the formula $LiAlSi_2O_6$ and occurs in monoclinic prismatic crystals, occasionally of very large size. It also occurs massive. Spodumene has a perfect prismatic cleavage often very noticeable; uneven to splintery fracture; brittle; hardness, 6.5–7.5; specific gravity, 3–3.2; luster, vitreous to dull; color, grayish- to greenish-white, green, yellow and purple. Its streak is white; it is transparent to translucent. Spodumene is characteristically a mineral of the pegmatites, and it is found in Sweden, Ireland, Madagascar and Brazil. In the United States it is found especially in the pegmatites of Oxford County, Maine; in the towns of Goshen, Huntington and Chesterfield in western Massachusetts; at Branchville, Connecticut; in North Carolina; in South Dakota in huge crystals and in San Diego and Riverside Counties in California.

The name spodumene is derived from the Greek meaning ash-colored, particularly appropriate for the slightly weathered varieties. Hiddenite, the beautiful emerald-green or yellow-green spodumene that is used as a gem, was named for W. E. Hidden. Kunzite, named in honor of George F. Kunz, is a transparent lilac to rose-colored spodumene from Madagascar and California, and recently as magnificent, large gem crystals of both purple and yellow color from the Hindu-Kush Mountains, Nuristan Province in Afghanistan. Beautiful gem stones are cut from such crystals, but its easy cleavage discourages its use as a wearable gem. Spodumene alters rather readily to a mass of albite and muscovite. The commercial use of spodumene is chiefly as a source of lithium compounds.

See also **Lithium.**

SPONDYLARTHROPATHIES. During the early part of the present century, all inflammatory diseases of or in the region of the joints were called "rheumatoid arthritis." Differences, such as the variations in sites involved, were simply assumed to be a reflection of the rather random character of a single disease. Dissatisfaction with this oversimplified view developed over the years and new names for specific manifestations of such conditions appeared. Rheumatoid spondylitis, for example, came into vogue to identify rheumatoid arthritis of the spine. Today there is an entire group of rheumatic diseases which are not directly related to rheumatoid arthritis. These diseases are called the *spondylarthropathies*. Ankylosing spondylitis is the most common disease of this group, which also includes Reiter's syndrome. Findings during and since the early 1970s have shown an immunogenetic connection between the spondylarthropathies and indeed a relationship with several other diseases. As one authority has observed, the group typified by ankylosing spondylitis may be distinguished from rheumatoid arthritis on historic, geographic, epidemologic, genetic, immunogenetic, clinical, immunologic, pathologic, radiologic, and therapeutic grounds.

In research in this area, human leukocyte antigen (HLA-B27) has been the key to developing the relationship of ankylosing spondylitis, for example, with certain other diseases. Statistics indicate that 90–100% of patients with ankylosing spondylitis are B27-positive; 70–90% of patients with endemic Reiter's syndrome (combination of urethritis, conjunctivitis, and arthritis, first discovered by Reiter in 1916); 80–90% with *Salmonella*-reactive arthropathy; 80% with epidemic Reiter's syndrome; 80% with *Yersinia*-reactive arthropathy; 50–70% with inflammatory bowel disease with sacroiliitis; 50–60% with psoriatic arthropy with sacroiliitis; 40–60% with juvenile chronic polyarthropathy; among other striking comparisons.

Ankylosing spondylitis, one time regarded as an uncommon disease, is now believed to occur at about the same rate as rheumatoid arthritis. Although the B27 gene occurs in about 7% of white people, only about 1.5% develop the disease. The disease is rare among black people. It is believed that available statistics may be low because of earlier misdiagnosis and inaccurate reporting. At one time, the disease was considered predominant among white males under 40 years of age, but it is now suggested that there may be a reasonable balance of occurrence in males and females.

If left untreated, ankylosing spondylitis ultimately renders the spine rigid. Fortunately, based upon the recently collected knowledge pertaining to the nature of the disease and, in combination with early treatment, the prognosis is quite good. Medication is directed toward relieving pain and decreasing inflammation. As of the early 1980s, because of the relative unavailability of detailed controlled studies defining the treatment of the disease, current therapy may be altered as new information is gained.

Psoriatic Arthropathy. The association of psoriasis with psoriatic arthropathy ("psoriatic arthritis") has been a subject of debate for many years. The disease is frequently seen and it is estimated that possibly 20% of individuals with psoriasis (principally with psoriatic nail disease) may develop psoriatic arthropathy. The occurrence in men significantly exceeds that in women. The disease may take the form of asymmetric involvement of both large and small joints. The so-called sausage-shaped digit is considered typical. Although a connection between psoriasis and psoriatic arthropathy has been made, much remains to be learned. It is known, for example, that some forms of psoriatic arthropathy are distinguished only with difficulty from rheumatoid arthritis and that the latter can be present in a patient with psoriasis coincidentally and with no connection between the two diseases.

SPONGES. See **Porifera.**

SPORE. A special type of reproductive cell that develops directly into a new plant. Spores are of many kinds. In the Thallophytes they may be asexual cells or zygotes. Thick-walled spores formed after the union of isogametes are called zygospores. Oöspores are fertilized eggs. In the higher plants spores are always produced by diploid plants and develop into haploid gametophytes. In Selaginella and the seed plants small spores, microspores, produce male gametophytes; large spores, megaspores, produce female gametophytes. Pollen grains in the seed plants develop from microspores.

Some bacterial cells become thick-walled spores. This allows them to survive unfavorable conditions. Fungi also develop spores. See **Fungus.**

SPOROPHYLL. A spore-bearing leaf. It actually bears the sporangia which contain the spores. It may be greatly modified in structure and appearance.

SPOROTRICHOSIS. This disease of worldwide distribution is caused by the dimorphic fungus *Sporothrix schenckii*, a fungus that thrives well in a variety of soils and decaying vegetation. It habitutates areas of temperate and tropical climates. Exceptionally rich sources of the fungus are found in sphagnum moss, barberry or rose thorns, some soils, and splinters from rotting wood. Infection usually results from subcutaneous inoculation of the infectious spores as the result of contacting a sharp object that contains the spores. The disease is only secondarily contracted by inhalation of spores. Incidence of the disease is higher among males than females. The disease usually is manifested by subcutaneous nodules with an overlying purplish or pinkish coloration. This is followed by involvement of the draining lymph nodes. The disease progresses slowly. The cutaneous-lymphatic form of sporotrichosis usually responds well to treatment with a saturated solution of potassium iodide administered 3 or 4 times daily. Iodide therapy may be required for approximately one month. There are some side effects and the treatment should be supervised by a physician. Where there is an intolerance to iodide therapy, amphotericin B, a microbial agent used in treating other fungus infections, may be administered. A small percentage of patients display pulmonary manifestations of the disease. The clinical signs are almost identical with those of tuberculosis.

R. C. V.

SPRAIN. A variety of injuries in or about a joint which occur when the movement of the joint is carried beyond its normal range, or forcibly in a direction where its range is limited. Sprains occur in the ankle, knee, wrist, elbow, and spine, in that order of frequency. An injury in which a sudden wrenching or twisting produces tearing of the ligaments or tendons is a common form of sprain. Displacement of cartilages between joints, tearing of muscles around the joint, and tearing of the synovial membrane are others. The symptoms of a sprain are those of inflammation with pain, swelling, and limitation of motion of the part.

Treatment consists of immobilization of the joint with an adhesive strapping or an elastic bandage, elevation of the extremity, and sometimes aspiration of the joint cavity if an effusion or outpouring of inflammatory fluid occurs.

SPRAY DRYING. A process used in the production of numerous chemical and food products. It is widely used in connection with the production of powdered milk and instant coffee preparations. The spray drying is unique among dryers in that it dries a finely divided droplet by direct contact with the drying medium (usually air) in an extremely short retention time (3 to 30 seconds). This short contact time results in minimum heat degradation of the dried product, a feature that led to the popularity of the spray dryer in the food and dairy industries during its early development. In the case of coffee extract, water in the feed will range from 50 to 70%.

Atomization. Inasmuch as the spray dryer operates by drying a finely divided droplet, the feed to the dryer must be capable of being atomized sufficiently to ensure that the largest droplet produced will be dried within the retention time provided. There are different requirements on the degree of atomization needed to result in the desired product. These factors including minimizing the fine and/or coarse fractions, controlling particle dryness, and controlling bulk density. All commercial atomizers, whether of the centrifugal-wheel, pressure-nozzle, or other types, will produce a particle-size distribution that follows a probability curve. As the total energy input increases, the average particle size will decrease and the particle-size distribution will improve, i.e., the spread between the largest and smallest particles will be less.

Centrifugal-wheel Atomizers. The wheel consists of a disk which is rotated at very high speed (1700-50,000 revolutions per minute). See Fig. 1. Feed generally is introduced to the center, with centrifugal force dispersing the feed and throwing out a thin film to the periphery. As the film leaves the disk, it breaks up into a thread, which in turn forms droplets. The disk is located in the hot-air stream so that even though droplets are thrown toward the wall of the dryer, the hot air travels cocurrently and dries the particle sufficiently to prevent wall build-up upon contact. A spray dryer with a wheel atomizer must be relatively large in diameter and shorter than a dryer with pressure nozzles.

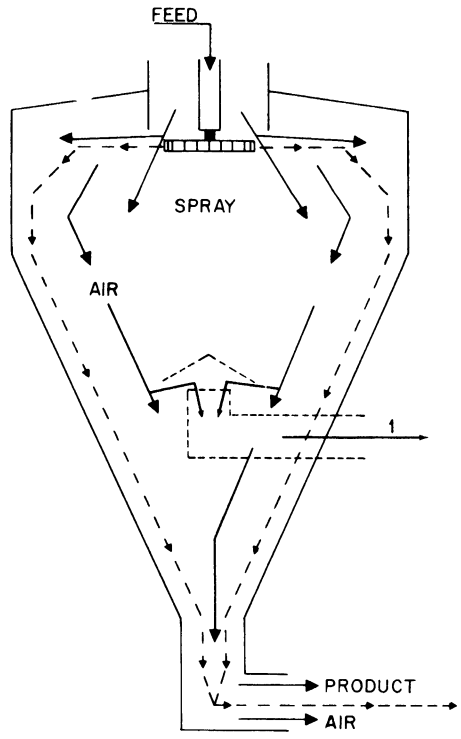


Fig. 1. Spray dryer with wheel atomizer: (1) Air outlet when drying chamber is used for initial separation.

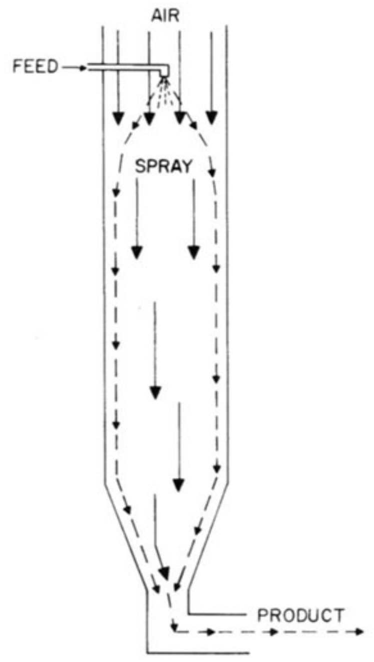


Fig. 2. Spray dryer featuring parallel flow.

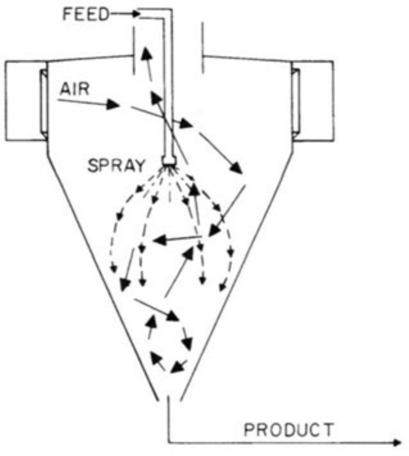


Fig. 3. Mixed-flow spray dryer.

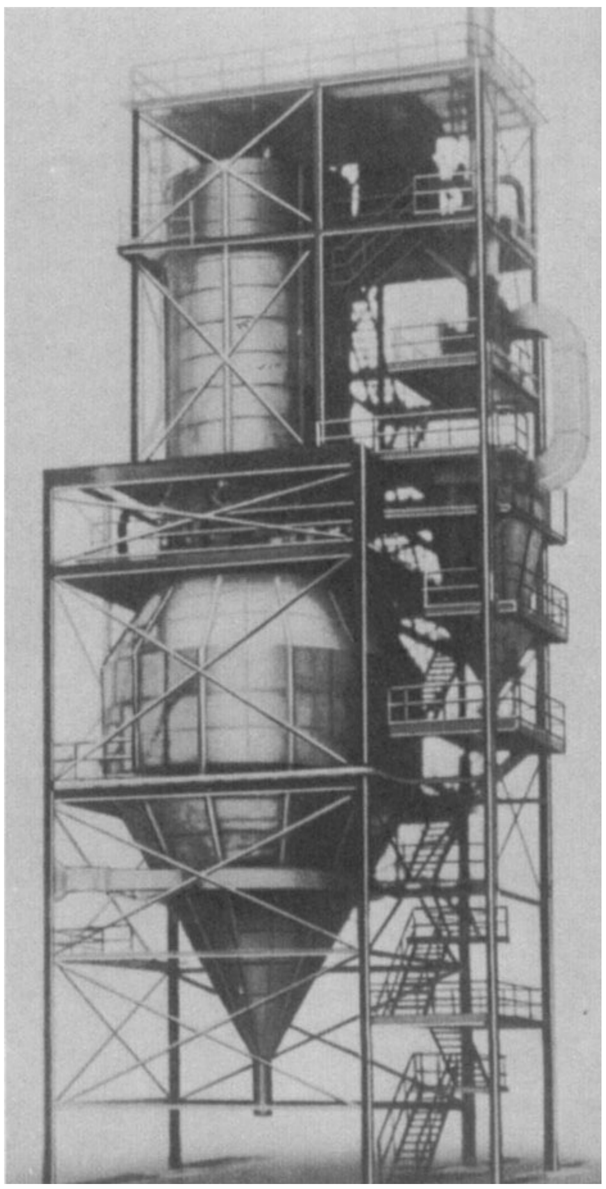


Fig. 4. Tall-form chamber employing nozzle atomization. System is particularly suited for dense particles requiring high-pressure atomization. (*Stork-Bowen.*)

Pressure-Nozzle Atomizers. This system consists of an orifice placed after a fixed mechanism, called a core, swirl chamber, or whizzer, depending upon the manufacturer. A high-pressure pump moves the feed to the nozzle body at a pressure of from 250 to 8000 pounds per square inch (17 to 544 atmospheres). The feed slurry is pumped through high-pressure piping to the whizzer, where a spin is imparted to the fluid before it enters the nozzle orifices. This results in a hollow-cone spray which throws droplets either cocurrent or countercurrent to the air flow.

The flow pattern is such that a cocurrent spray dryer must be relatively long and small in diameter (Fig. 2), whereas a countercurrent dryer is shorter and larger in diameter. A third type, sometimes referred to as a mixed-flow dryer (Fig. 3), uses an air pattern similar to a cyclone collector, i.e., the spray is introduced at the upcoming air stream (countercurrent) and the particles transfer to the air sweeping the wall (cocurrent).

A multistory, tall-form dryer chamber using nozzle atomization is shown in Fig. 4.

References

Charm, S.E.: "Fundamentals of Food Engineering," 3rd edition, AVI, Westport, Connecticut, 1978.

Considine, D. M. (editor): "Foods and Food Production Encyclopedia," Van Nostrand Reinhold, New York, 1982.

Flink, M. M.: "Energy Analysis in Dehydration Processes," *Food Technology*, **31**, 3, 77-84 (1977).

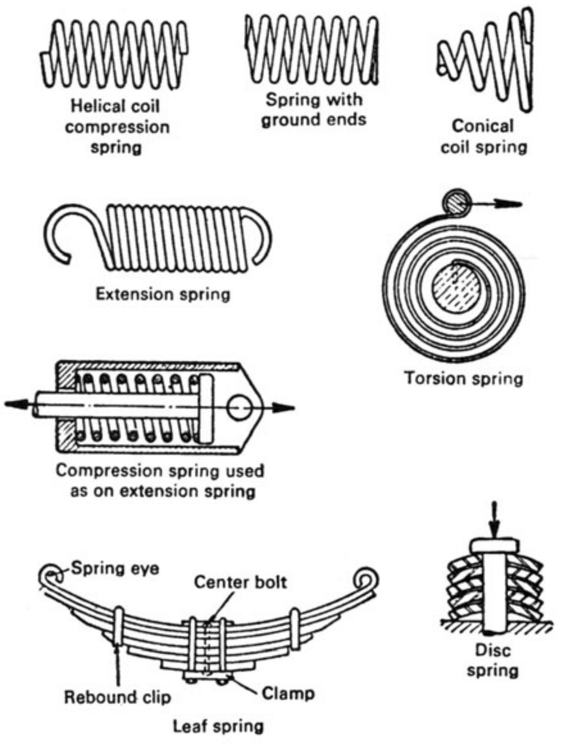
Harper, W. J., and C. W. Hall: "Dairy Technology and Engineering," AVI, Westport, Connecticut, 1976.

SPREADING COEFFICIENT. A thermodynamic expression for the work done in the spreading of one liquid on another. It is the difference between the work of adhesion between the two liquids and the work of cohesion of the liquid spreading, which may be expressed by the equation

$$F_s = \gamma_B - \gamma_A - \gamma_{AB}$$

where F_s is the spreading coefficient, γ_B is the surface tension of the stationary liquid, γ_A is the surface tension of the spreading liquid, and γ_{AB} is the interfacial tension between the liquids.

SPRING. Devices used to absorb energy or shock as in automobile springs; to serve as a source of power as in clocks or watches; and to provide a force to maintain pressure between contacting surfaces as in friction clutches. Springs with ground ends are generally more satisfactory than those with plain ends; compression springs are more desirable for heavy loads than extension springs, because of the possibility of stress concentration in the loop of the extension spring. Compression springs can, however, be employed for tensile loading. Conical coil springs, if properly designed, may be compressed flat under load. Disk springs represent a recent development that is being extensively employed for heavy loads. Laminated or leaf springs are used in vehicles of various types, although coil springs are now being used in automotive applications. Coil springs may be made of square, rectangular, or round wire.



Types of springs.

SPRING CLOCK. A clock that uses a mainspring to power its balance wheel. Until the fifteenth century, all European clocks were weight-driven. Early spring clocks, developed before the seventeenth century invention of the balance wheel and hairspring by Huygens, had no effective control over the mainspring's progressive loss of power as it unwound. Consequently, first the *fusee* and later the *stackfreed* were developed for equalizing the spring's declining force.

The term fusee is derived from the Latin *fusata*—a cone-shaped device spirally wound with a thread, attached to the spring, designed to compensate through leverage for the declining power of the unwinding clock spring. The term stackfreed is probably derived from the German—a device for compensating through leverage for the declining force of an unwinding clock or watch spring, employing a pinion fixed to the arbor around which the spring is coiled and a second spring with an eccentric cam.

Although not accurate, the resulting clocks were useful because, unlike weight-driven clocks, the spring clocks worked continuously while and after being moved; that is, they were portable. Huygens used the hair-spring to control the periodic oscillation of the balance wheel which, in turn, determined the periodic freeing of the pallet from the escape wheel. The latter rotated to work the clock's wheel train which operated the hands on the dial.

The eighteenth century English horologist John Harrison designed and built the first accurate spring clock for use as a ship's chronometer. Harrison used a mainspring-powered balance wheel movement (incorporating a fusee) set in a watchlike case about 5 inches (12.5 centimeters) in diameter. In 1770, Thomas Mudge of England invented the detached lever escapement that freed the balance wheel from direct contact with the escape wheel, thus completing the basic mechanical clock movement. This basic movement (without a fusee), versions of which still are used in clocks, was miniaturized in the twentieth century for use in small jeweled-lever travel clocks and wristwatches. The mainspring-powered balance wheel normally oscillates two and one-half times a second, causing the jeweled lever escapement to tick five times a second.

Spring-operated timing devices cover a range of intervals which may vary from a fractional second to a week, or a month or more. The typical lever-type escape movement used today is dependent for its accuracy on precision of manufacture, temperature compensation, and the effect of external loads. Spring movements are limited to driving light loads because of low torque and lack of constant torque. Accuracy may be within 0.2 seconds absolute in a good stopwatch and 5 seconds per day in a good clock movement.

William O. Bennett, John J. Carpenter, Frank Dostal, and E. Van Haaften, New York.

SPRING INDEX AND SPRING RATE. The spring index is the ratio between the mean diameter of the coil of a helical compression spring and the diameter or radial thickness of the spring wire or strip. The spring index is used as a modifier in computing the safe load a spring may carry. For the majority of spring types, the deflection of the spring is proportional to the load on it. The ratio of load to deflection, which may have the units of pounds per inch or inch-pounds per radian, is known as the spring rate, spring constant, or spring scale.

SPRUCE TREES. Members of the family *Pinaceae* (pine family), these trees are of several species. Spruces are of the genus *Picea*. In terms of tree population and regional areas covered, the spruces exceed the fir trees. See also **Fir Trees.** Some of the species of spruce are extensively used for pulp wood in paper production. The fiber length of the Sitka spruce is a little over 3 millimeters ($\frac{1}{8}$ inch) in length, putting it just behind the Douglas fir and the longleaf pine in this respect; and considerably ahead of the jack pine and lodgepole pine, all used as pulp woods.

Important species of spruce trees not listed on accompanying tables include:

Chinese spruce	*P. asperata*
Colorado spruce	See Blue spruce
Dragon spruce	See Chinese spruce
Eastern Himalayan weeping spruce	*P. Spinulosa*
Himalayan weeping spruce	*P. smithiana*
Japanese tiger-tail spruce	*P. polita*
Likiang spruce	*P. likiangensis*
Oriental spruce	*P. orientalis*
Sargent spruce	*P. brachytyla*
Schrenk's spruce	*P. schrenkiana*
Serbian spruce	*P. omorika*
Weeping spruce	See Brewer's spruce
Yezo spruce	*P. jezoensis*

Spruce trees are considered very hardy. They are evergreen trees, often conical. The record spruce trees in the United States are detailed in Table 1. See also **Conifers.**

The white spruce (sometimes called cat spruce or skunk spruce) is a tall tree, usually attaining a height of 70 to 100 feet (21 to 30 meters) in favorable circumstances, somewhat resembles the balsam fir in color and contour. The numerous branches tend to be pendulous. When the needles are bruised, they exude a pungent and unpleasant odor. In addition to pulp wood for paper production, the light, soft, straight-grained and pale yellow wood is used for interiors, flooring, and general construction. The tree ranges from Labrador, Newfoundland, and Nova Scotia all the way west across Canada to British Columbia and is found in the northern parts of the northeastern United States, extending

TABLE 1. RECORD SPRUCE TREES IN THE UNITED STATES[1]

	Circumference[2]		Height		Spread		
Specimen	Inches	Centimeters	Feet	Meters	Feet	Meters	Location
Black spruce (1972) (*Picea mariana*)	60	152	83	25.2	20	6.1	Minnesota
Blue spruce (1982) (*Picea pungens*)	167	424	148	45.1	34	10.4	Colorado
Brewer spruce (1964) (*Picea brewerana*)	164	417	170	51.8	39	11.9	Oregon
Englemann spruce (1970) (*Picea engelmannii*)	290	737	179	54.6	43	13.1	Idaho
Norway spruce (1976) (*Picea abies*)	185	470	108	32.9	55	16.8	New Hampshire
Red spruce (1965) (*Picea rubens*)	165	419	110	33.5	50	15.2	North Carolina
Sitka spruce (1973) (*Picea sitchensis*)	630	1600	216	65.8	93	28.3	Oregon
White spruce (1975) (*Picea glauca*)	116	295	128	39.0	25	7.6	Minnesota

[1]From the "National Register of Big Trees," The American Forestry Association (by permission).
[2]At 4.5 feet (1.4 meters).

westward at that latitude as far as the Rocky Mountains. It is also found in the mountainous regions of Massachusetts and Connecticut.

The timber of the red spruce is quite strong. The tree sometimes attains a height of about 100 feet (30 meters). Near the center of the tree, the branches are essentially spread horizontally, but the upper branches ascend at an angle of about 45°. This tree ranges from New foundland westward through Pennsylvania and on to Minnesota. It is found in the White Mountains of New England as high as 5000 feet (1500 meters) and in the Adirondacks and Alleghany Mountains to about 4000 feet (1220 meters). The tree ranges along the Alleghany Mountains as far south as Georgia. Some of the better timber from this tree is used for piano sounding boards, but the majority is used for pulp wood, for sheathing and flooring in building construction. Early in this century, the red spruce was threatened by over-harvesting, averted by the passage of the Weeks Act in 1911 for the purchase of national forest at the headwaters of navigable streams. This provided government supervision and control over about one-hundred thousand acres in the White Mountain region.

Closely related, but smaller than the red spruce, the black spruce (or bog spruce) is a smaller tree found in the mountains of the southern states. This tree ranges widely from Labrador, Newfoundland, and Nova Scotia westward past Hudson Bay and on across Canada into southern Alaska. Occurrence in the south is essentially dictated by the presence of swampy regions where it thrives. Black spruce is also a pulp wood and is also used for heavy construction, such as piles, posts, and ship construction.

The Sitka spruce is a western tree, ranging from northern California into Alaska and at elevations ranging from sea level to about 3000 feet (900 meters). Normally, under favorable conditions, the tree may reach a height of some 150 to 180 feet (45 to 54 meters). A Sitka spruce is illustrated in the accompanying photo. Parts of fallen Sitka spruces have been found, indicating the trees once approached 300 feet (90 meters) in height. The rugged characteristics of this tree have made it a favorite for planting forests in northern Europe.

The Englemann spruce is also a western tree and is found at high elevations from northwestern Canada (Alberta and British Columbia) southward in a wide swath through Montana, Idaho, Washington, and Oregon to Arizona and New Mexico. The timber from the tree finds general construction use, but is not as strong as most timbers. The blue spruce (sometimes called Colorado blue spruce or silver spruce) is of light coloration, with a distinctive light bluish-gray color, appearing silverish in some light. The tree is a favorite for gardens and landscaping. Normally, the tree does not attain a height much in excess of 40 feet (12 meters) under favorable conditions, but there are exceptions as will be noted from Table 1. The natural region of the tree includes the Colorado and eastern Utah Rocky Mountains and northward into Wyoming. However, the tree has been planted widely throughout the United States. The color tends to vary with location, sometimes becoming more of a bluish-green or light green.

Brewer's or the weeping spruce is also a western tree, found from northwestern California northward into Oregon. Some trees also are found in the Coastal range on the northern slopes. The tree normally attains a height of less than 75 feet (22.5 meters), but can well exceed this figure as shown in Table 1. The tree branches downward to the ground, giving it the weeping appearance. The tree prefers altitudes between 4000 and 8000 feet (1200 and 2440 meters).

The tallest native tree of Europe is the Norway spruce, attaining a height of about 200 feet (60 meters) under favorable conditions. It is the common Christmas Tree of Europe. After the Ice Ages had eliminated the tree from Britain, it was absent until about the year 1500 when it was reintroduced to the British Isles. The tree ranges from Norway southward and eastward, reaching the Italian Alps. The tree was introduced into North America many years ago and can be found in parks and private grounds of northern cities. Another rugged European spruce is the Serbian spruce which is mainly found in the mountainous regions of eastern Yugoslavia. It is a very fast growing tree. The oriental spruce actually is not from the orient, but is native to the Caucasus mountains. It is a popular garden species in Europe.

Asian spruces include the Sargent spruce which is found in western China. Some authorities compare it favorably with the Norway spruce. Because of its attractive flowers, the Likiang spruce, found in China's southwest Yunnan province, is planted mainly for reasons of decor. The Himalayan weeping spruce is often compared with the North American weeping spruce (Brewer's spruce). Schrenk's spruce is found in central Asia.

The engineering characteristics of some spruce timber are given in Table 2.

Spruce Budworm For many years, the spruce budworm (*Choristoneura fumiferana*) has been recognized as one of nature's most destructive forest insects. The first documented epidemic occurred in 1704 and since there have been a number of epidemics, each lasting from 5 to 11 years. Authorities believe that the insect has been infecting the spruce-fir forest at intervals of 50 to 100 years as far back as the Ice Age. A major epidemic occurred over the years 1910 to 1920 that spread through Quebec, New Brunswick, Maine, and northern Minnesota, with an estimated destruction of 225 million cords of pulpwood. In some areas, 90% of the fir trees were reported killed during that epidemic. During relatively recent decades, widespread aerial spraying of chemical insecticides was the remedial measure of choice. But, despite extremely large expenditures for materials and personnel for undertaking such massive spray projects, chemical pesticides proved only partially successful in suppressing the budworm. In the early 1980s, forest managers turned to integrated pest management by which all available necessary control techniques are consolidated into a con-

TABLE 2. ENGINEERING DATA ON SPRUCE TREES

Common Name for Species	Green Condition			Air Dried (12% Moisture)		Maximum Crushing Strength (Parallel to Grain)		Maximum Tensile Strength (Perpendicular to Grain)	
	Moisture Content (Percent)	Weight/ Cu. Foot (Pounds)	Weight/ Cu. Meter (Kilograms)	Weight/ Cu. Foot (Pounds)	Weight/ Cu. Meter (Kilograms)	(Psi)	(MPa)	(Psi)	(MPa)
Eastern spruce	45	34	545	28	449	2600	17.9	200	1.4
Englemann spruce	80	39	625	23	368	2190	15.1	240	1.7
Sitka spruce	42	33	529	28	449	2670	18.4	250	1.7

SOURCE: U.S. Forest Products Laboratory.

Sitka spruce located at Seaside, Oregon. (*W. Gucterian, Portland, Oregon.*)

certed program to manage insect populations in ways that avoid or reduce economic damage and minimize adverse environmental side effects. An important part of integrated pest management is a study of the dynamics of the spruce budworm itself. What is its role and how does it interact in the forest ecosystem?

As pointed out by M. J. Jones (*American Forests*, 18–23, June 1980), "The budworm itself is part of an integrated system, a natural cycle that assures the continual regeneration of the spruce-fir forest. Balsam fir and red spruce enjoy a shared role in the coniferous stands north of the pine belt and south of the fir belt in the east. These two species differ in character, yet cooperate in returning land to a coniferous state after a major disturbance, such as windthrow, logging and insect attack. Balsam fir begins to produce seed early (at 15 years) and often (every 2 years). Fir seedlings usually are abundant and grow rapidly when competing trees are removed. Red spruce begins to produce seed later (at 25 years) and less frequently (every 3 to 8 years), and does not grow up as rapidly. Thus, a disturbance in a spruce-fir stand tends initially to shift the composition balance toward fir. And balsam fir is the preferred food of the spruce budworm, despite its name. To a lesser extent, the budworm also attacks red, white, and black spruce, and occasionally tamarack and hemlock. In the very act of eating, the budworm prepares for future meals. In stands where it reaches epidemic levels, the bug destroys the host species in such a way that ensures the development of a new stand of that same species—for future infestation. The factors that contribute to an outbreak are not completely understood, but it is believed that for budworm populations to explode, mature and overmature fir must be combined with a series of warm, dry summers. Where these conditions coincide, budworms thrive and multiply tenfold each year. At the same time, the fir produces prolific seed and establishes a new stand."

In addition to other biological approaches used in recent years, pheromones (sex attractants) have been synthesized for use as bait in trapping amorous males. Pheromones and other biological controls are discussed in entry on **Insecticide and Pesticide Technology.**

SPRUE (Tropical). A disease that mainly occurs in India, Puerto Rico, and Vietnam. The disease features a malabsorption of two or more substances—fats, proteins, carbohydrates, minerals, vitamins, and even water. These losses occur in the feces in an acute diarrhea with pale bulky stools. Steatorrhea is present in 95% of cases. There is abdominal distention with colic, weakness and wasting, megaloblastic anemia and edema.

The disease is now considered to be caused by colonization of the upper intestinal tract by coliform bacilli. This, in turn, produces lesions based on the villi of the jejunum and ileum which broaden and fuse.

Normally self-limiting, sprue has caused mortality up to 35% in some villages in India. Treatment involves antibiotics—tetracycline and ampicillin—together with folic acid and Vitamin B_{12} therapy.

R. C. V.

SPUR-FOWL. See **Partridge.**

SPUR GEARING. This form of toothed gearing is used for transmitting power between shafts whose axes are parallel. The velocity ratio of a spur gear set is the ratio of the number of revolutions of one gear to the number of revolutions of the other. The pitch circles of a pair of spur gears are those imaginary circles that are equivalent to the peripheries of a pair of friction wheels that would operate without slipping at the same velocity ratio and center distance as the gears themselves. The point of tangency P of the pitch circles on the line of centers is the pitch point of the gearing. The smaller of the two gears is usually referred to as the pinion.

SPURIOUS CORRELATION. Correlation that is misleading. Take two variables x and y which are independent. Consider a third variable z which is correlated with x and with y. Form $v = f(x, z)$ and $w = f(z, y)$, then frequently v and w will show considerable dependence which may be traced to the correlation of x with z and y with z. In general, then, spurious correlations may be defined as correlation which is introduced by other variables rather than the ones under study.

The real question at issue in correlation is actually this: Are the variables in which we are interested x and y or v and w? If the variables are x and y then the correlation of y with w is spurious. If the casual variables, however, are v and w and not x and y, then the correlation of v and w is valid and not spurious. The term "spurious" is perhaps itself misleading. It does not imply that the correlation does not exist; only that it may be due to a rather circuitous train of casual influences.

Sir Maurice Kendall, International Statistical Institute, London.

SPURIOUS RADIATION. Any radiation from a transmitter other than that produced by the carrier and its normal sidebands. A radiated harmonic of the carrier is one example of a spurious radiation.

SPUTNIK. The first artificial satellite, one of a series of Russian earth-orbiting satellites, launched on October 4, 1957.

SPUTTERING. 1. In a gas discharge, material is removed, as though by evaporation, from the electrodes, even though they remain cold. This phenomenon is known as sputtering. 2. The term is also used for the corresponding phenomenon when the discharge is through a liquid. In the first case, sputtering is a nuisance that limits the life of a device; in the second case, it is put to work to make colloidal solutions of metals. 3. A result of the disintegration of the metal cathode in a vacuum tube due to bombardment by positive ions. Atoms of the metal are ejected in various directions, leaving the cathode surface in an abraded and roughened condition. The ejected atoms alight upon and cling firmly to the tube walls and other adjacent surfaces, forming a blackish or lustrous metallic film. This effect is often utilized to form very fine-grained coatings of metal upon surfaces of glass, quartz, etc., purposely exposed to the sputtering. Films of different metals can be obtained by using cathodes made of these metals. Glass plates may be thus silvered, or suspension fibers of spun quartz rendered conducting for use in electrometers, etc.

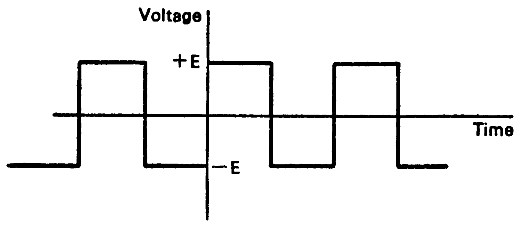
SQUALL. See **Fronts and Storms.**

SQUARE AND SQUARE ROOT. The square of a number or quantity is the product of that number or quantity when multiplied by itself. Hence, 4×4 yields 16, the latter being the square of 4. Similarly, the number 4 is termed the square root of the number 16. The process of raising any number to any integral *power*, as in squaring (a power of 2) or in cubing (a power of 3) is known as *involution*. The process of extracting roots, that is, of finding that 4 is the square root of 16, or of finding that 3 is the cube root of 27, is known as *evolution*. See also **Exponent; Number Theory; Radical (Mathematics);** and **Root (Mathematics).**

SQUARE LAW MODULATOR. A device whose output is proportional to the square of its input. The carrier and modulating signal are added in the input to produce a modulated carrier in the output.

SQUARE WAVE. A square wave, as the name indicates, has the wave form shown in the figure.

Any periodic wave, regardless of its shape, may be analyzed into a series of sine and cosine components whose frequencies are harmonically related. The number of these components will be determined by the shape of the wave, but in general the sharper the corners of the original wave the more component terms. Thus a square wave will require a wide range of frequencies to express it. These components are not mere mathematical fictions but are true electrical components in the case of an electric wave. They may be separated and examined by means of proper filter circuits. Since a square wave will contain a long series of frequencies it may be used for rapidly determining the frequency response of a piece of equipment by applying the wave to the input and noting the distortion of the output wave. The distortion is due to certain frequencies of the original wave being attenuated or amplified out of proportion in passing through the circuit. Thus the necessity of making a laborious series of tests at various frequencies using sine waves is avoided. When an operator is properly trained in interpreting the results of such testing it offers a rapid means of checking amplifiers, networks, etc. These square waves may be generated by a variety of electronic circuits.



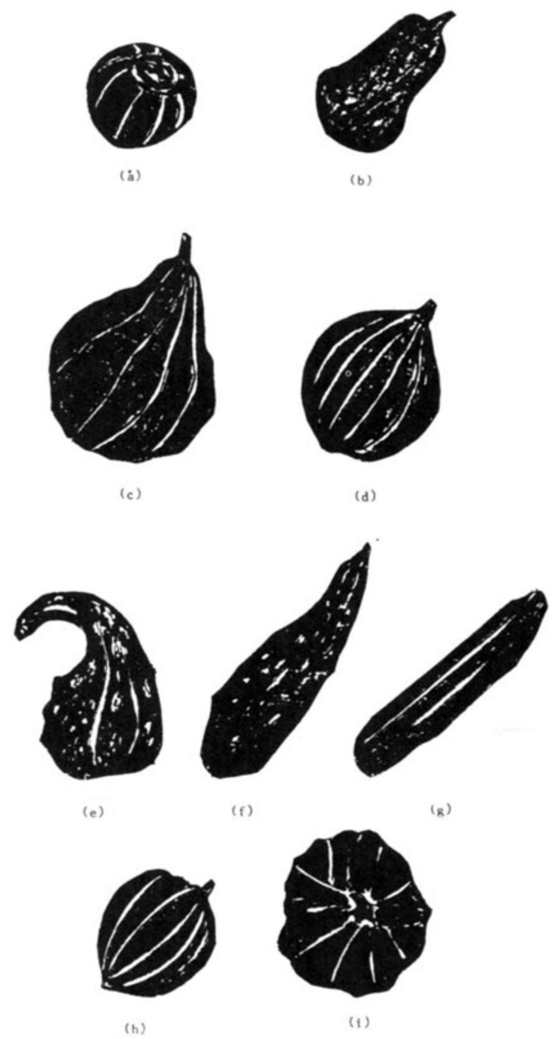
Square wave.

SQUASH. Of the family *Cucurbitaceae* (gourd family), squash plants are of three major species and one minor species of the genus *Cucurbita*. These several species, plus the designations of summer and winter squash, tend to complicate a classification of these plants. There is not a direct relationship between species or growth pattern and whether a plant is a summer or a winter variety. Although a majority of squash plants assume an indeterminate growth pattern as vinelike, tendril-bearing herbs, some take the more determinate form of bushes or semibushes. In terms of food value, the winter squash is rated very high among all vegetables. Winter squash usually is baked or used in pies. Varieties of winter squash more closely resemble the pumpkin, and canning and freezing processes are very similar. Many, but not all varieties of winter squashes are members of the species *Cucurbita maxima*. Some also are members of *C. moschata* and *C. mixta*. Varieties of *C. maxima* and *C. moschata* can be crossed artificially, but such crossing does not occur naturally in the field.

Specific Varieties. The genus *Cucurbita* is indigenous to the Americas. The largest concentration of wild species is in Mexico, in a vast area from just south of Mexico to the border between Mexico and Guatemala. The cultivated species *C. pepo, C. mixta,* and *C. moschata* are North American, whereas *C. maxima* is from South America. Because of the uncharted movement of pre-Columbian peoples and their crops in the Americas, it is difficult to pinpoint the exact area of origin of the cultivated species. One authority suggests that *C. pepo* is native to the southwestern United States and northern Mexico. *C. moschata* and *C. mixta* are lowland species ranging from Vera Cruz, Mexico, southward through Central America. In this area, cultivars of *C. mixta* are much used for their tasty, edible seeds.

Through years of cultivation, research, and experimentation, numerous varieties of squash have been developed. See accompanying illustration.

Pollination. Squash plants have unisexual flowers. In order to produce fruit, squashes require cross-pollination—that is, the transfer of pollen from the anthers of the male flower to the stigma of the female flower. Inadequate pollination results in reduced yield and misshapen fruit. Bees provide the primary means of pollen transfer, and often beehives are used. Experienced growers will have one hive for every 3 to 5 acres (1.2 to 2 hectares) of squash plants. Beehives are placed in the center of the field so that bees do not have to travel more than a few hundred feet to feed. Obviously, there should be no insecticide applications during the flowering period.



Various types of squash. Winter squash: (a) Buttercup; (b) Butternut; (c) Hubbard; (d) Gold Nugget. Summer squash: (e) Crookneck; (f) Straightneck; (g) Zucchini; (h) Acorn; (i) Scallop.

Nutritional Aspects. Winter squash is high in vitamin A. Summer squash is a good source of vitamin C. Squash has fair amounts of iron. It is low in sodium and protein. Squashes, particularly the summer varieties, are considered low-calorie foods.

Prepared squash is one of the more popular baby foods. See accompanying table.

AVERAGE NUTRIENT VALUES OF SQUASH BABY FOOD[1]

Nutrients	Per 100 grams
Calories	26
Protein (grams)	0.8
Carbohydrate (grams)	5.1
Fat (grams)	0.2
Crude fiber (grams)	0.8
Total solids (grams)	7.2
Calcium (milligrams)	19
Phosphorus (milligrams)	15
Iron (milligrams)	0.2
Sodium (milligrams)	108
Potassium (milligrams)	180
Vitamin A (international units)	1270
Thiamine (milligrams)	0.02

Note: 100 grams = 7 tablespoons.
[1]Gerber Products Company, Fremont, Michigan.

SQUASH BORER (*Insecta, Lepidoptera*). The larva of a moth, *Melittia satyriniformis*, which bores in the root and stem of squash vines, often killing the plant. The moth is a beautiful species with olive fore wings, transparent hind wings, and legs tufted with orange and black.

The larva thrusts waste material out of holes in the stem. When detected by this means it can be cut out of the plant and killed. In large fields where hand control is impossible, deep plowing when the vines are dead kills many of the insects. Other methods depend on an accurate knowledge of the time of deposition of the eggs. Spraying is effective during that period.

SQUASH BUG (*Insecta, Hemiptera*). A true bug, *Anasa tristis*, about $\frac{5}{8}$ inch (16 millimeters) long, gray-brown above and mottled with yellow below, showing red in flight from areas beneath the wings. It sucks the juices of pumpkins, squash, and related vines and spreads a bacterial wilt.

The adults hibernate beneath debris on the ground and may be trapped under pieces of board and destroyed.

SQUID. See **Mollusks.**

SQUIRRELFISHES (*Osteichthyes*). Of the order *Berycomorphi*, family *Holocentridae*, squirrelfishes, also known as soldierfishes, are usually of some shade of bright red and prefer tropical reefs. There are about 70 species, well distributed throughout world tropical waters. The largest (*Holocentrus spinifer*) attains a length of about 2 feet (0.6 meter) and is found in the eastern Pacific. *Ostichthys japonicus* (deep-water squirrelfish) differs from other fishes of the family by preferring deeper water. The fish is marketed commercially in Hawaii under the name *menpachi* and is considered a premium food fish. It has been reported that squirrel-fishes can produce sounds.

SQUIRRELS AND OTHER SCIUROMORPHS (*Mammalia, Rodentia*). This large family of rodents also includes the chipmunks, gophers, spermophiles, prairie dogs, woodchucks, marmots, and whistlers, among others.

The squirrel is an arboreal or terrestrial rodent with a long bushy tail. Representatives of the true squirrels are found on all continents but Australia. North America has ten species of true squirrels, some widely distributed and extremely variable, and two species of flying squirrels. The little red squirrel, *Sciurus hudsonicus*, is the most widely known, with the gray, *S. carolinensis*, and fox, *S. niger*, also well known and widely distributed. The latter two animals are valued as small game, many people considering the flesh to be excellent. These squirrels may gather as much as 20 pounds (9 kilograms) of nuts per season for storage to be used during the winter months. Some species also accumulate thick layers of fat in preparation for winter.

Flying squirrels belong to the genus *Sciuropterus* and are distinguished from true squirrels and related forms by the presence of folds of skin stretching from the front to the hind legs along the sides of the body. These folds, held extended by the legs, support the animal in the air during long gliding leaps. See Fig. 1. The African flying squirrels belong to a different family, the *Anomaluridae*. The sciuroptera are found from Scandinavia eastward to Japan and from northern Canada southward to Honduras. Large numbers are found in India and the Malayan regions. Size ranges widely, but some species approach 3 feet (.9 meter) in length. General coloration is a fox-like red, or red spotted with white or black. In the larger species, the fold of the gliding membrane connects the hind limbs with the tail. The smaller species have the membrane along the side from front to back feet. All species are omnivorous and have large, dark eyes. It has been reported that glides of up to 200 feet (60 meters) from high to low tree branches occur.

Chipmunks are small burrowing rodents of squirrel-like appearance, but with the tail shorter and not bushy. They are brown or grayish with longitudinal stripes on the back or sides. They are omnivorous and, while not usually troublesome, sometimes do destroy flowering bulbs during the winter. Chipmunks belong to several genera and numerous species and subspecies. Some of the western species are called golden chipmunks or rock squirrels and others antelope chipmunks or ground

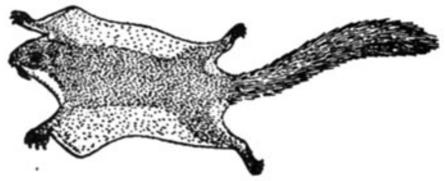


Fig. 1. Flying squirrel.

squirrels. They are related to the ground squirrels and gophers. While predominantly North American, some species are found in Siberia.

Pocket gophers are stout-bodied burrowing animals of several species, found throughout the United States. They have furlined cheek pouches opening at the sides of the mouth. So-called ground squirrels are slender burrowing species of the gopher and are found in the central and western states. Gophers are injurious to crops. The pocket gophers eat roots, and the ground squirrels are more injurious to grain. The term *gopher* is from the French, meaning "to tunnel." Gophers live in tunnels. The entrance to a tunnel is hidden, but the presence of a large mound of dirt nearby indicates the proximity of a tunnel. The tunnels have numerous entries and ports of escape and can be long and complex. Gophers are particularly destructive in dry seasons, when they gnaw roots for their moisture content. In size, the gopher is comparable to a large rat, but with larger proportioned claws. Because of the grinding sound made when cutting roots, the animal sometimes can be heard while working.

The spermophile is a small slender animal with short legs, small external ears, a short hairy tail, and large cheek pouches. They are also called ground squirrels or gophers. Most of the several species of spermophiles are confined to limited areas in the arid lands of the western United States. The striped species, commonly called gopher, ranges from central Ohio to the Rocky Mountains and from Canada to Texas. This species is brown with alternating clay-yellow stripes and rows of spots on the back and sides. The animal is destructive to crops and known for damaging lawns.

The woodchuck is a large heavy-bodied animal of the Northern Hemisphere. See Fig. 2. Woodchucks have short stout legs and are powerful burrowing animals, penetrating many feet into the ground. They also climb readily, although somewhat clumsily. They eat vegetation of many kinds and sometimes become troublesome in fields and gardens. North America has three species of woodchucks, the common woodchuck, *Marmota monax*, the yellow-bellied woodchuck, *M. flaviventris*, and the whistler, *M. caligata*. The first occurs from Kansas to Georgia, northward to Alaska and Hudson Bay. The second ranges from the Rocky Mountains to the Pacific, and the third is also western, ranging from Montana and Washington to Alaska. The names *marmot* and *groundhog* are also generally applied to them. The Old World species are widely distributed in Europe and Asia, where they are more commonly called marmots. Some of these include the bobac, *M. bobac*, and the alpine marmot, *M. marmota*. The fur is sparse and rather coarse, but it is used to a limited extent. The flesh of woodchucks is eaten, but it is inferior to that of other common rodents. While its flavor has been described as good, the flesh is coarse in texture, as compared with that of squirrels and rabbits.

Fig. 2. Eastern woodchuck. (*W. Goodpaster.*)

The prairie dog is a burrowing animal found chiefly west of the Mississippi River but has been introduced into a few eastern states. The animal is small and stout-bodied, with shallow cheek pouches. The prairie dog is a plant feeder and in settled regions can damage crops severely. The animal is also called the prairie marmot. The prairie dog ranges from 10 to 12 inches (25 to 30 centimeters) in length, plus a tail of 2 to 4 inches (5 to 10 centimeters) in length. Because of their damaging ways, the population is small due to extensive extermination.

STABILITY. In general, the tendency to remain in a given state or condition, without spontaneous change; and thus that attribute of a system which enables it to develop restoring forces between its elements, equal to or greater than the disturbing forces, so as to restore a state of equilibrium between the elements. Thus, a body of air is in a stable state if, when displaced somewhat from its original position, it tends to return thereto. A chemical compound is said to be stable if it is not readily decomposed.

In meteorology, *static stability* (also called *hydrostatic stability, vertical stability, convectional stability*) is the stability of an atmosphere in hydrostatic equilibrium with respect to vertical displacements, usually considered by the parcel method. The criterion for stability is that the displaced parcel be subjected to a buoyant force opposite to its displacement, e.g., that a parcel displaced upward be colder than its new environment. This is the case if $\gamma < \Gamma$, where γ is the environmental lapse rate and Γ the process lapse rate, dry-adiabatic for unsaturated air and saturation-adiabatic for saturated air.

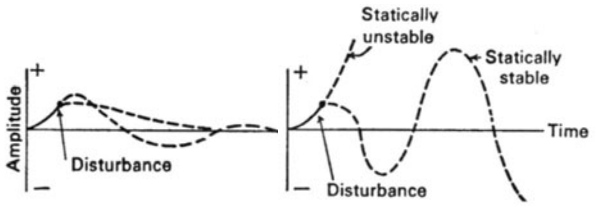
Neutral stability (also called *indifferent stability, indifferent equilibrium*) is the state of an unsaturated or saturated column of air in the atmosphere when its environmental lapse rate is equal to the dry-adiabatic lapse rate or the saturation-adiabatic lapse rate, respectively. Under such conditions, a parcel of air displaced vertically will experience no buoyant acceleration.

Conditional stability is a condition of a negative feedback system which causes it to be stable (nonoscillating) for certain values of gain, and unstable for other values.

STABILITY (Mechanical). Mechanical stability is that property of a body which causes it to develop forces opposing any position or motion disturbing influence. The subject may be divided into *static stability* and *dynamic stability*. The former is concerned with the production of the restoring forces, the latter with the oscillations that are set up in the system as a result of the restoring forces.

Another classification is into (1) *positive stability* when the displaced object returns to an initial state of equilibrium after a temporary disturbance, (2) *neutral stability* when the object tends to remain in a definite position but when disturbed may come to rest in a new position, (3) *negative stability* (i.e., instability), when the object assumes an entirely new position when disturbed from its initial state. A simple damped pendulum illustrates the first; a sphere on a horizontal plane, the second; while a slender cylinder standing vertically on end is a case of negative stability.

Let it be assumed that an object at rest or in a state of uniform motion receives a disturbing force. Depending on the kind of stability possessed, it might react with one of the motions shown in the accompanying figure. If it is dynamically stable as well as statically stable, its



(*Left*) Positive stability (both statically and dynamically stable); (*Right*) negative stability (both dynamically unstable).

motion-time history may be one of diminishing oscillation or of simple subsidence, depending on the magnitude of damping, and inertial effects. Dynamic instability may occur with either static stability or static instability. These lead to divergent oscillation, or to complete divergence.

STABILITY (System). For control and feedback system analysis, a system is stable only if all roots of its characteristics equation lie to the left of the imaginary axis of the s-plane. Thus, if the roots of the characteristic equation can be determined, the question of stability is answered. If, however, the characteristic equation is of high order, the determination of roots may involve a great deal of work. In cases such as this the existence of roots in the right half of the s-plane may be determined using either the Routh or Hurwitz criterion. The methods are as follows:

Routh Criterion. If, as is the case with all lumped parameter systems, the characteristic equation can be written in polynomial form

$$a_0 + a_1 s + a_2 s^2 + \cdots + a_n s^n = 0$$

the array

$$\begin{matrix} a_0 & a_2 & a_4 & \cdots \\ a_1 & a_3 & a_5 & \cdots \\ A_1 & A_2 & A_3 & \cdots \\ B_1 & B_2 & \cdots & \cdots \\ C_1 & \cdots & \cdots & \cdots \end{matrix}$$

is set up, in which $A_1 = (a_1 a_2 - a_0 a_3)/a_1$, $A_2 = (a_1 a_4 - a_0 a_5)/a_1$, $A_3 = (a_1 a_6 - a_0 a_7)/a_1$, etc., and succeeding rows are obtained from the preceding two rows in the same manner as the third is obtained from the first two. The rows will be found to get shorter by one element every two rows and to be $(n + 1)$ in number.

The system is stable if and only if all the elements of the first column have the same sign. (In fact, the number of changes of sign on going down the first column is the number of roots of the characteristic equation in the right half-plane.)

If a zero is formed in the first column before the array is complete (which prohibits the calculation of further rows), it may indicate either instability or critical stability. If the zero is preceded by two rows consisting of the same number of elements and such that the ratio of corresponding elements in the two rows is constant, roots on the imaginary axis are indicated; moreover, in this case, the elements of either of these rows, read from right to left, successively multiplied by s^0, s, s^2, etc., added and equated to zero will give the equation whose roots are the roots of the characteristic equation lying (in conjugate pairs) on the imaginary axis. If, on the other hand, the zero is not preceded by two rows with proportional elements, then there is at least one root in the right half-plane and the system is unstable.

A necessary but, except when $n = 1$ or 2, an insufficient condition for stability is that all the polynomial coefficients shall be nonzero and all have the same sign. This condition is always implicitly contained in the Routh criterion.

Hurwitz Criterion. Using the same polynomial form of the characteristic equation as before, the array

$$\begin{matrix} a_1 & a_0 & 0 & 0 & 0 & \cdots & \cdots \\ a_3 & a_2 & a_1 & a_0 & 0 & 0 & \cdots \\ a_5 & a_4 & a_3 & a_2 & a_1 & a_0 & \cdots \\ a_7 & a_6 & \cdots & \cdots & \cdots & \cdots & \cdots \end{matrix}$$

is drawn up, extending for $(n - 1)$ rows and columns, absent coefficients being replaced by zeros.

The system is stable if and only if:

1. All the coefficients of the polynomial are of the same sign, which must be assumed positive; and
2. The values of the determinants of order 2 to $(n - 1)$ and having the top left array element as their top left element are also positive.

The application of either the Routh or Hurwitz criteria is a valuable tool in the analysis of system stability. However, there are limitations that must be recognized. To use either of these techniques the system equation must be known. In many cases, and this is especially true in process control applications, the system equation is not available. In applying these techniques little information can be obtained concerning the relative stability of the system. Furthermore, it is difficult to determine the effect on stability of individual parameters for design considerations. Due to these limitations, the stability of a system is normally determined from frequency response curves using the Nyquist stability criterion. The frequency response curves may be obtained from the system equations or determined experimentally. A general description of the Nyquist method is as follows.

Nyquist Criterion; 1 (for a control system). It is assumed that if $G(s)$ is the output to error transfer function, then as $s \to \infty$, $G(s) \to 0$. This is normally the case in any practical system.

If the system is stable on closed loop, the vector locus of $G(i\omega)$ drawn from $\omega \to -\infty$ to $\omega \to +\infty$ in the sense of increasing ω (i.e., the vector locus of $G(s)$ corresponding to following the imaginary axis in the s-plane) encircles the point $(-1, 0)$ P times counterclockwise, where P is the number of poles of $G(s)$ to the right of the imaginary axis, multiple poles counting according to their order.

The following points should be noted:

1. If $P \neq 0$ the system is unstable on open loop, i.e., with the (-1) feedback from output to error broken.
2. The value of P may be found, in lumped parameter systems, by writing $G(s) = N(s)/D(s)$, where N and D are polynomials; P is then the number of zeros of $D(s)$ to the right of the imaginary axis, which may be found most simply by the Routh criterion.
3. If $G(s)$ has poles *on* the imaginary axis, these must be circumvented in the s-plane by infinitesimal semicircles in the right half-plane having these poles as centers. These indentations will correspond in the plane of $G(s)$ to a clockwise rotation through 180° at indefinitely large radius for every simple pole so encountered, multiple poles again counting according to their order. These infinite semicircles must be taken into account in assessing the number of encirclements of the point $(-1, 0)$.
4. Since $G(-i\omega)$ is the complex conjugate of $G(i\omega)$, the locus of $G(s)$ for negative real frequencies is the mirror image, in the real axis of the $G(s)$ plane, of the locus for corresponding positive real frequencies.
5. If and only if $P = 0$, an adequate simpler criterion is that the locus of $G(i\omega)$ drawn in the sense of increasing ω, shall leave the point $(-1, 0)$ to its left. See **Feedback.** This same concept may be applied in using the Bode plot. In this context the system gain must be less than 1.0 when the phase angle is equal to minus 180°. It is easily seen that this corresponds to the point $(-1, 0)$ on the Nyquist plot.

Nyquist Criterion; 2 (for a general feedback system). It is, of course, always possible to reduce the basic equations of the system

$$a_{1j}(s)Q_1(s) + a_{2j}(s)Q_2(s) + \cdots + a_{nj}(s)Q_n(s) = f_j(s), j = 1, 2, \cdots n$$

by eliminating all but two of the Q's, to a pair of equations relating these two quantities only, which can be put in the form

$$Q_1(s) - \beta s) Q_2(s) = F_1(s)$$
$$Q_2(s) - \mu(s) Q_1(s) = F_2(s)$$

This, in effect, reduces the system to a single-loop system of loop transfer function $\mu(s)\beta s)$. Since the characteristic equation is

$$\Delta(s) = 1 - \mu(s)\beta s) = 0$$

the above formulation of the Nyquist criterion may be applied, replacing $G(s)$ in the preceding section by $-\mu(s)\beta s)$. Thus the system will be stable provided that:

1. The locus of $-\mu(i\omega)\beta i\omega)$ encircles the point $(-1, 0)$ P times counterclockwise, or
2. The locus of $+\mu(i\omega)\beta i\omega)$ encircles the point $(+1, 0)$ P times counterclockwise, or
3. The locus of $1 - \mu(i\omega)\beta i\omega)$ encircles the origin P times counterclockwise, P being in each case the number of simple poles of

$\mu(s)\beta s)$ lying in the right-hand half-plane, any poles on the imaginary axis being circumvented as explained above.

In a multiloop system, however, μ or β or both may comprise subsidiary, possibly unstable loops, so that in general $P \neq 0$. To avoid the reduction of the equations to the above form and the subsequent finding of the value of P by the Routh criterion, the Nyquist method may be extended as follows.

We start from the premise that a passive system is necessarily stable. In other words, if, in the given system, the active elements are made inactive, we have a stable system. We then imagine each of the active elements to be activized one at a time until the system is back to its normal state, investigating at each stage the behavior of the return difference for the reactivized element. (The order in which the elements are reactivized is entirely a matter of convenience.)

Rendering an element inactive has the effect, in the set of general system equations, of making some parameter of that element zero. In the case of an electron tube, the amplification factor or the mutual conductance vanish; in the case of a rotary machine amplifier, the armature voltage per ampere-turn vanishes, and analogously for other types of active elements. Further, it can be shown that the return difference for any parameter is the ratio of the values assumed by the system determinant when the parameter has its normal value and when this value is made zero, provided only that the determinant is a linear function of the parameter.

We suppose then that there are m active elements and, having placed these in some arbitrary order, we denote by Δ_r the value of the system determinant when the first r of these elements are activized, the remainder being inactive. Starting from the completely inactive system, the return difference for the first element is Δ_1/Δ_0. The return difference for the second element (the first one remaining active) is Δ_2/Δ_1. The return difference for the third (the first two remaining active) is Δ_3/Δ_2 and so on until finally the return difference for the mth element (all other remaining active) is Δ_m/Δ_{m-1} in which, furthermore, $\Delta_m \equiv \Delta$, the normal system determinant. Now the number of counterclockwise encirclement of the origin made by the Nyquist locus of the return difference of the rth element Δ_r/Δ_{r-1} is $(z_{r-1} - z_r)$, where z_r denotes the number of zeros of Δ_r in the right half-plane. Hence the sum of such encirclements made by all the successive return difference loci is $(z_0 - z_1) + (z_1 - z_2) + \cdots + (z_{m-1} - z_m) = z_0 - z_m$. But z_0 is zero since the system is then passive. For stability, moreover, z_m must be zero. Hence, *for stability of the normal system, the sum of the encirclements of the origin made by the various return difference Nyquist loci must be zero.*

Additional Reading

Higgins, S. P., Jr., and J. M. Nelson: "Process Control Techniques," in *Process Instruments and Controls Handbook*, 3rd Ed. (D. M. and G. D. Considine, Eds.), McGraw-Hill, New York, 1985.

Matley, J., et al.: "Practical Process Instrumentation and Control," Vol. II, McGraw-Hill, New York, 1986.

Osborne, R. L.: "Fundamentals of Automatic Process Control," in *Process Instruments and Controls Handbook*, 3rd Ed. (D. M. and G. D. Considine, Eds.), McGraw-Hill, New York, 1985.

Shinskey, F. G.: "Process Control Systems," 2nd Ed., McGraw-Hill, New York, 1979.

STABILIZATION (Ship). In addition to the obvious method of suitable design of the hull, a number of specific devices have been developed for damping the motions of a ship among the waves, especially the rolling motion. Ships have been built carrying large gyroscopes, but in general this method proved too costly in money and added weight for the effect obtained. A more widely used method is that of *anti-rolling tanks*, also called *water chambers*, which are usually placed in stabilizing bulges on each side of the hull. By controlling the flow of water from the tanks on one side to those on the other, so that the period of the contained water is proportionate (ideally about 70%) to the ship's period of roll, the maximum stabilization can be obtained. (Its amount is much less than that of a gyroscope, but far less costly.) Control of the water flow may be effected by a valve-controlled air line connecting the upper parts of the U-tubes connecting the tanks. This control can also be used to shut off the system in wave systems of very irregular pattern, which cause erratic behavior of the ship when the water is shifted among the anti-rolling tanks.

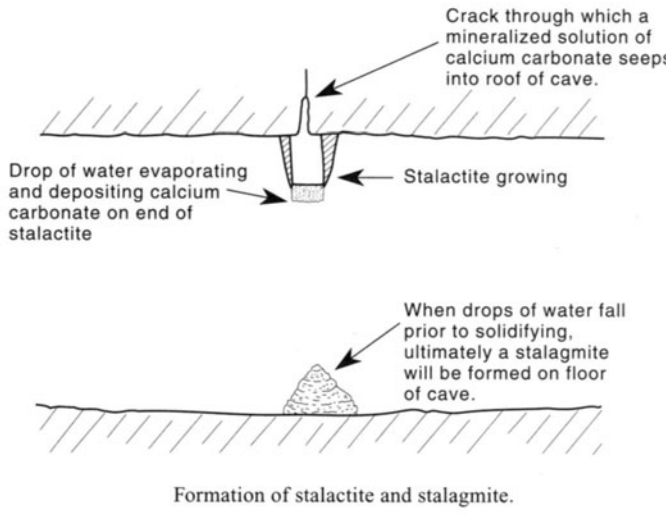
STABLE FLY (*Insecta, Diptera*). Similar to the house fly in appearance, the small stable fly, *Stomoxys calcitrans* (Linne), is a severe pest against horses and mules, usually biting the legs of the animals. It also attacks cattle, hogs, goats, sheep, rabbits, dogs, rats, and humans. The insect is widely distributed throughout the United States. In the northern states, the insect winters over as larvae and pupae, usually in wet areas near straw and manure. In the southern states, the fly can be found in all stages all year. Where stable flies are abundant, animals can lose a substantial amount of blood during the course of a day because each stable fly will require from 1 to 2 drops of blood each time it bites. The bites are painful and irritating and generally contribute to deterioration of the health of the animals attacked. Dairy cows yield less milk; cattle lose weight; horses can become unmanageable. The fly is also suspected of carrying a number of diseases, such as anthrax, leprosy, surra, and swamp fever (infectious anemia of horses), although no solid proof of this has been available.

Prevention through cleanliness of animal quarters cannot be overstressed as an effective measure. Highly valued animals can be protected with blankets and coverings. Fly traps, mechanical or electrical, are suggested for dairy barns and other areas frequented by the flies. Numerous commercial repellent formulations are available—for use in disinfecting animal quarters as well as for spraying on affected parts of the animals once or twice per week.

STACK. In geology, a rock pillar or monument that occurs relatively close to marine cliffs which have usually been developed in hard, horizontally bedded, but jointed, sedimentary rocks. In some cases, a stack is the remaining pillar of an arch. Both sea arches and stacks are well developed on the coast lines of the Gaspé Peninsula and the North East Highlands of Scotland.

STAGE. 1. A self-propelled separable element of a rocket vehicle. 2. A step or process through which a fluid passes, especially in compression or expansion. 3. A set of stator blades and a set of rotor blades in an axial-flow compressor or a turbine (see **Turbine (Steam)**); an impeller wheel in a radial-flow compressor.

STALACTITE AND STALAGMITE. A *stalactite* is a deposit of a mineralized solution, commonly calcium carbonate, which hangs like an icicle from the roof or wall of a limestone cavern. See accompanying illustration. The formation of the stalactite usually is quite a slow process. Corresponding columnar structures built upward from the floors of caves beneath the stalactites are called *stalagmites*. Stalactite is derived from the Greek, meaning to fall in drops. Stalagmite, also derived from the Greek, means that which drops. When a stalactite from the top of a cave and a stalagmite from the floor of the cave join, the resulting singular structure is called a *column*.


Crack through which a mineralized solution of calcium carbonate seeps into roof of cave.
Drop of water evaporating and depositing calcium carbonate on end of stalactite
Stalactite growing
When drops of water fall prior to solidifying, ultimately a stalagmite will be formed on floor of cave.

Formation of stalactite and stalagmite.

STALAGMOMETER. An apparatus for the measurement of surface tension by the drop-weight method.

STALING (Bakery and Food Products). Most authorities agree that staling of bakery products commences immediately after the product leaves the oven and that it is a reasonably progressive process, not involving a delayed start or of a cyclic nature. Post-processing conditions obviously can favor or be unfavorable to the process of staling. In bread, the moisture content of the crumb averages about 45% (weight) when the loaf leaves the oven. The crust, on the other hand, has a moisture content of 5–7%. Under these conditions (lack of equilibrium), the crust loses moisture on the outside to the atmosphere, while it picks up moisture from the crumb on the inside of the loaf. The result is a toughening of the crust, accompanied and aided by formation of a soggy undersurface. The crumb, of course, becomes drier and drier with time. As moisture is expelled to the atmosphere, along with other volatiles, flavor decreases. Flavor loss also is the result of essential deactivation of certain flavoring components when they are absorbed by the inner solids. It is also likely that unexplained chemical reactions occur (oxidation, etc.) that alter the chemical structure and hence aroma and flavor of the flavoring components. With some products, the staling processes are not necessarily fully irreversible. Some breads, such as Italian breads, can be moistened and placed in an oven and returned to a reasonable degree of organoleptic acceptability. Some bakers purposely underbake their product to slow up the staling process.

Many authorities believe that starch is in some way the key to most of the staling phenomena. A tie-in with starch is evident in a practical way by observing that bread put in the refrigerator tends to become stale sooner than when left at a reasonable room temperature; but if the bread is placed in a freezer, it may be retained for long periods, but with some evidence of staleness once the bread is thawed. It is known that starch retrogrades at a faster rate as the temperature is lowered, but beyond a certain point the starch degradation is greatly slowed at very low temperatures and hence the staling rate is slowed. Various additives have alleviated the staling process to some degree.

STALL. A term in common use to describe the condition of an aerodynamic burble upon a wing. An airplane that has "stalled" has had the streamline flow over the upper wing surface partly destroyed. The lift is partially lost so weight and lift are out of equilibrium. Hence, a stall is followed by a nosing down motion accompanied by rapid loss of altitude. The term "stall" here applies to an aerodynamic phenomenon. The term is also used when an aircraft engine ceases to operate, when it is said to stall.

See also **Aerodynamics;** and **Airplane.**

STANDARD CONDITIONS. Many physical and chemical phenomena and substances are defined in terms of *standard* conditions. In some instances, a temperature commonly prevailing in a chemical laboratory may be selected. Thus, when comparing a number of substances, such as their index of refraction, one may find lists in handbooks that give these values as measured by a given temperature and pressure. The Smithsonian Tables, for example, include such data for scores of substances. The researcher then can seek formulas for converting such values to other temperatures/pressures. Interpolations in some instances can be linear over a wide range of values, or the relationships may be nonlinear.

In the case of gases, properties may be tabulated in terms of their existence at 0°C and 760 mm pressure. To determine the volume of a gas at some different temperature and pressure, corrections derived from known relationships (Charles', Amonton's, Gay-Lussac's, and other laws) must be applied as appropriate. In the case of pH values given at some measured value (standard for comparison), the same situation applies. Commonly, lists of pH values are based upon measurements taken at 25°C. The pH of pure water at 22°C is 7.00; at 25°C, 6.998; and at 100°C, 6.13. Modern pH instruments compensate for temperature differences through application of the Nernst equation.

Standard conditions are not necessarily consistent with standards definitions. The careful researcher will always take note of the conditions stated for determining values in a tabulated list.

STANDARD DEVIATION. The standard deviation of a probability distribution is the square root of its variance. It is the most useful measure of dispersion.

STANDARD ERROR. A name often given to the standard deviation of a sampling distribution.

STANDARD FREE ENERGY INCREASE. Often referred to as standard free energy. The increase in Gibbs free energy (see **Free Energy**) when the reactants in a chemical change, all in their standard states (e.g., unit concentration, or at 1 atmosphere pressure) are converted into the products in their standard states. Given by

$$\Delta G^\circ = -RT \ln K$$

where R is the gas constant, T, the absolute temperature, K, the equilibrium constant. See also **Free Energy Change.**

STANDARD STATE. The stable form of a substance at unit activity. The stable state for each substance of a gaseous system is the ideal gas at 1 atmosphere pressure; for a solution it is taken at unit mole fraction; and for a solid or liquid element it is taken at 1 atmosphere pressure and ordinary temperature.

STANDARD UNIT (Statistics). A variate may be changed to standard units by the transformation

$$t = \frac{x - \bar{x}}{\sigma_x}$$

where t is in standard units, x is the variate, $\bar{x}$ is the mean, and σ_x is the standard deviation of the distribution. In such cases t has a mean of zero and a standard deviation of 1 and is said to be in standard measure.

STANDING-WAVE RATIO. Any transmission line such as a waveguide or an acoustic transmission system, unless terminated by its characteristic impedance, will exhibit a superposition of standing and progressive waves. The standing-wave ratio is a measure of the relative amplitudes of the two types of wave and is defined as the ratio of the maximum amplitude of pressure (or voltage) to the minimum amplitude of pressure (or voltage) measured along the path of the waves. Thus, at a given frequency in a uniform waveguide the standing-wave ratio is the ratio of the maximum to the minimum amplitudes of corresponding components of the field (or the voltage or current) along the waveguide in the direction of propagation. Alternatively, the standing-wave ratio may be expressed as the reciprocal of the ratio defined above.

STANNITE (Mineral). This mineral is a sulfo-stannate of copper and iron, sometimes with some zinc, corresponding to the formula, Cu_2FeSnS_4. It is tetragonal; brittle with uneven fracture; hardness, 4; specific gravity, 4.3–4.5; metallic luster; color, gray to black, sometimes tarnished by chalcopyrite; streak, black; opaque. The mineral occurs associated with cassiterite, chalcopyrite, tetrahedrite, and pyrite, probably the result of deposition by hot alkaline solution. Stannite occurs in Bohemia; Cornwall, England; Tasmania; Bolivia; and in the United States in South Dakota. It derives its name from the Latin word for "tin," *stannum.*

STAR. There is probably no one class of physical objects that has attracted more popular and scientific attention throughout the ages than have the stars. A small portion of the mass of mythological material that is associated with these bodies will be found in the various articles dealing with the individual constellations and a few of the brighter stars.

A star, as distinguished from a *planet*, is a self-gravitating object capable of generating its own energy by nuclear processes, or the remnant (white dwarf or neutron star) of such an object. The *protostellar phase* would be seen as the formative part of the life of a star. The contraction of the gas cloud leads to core temperatures high enough to ignite the central engine. A lower mass limit of 0.08 solar masses ap-

TABLE 1. PHYSICAL CHARACTERISTICS OF SOME TYPICAL STARS[a]

Star	Spectral Class	Temperature in °K	Density in Terms of Water	Referred to Sun as Unity: Luminosity	Mass	Diameter
			Giants			
Antares	M1 Ib	3,100	0.0000003	3,500	30	480
Aldeberan	K5 III	3,300	0.00002	90	4	60
Arcturus	$K2_p$ III	4,100	0.0003	100	8	30
Capella	G8 III?+F	5,500	0.002	150	4.2	12
β Centauri	B III	21,000	0.02	3,100	25	11
			Main Sequence			
Vega	A0 V	11,200	0.1	50	3	2.4
Sirius A	A1 V	11,200	0.4	26	2.4	1.8
Altair	A7 V	8,600	0.6	9.2	2	1.4
Procyon	F5 V	6,500	1.2	5.4	1.75	1.7
α Centauri A	G2 V	6,000	1.1	1.12	1.1	1.2
The Sun	G2 V	6,000	1.4	1	1	1.0
70 Ophiuchi A	K0 V	5,100	0.9	0.42	0.9	1.0
61 Cygni A	K7 V	3,800	1.3	0.21	0.6	0.7
Krueger 60A	M3 V	3,300	9	0.002	0.3	0.3
			White Dwarfs			
Sirius B	F	7,500	27,000	0.10	0.96	0.034
O_2 Eridani B	A0	11,000	64,000	0.003	0.44	0.019

[a]For a general discussion of Population I and Population II stars, see **Giant and Dwarf Stars.**

TABLE 2. THE TEN BRIGHTEST STARS
(*Courtesy* van de Kamp *Publications of the Astronomical Society of the Pacific*, 1953, vol. 65)

Number	Name	Right Ascension 1900	Declination 1900	Visual Magnitude and Spectrum: A	B	C	Annual Proper Motion	Parallax		Distance in Light Years	Visual Absolute Magnitude: A	B	C	Visual Luminosity ⊙: A	B	C
1	Sirius	$6^h 40^m.7$	−16° 35′	−1.6 A0	7.1 A5	—	1″.32	0″.375	±″.004	8.7	+1.3	+10.0	—	23	0.008	—
2	Canopus	6 21 .7	−52 38	−0.9 F0	—	—	0 .02	.018	6	180:	−4.6:	—	—	5,200:	—	—
3	α Centauri	14 32 .8	−60 25	+0.3 G0	1.7 K5	11 M	3 .68	.760	5	4.29	+4.7	+ 6.1	+15.4	1.0	.28	.000052
4	Vega	18 33 .6	+38 41	0.1 A0	—	—	0 .35	.123	5	26.5	+ .5	—	—	48	—	—
5	Capella	5 9 .3	+45 54	0.2 G0	10.0 M1	13.7 M5	0 .44	.073	4	45	− .5	+ 9.3	+13.0	120	.014	.0005
6	Arcturus	14 11 .1	+19 42	0.2 K0	—	—	2 .29	.090	5	36	0.0	—	—	76	—	—
7	Rigel	5 9 .7	− 8 19	0.3 B8p	—	—	0 .01	.005:		650:	−6.2:	—	—	23,000:	—	—
8	Procyon	7 34 .1	+ 5 29	0.5 F5	10.8	—	1 .25	.288	4	11.3	+2.8	+13.1	—	5.8	.00044	—
9	Archernar	1 34 .0	−57 45	0.6 B5	—	—	0 .09	.023	42	140:	−2.6:	—	—	800:	—	—
10	β Centauri	13 56 .8	−59 53	0.9 B1	—	—	0 .04	.016	11	200:	−3.1:	—	—	1,300:	—	—

TABLE 3. STARS NEARER THAN FIVE PARSECS
(*Courtesy* van de Kamp *Publications of the Astronomical Society of the Pacific*, 1953, vol. 65)

Number	Name	Right Ascension 1950	Declination 1950	Parallax	Distance in Light Years	Cross Proper Motion	Motion KM/sec	Position Angle	Radial Velocity KM/ sec	Visual Magnitude and Spectrum: A	B	C	Visual Absolute Magnitude: A	B	C
1	Sun	—	—	—	—	—	—	—	0	−26.9 G0	—	—	4.7	—	—
2	α Centauri	$14^h 36^m.2$	−60°38	0″.745	4.3	3″.68	23	281°	−25	0.3 G2	1.7 K5	11 M5e	4.7	6.1	15.4
3	Barnard's star	17 55 .4	+4 33	0 .552	6.0	10 .30	90	356	−108	9.5 M5	*	—	13.2	*	—
4	Wolf 359	10 54 .2	+7 20	0 .429	7.7	4 .71	54	235	+13	13.5 M6e	—	—	16.6	—	—
5	Luyten 726–8	1 36 .4	−18 13	0 .367	7.9	3 .35	38	80	+29	12.5 M6e	13.0 M6e	—	15.6	16.1	—
6	Lalande 21185	11 0 .6	+36 18	0 .398	8.2	4 .78	57	187	−86	7.5 M2	*	—	10.5	*	—
7	Sirius	6 42 .9	−16 39	0 .375	8.7	2 .32	16	204	−8	−1.6 A1	7.1 wd	—	1.3	10.0	—
8	Ross 154	18 46 .7	−23 53	0 .345	9.3	0 .72	9	106	−4	10.6 M5e	—	—	13.3	—	—
9	Ross 248	23 39 .4	+43 55	0 .316	10.3	2 .58	23	176	−81	12.2 M6e	—	—	14.7	—	—
10	ϵ Eridani	3 30 .6	−9 38	0 .303	10.8	0 .97	15	271	+15	3.8 K2	—	—	6.2	—	—

TABLE 3. *(continued)*

Number	Name	Right Ascension 1950	Declination 1950	Parallax	Distance in Light Years	Cross Proper Motion	Motion KM/sec	Position Angle	Radial Velocity KM/sec	Visual Magnitude and Spectrum A	B	C	Visual Absolute Magnitude A	B	C
11	Ross 128	11 45 .1	+1 7	0 .301	10.9	1 .37	22	151	−13	11.1 M5	—	—	13.5	—	—
12	61 Cygni	21 4 .7	+38 30	0 .293	11.1	5 .22	84	52	−64	5.6 K6	6.3 M0	*	7.9	8.6	*
13	Luyten 789–6	22 35 .7	−15 37	0 .305	11.2	3 .27	53	46	−60	12.2 M6	—	—	14.5	—	—
14	Procyon	7 36 .7	+5 21	0 .288	11.3	1 .25	20	214	−3	0.5 F5	10.8 wd	—	2.8	13.1	—
15	ϵ Indi	21 59 .6	−57 0	0 .285	11.4	4 .67	77	123	−40	4.7 K5	—	—	7.0	—	—
16	Σ 2398	18 42 .2	+59 33	0 .280	11.6	2 .29	38	324	+1	8.9 M4	9.7 M4	—	11.1	11.9	—
17	Groombridge 34	0 15 .5	+43 44	0 .280	11.7	2 .91	49	82	+14	8.1 M2e	10.9 M4e	—	10.3	13.1	—
18	τ Ceti	1 41 .7	−16 12	0 .275	11.8	1 .92	33	297	−16	3.6 G4	—	—	5.8	—	—
19	Lacaille 9352	23 2 .6	−36 9	0 .279	11.9	6 .90	118	79	+10	7.2 M2	—	—	9.4	—	—
20	BD + 50° 1668	7 24 .7	+5 29	0 .268	12.4	3 .73	67	171	+26	10.1 M4	—	—	12.2	—	—

*The stars nearest the sun are often referred to as nearby stars. They are generally main sequence dwarf stars.

pears to be required in order to initiate the simplest nuclear reactions, while the upper stable mass may be as great as several hundred solar masses.

The methods for determining the physical characteristics of the stars will be found under such titles as: stellar parallax, stellar magnitude, spectral class, binary stars, variable stars, cepheids, giant and dwarf stars, etc. The characteristics of a typical main-sequence G-type star will be found in the articles dealing with the sun and various solar characteristics. The source of the energy radiated by the stars is discussed in the article on the sun, and under carbon cycle and proton-proton chain. See Tables 1, 2, and 3.

Numerous classes and specific stars are described in various entries throughout this encyclopedia. Consult alphabetical index.

Steven N. Shore, former Director, Astrophysics Research Center and Assoc. Prof. of Physics, New Mexico Institute of Mining and Technology, Socorro, New Mexico.

STAR CATALOGUES. Any listing of stars, usually arranged in order of increasing right ascension. Originally, star catalogues were intended merely for the purpose of providing accurate positions of the stars for use by navigators, but many modern catalogues are designed to provide particular characteristics of the stars.

The oldest existing star catalogue is contained in the Almagest of Ptolemy issued about 137 A.D. The Almagest catalogue was the standard one used throughout the Middle East and Latin West to the end of the 15th Century. Tycho Brahe's catalogue of 1580 marks the dawn of the modern era of star catalogues, and since that time many others have appeared. Probably the most comprehensive catalogue issued is the Bonner Durchmusterung, which first appeared about 1850, together with the various extensions which have since been published. During the last half of the nineteenth century the Astronomische Gesellschaft sponsored a catalogue of accurate positions of the majority of the stars contained in the Bonner Durchmusterung.

With application of photography to astronomy, several projects have been launched for obtaining comprehensive star catalogues. By far the most ambitious of all of the photographic catalogues is the so-called Astrographic Catalogue, which was started as a cooperative effort of 18 observatories in 1887. It is not entirely completed. However, there has been completed a photographic survey of the sky known as the Sky Atlas, the most recent such atlas being the Palomar Sky Atlas, and European Southern Observatory atlas.

As part of the Space Telescope project, a general catalogue of potential guide stars is being prepared. This will also serve as a permanent archival data set for future magnitude and positional studies. The European Space Agency is building an astrometric satellite (*HIPPARCHOS*) designed to observe positions and motions of several hundred thousand stars over the next decade.

Not all catalogues are restricted to positional and magnitude data. The largest spectroscopic catalogue is the Henry Draper project, first completed in the 1920s and currently being revised. The document contains information on the spectral characteristics of over 200,000 stars.

See also **Bonner Durchmusterung.**

Steven N. Shore, former Director, Astrophysics Research Center and Assoc. Professor of Physics, New Mexico Institute of Mining and Technology, Socorro, New Mexico.

STARCH. Chemically, starch is a homopolymer of α-D-glucopyranoside of two distinct types. The linear polysaccharide, amylose, has a degree of polymerization on the order of several hundred glucose residues connected by alpha-D-(1 → 4)-glucosidic linkages. The branched polymer, amylopectin, has a DP (degree of polymerization) on the order of several hundred thousand glucose residues. The segments between the branched points average about 25 glucose residues linked by alpha-D-(1 → 4)-glucosidic bonds, while the branched points are linked by alpha-D-(1 → 6)-bonds. See Fig. 1.

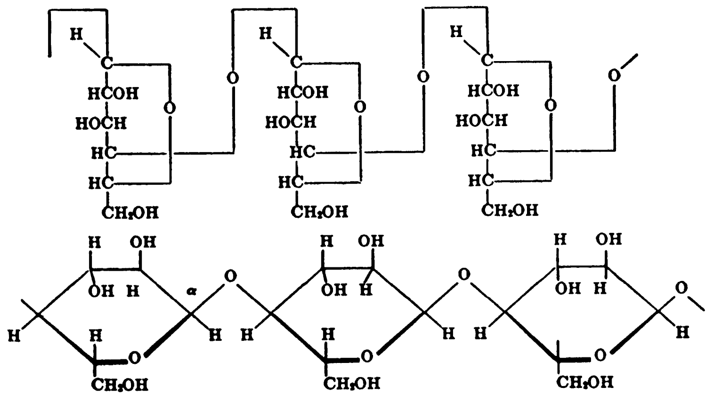


Fig. 1. A segment of the starch molecule.

Most cereal starches are made up of about 75% amylopectin and 25% amylose molecules. However, root starches are slightly higher in amylopectin, while waxy corn* and waxy milo starch contain almost 100% amylopectin. At the other extreme, high amylose corn starch and wrinkled pea starches contain 60–80% amylose. The molecules of amylose and amylopectin are synthesized by enzymes inside the living cell in plastids known as amyyloplasts and are deposited as starch granules. These granules are microscopic in size, ranging from 3–8 micrometers in diameter for rice starch up to 100 micrometers for the larger potato starch granules. Corn starch usually falls in a range of 5–25 microme-

* With exception of North America, where the plant is called *corn*, other English-speaking people call it *maize*. French = *mais*; Spanish = *maiz*.

ters. An experienced observer usually can identify the genetic origin of a sample of starch by the size and shape of the granules. The granules are insoluble in cold water, but swell rapidly when heated to the gelatinization temperature range for the particular starch involved. As the granules swell, they lose their characteristic cross under polarized light and imbibe water rapidly until they are many times their original size. Upon continued heating or mechanical shear, the swollen granules begin to disintegrate and the viscosity, having reached a maximum, begins to decrease. However, there usually are some granules and some segments of granules that do not completely disperse in aqueous systems even under the most stringent conditions.

As the partially dissolved paste is cooled, the hydrated molecules and segments of granules begin to precipitate. In a dilute system (approximately 1%), the segments and molecules retrograde or precipitate. At higher concentrations, sufficient intermolecular and intersegment bonds form to fix the entire system into three-dimensional gel. The rigidity of this gel is affected by many factors, but the amylose content is perhaps the most significant. High amylose starches, when thoroughly cooked, form very rigid gels. Waxy corn or waxy milo starch paste form little, if any, gel structure when cooled.

While some wheat and potatoes are processed in the United States, over 90% of all starch is produced from corn in what is called the *corn wet milling industry*. Close to one-quarter of a billion bushels of corn, representing about 5% of the total corn crop, is converted into wet-process products. The corn refining process is illustrated in Fig. 2. Shelled corn is delivered to the wet-milling plant in boxcars containing an average of 2,000 bushels (50.8 metric tons) per car, and unloaded into a grated pit. The corn is elevated to temporary storage bins, and then to scale hoppers for weighing and sampling. The corn passes through mechanical cleaners designed to separate unwanted substances, such as pieces of cobs, sticks, and husks, as well as metal and stones. The cleaners agitate the kernels over a series of perforated metal sheets; the smaller foreign materials drop through the perforations, while a blast of air blows away chaff and dust, and electromagnets draw out nails and bits of metal. Coming out of the storage bins, the corn is given a second cleaning before going into very large "steep" tanks.

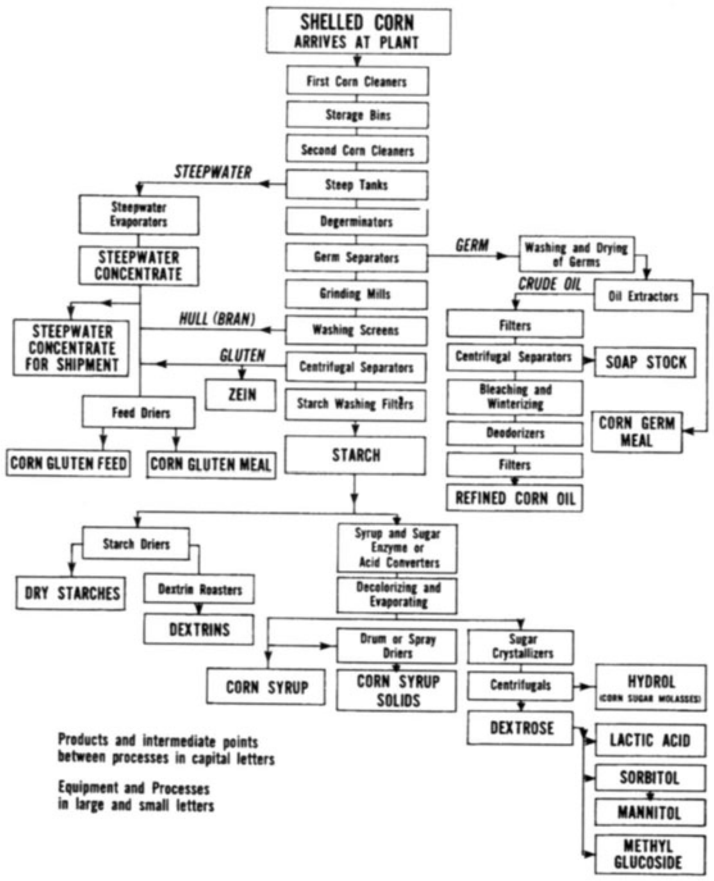


Fig. 2. The corn (maize) refining process. (*Corn Refiners Association, Inc.*)

At this point, the use of water becomes an essential part of the corn refining process. The cleaned corn is typically moved into large wooden or metal tanks holding 2,000 to 6,000 bushels (50.8 to 152.4 metric tons), and soaked for 36 to 48 hours in circulating warm water 49°C (120°F) containing a small amount of sulfur dioxide to control fermentation and to facilitate softening. At the end of the steeping process, the steepwater contains much of the soluble protein, carbohydrates, and minerals of the corn kernel, and is drawn off as the first by-product of the process. Steepwater, unmodified and modified, is an essential nutrient for production of antibiotic drugs, vitamins, amino acids, and fermentation chemicals. See Fig. 3. It is also an effective growth supplement for animal feeds.

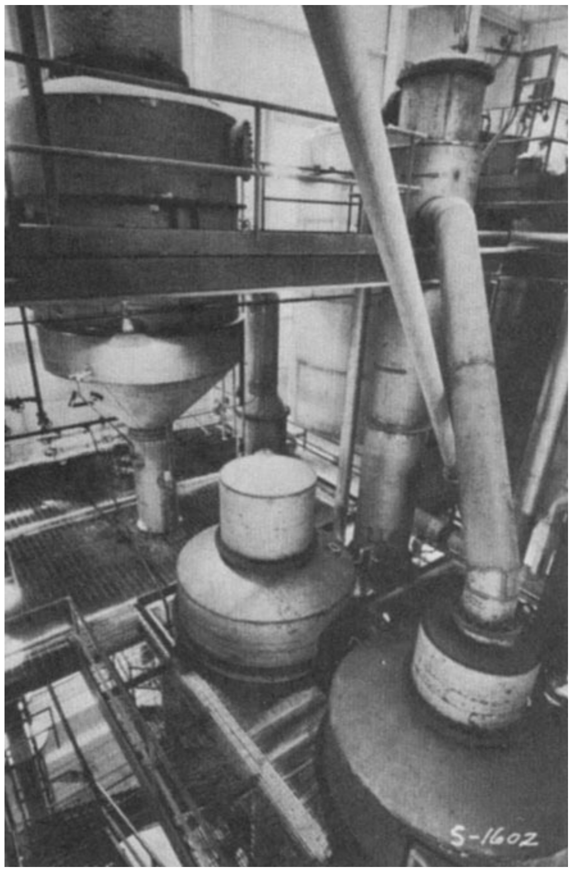


Fig. 3. Triple-effect steepwater evaporator. The third effect (forced circulation) is shown in background; second effect (falling-film, recirculating) is middle unit. The first effect vapor head is shown in foreground. (*Swenson, Whiting Corp.*)

From the steeps, the softened kernels go through degerminating mills, which are designed not for fine grinding, but rather for tearing the soft kernels apart into coarse particles, freeing the rubbery oil-bearing germ without crushing it, and loosening the bran. The wet, macerated kernels then are sluiced into flotation tanks, called germ separators, or centrifugal hydrocyclones. The germs, lighter than the other components of the kernel, float to the surface, and are skimmed off. By oil expellers or extractors (heat and pressure) and by means of solvents, practically all of the oil is removed as another by-product to be settled, filtered, refined, and otherwise processed into clear, edible oil for salad dressing and frying, and "corn oil foods" or "soap stock" for soap manufacture. The residue of the germ, after oil-extraction. is ground and marketed as corn germ meal, or may become a part of corn gluten feed or meal.

The remaining mixture of starch, gluten, and bran (hull), which is finely ground, is washed through a series of screens to sieve the bran from the starch and gluten. The hull becomes part of corn gluten feed.

The remaining mixture of gluten and starch is pumped from the shakers to high speed centrifugal machines, which, because of the difference in specific gravity, separate the relatively heavier starch from the lighter

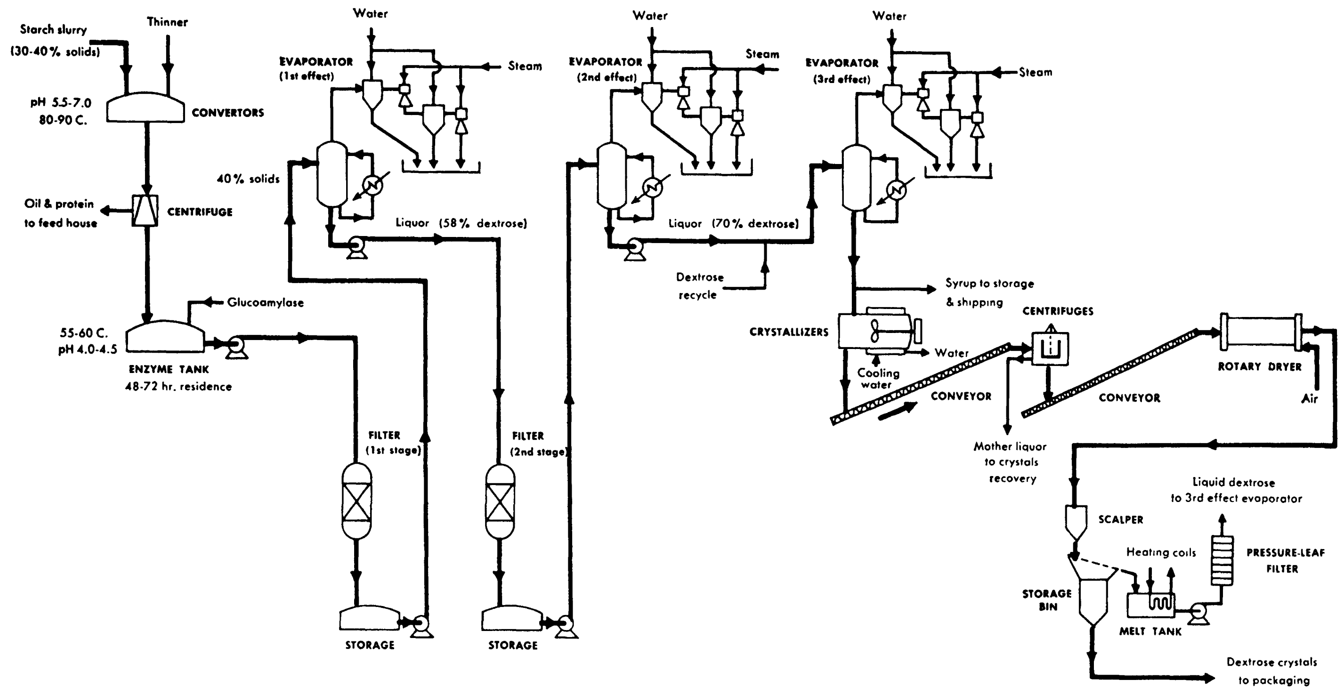


Fig. 4. Enzyme process for converting starch into dextrose. (*A. E. Staley Mfg. Co.*)

gluten. After further processing, the protein-rich gluten is marketed as such, or becomes corn gluten meal, or may be mixed with steepwater, corn oil meal, and hulls to become corn gluten feed. Gluten may also be made to yield a highly versatile protein, *zein*; amino acids, such as glumatic acid, leucine, and tyrosine; and xanthophyll oil, for poultry rations.

Having been separated from the kernels, the starch is now ready for washing, drying, or further processing into numerous dry-starch products, or into dextrin, or for conversion into syrup and sugar. From a 56-pound (25 kilograms) bushel of corn, approximately 32 pounds (14.5 kilograms) of starch result, about 14.5 pounds (6.6 kilograms) of feed and feed products, about 2 pounds (0.9 kilogram) of oil, the remainder being water.

Starch Conversion. More than half of the total production of starch is converted into syrup dextrins or dextrose by acid hydrolysis and/or enzyme action or heat treatment.

Starch, mixed with water, and heated in the presence of weak hydrochloric acid, breaks down chemically by hydrolysis. If the hydrolysis or conversion of corn starch is interrupted before final conversion, a non-crystallizing corn syrup is obtained. Many varieties may be made by supplemental use of enzymes to meet specific functional requirements. The solids content is varied to suit the requirements of the users. Corn syrup is used in a wide variety of food products, including baby foods, breakfast foods, cheese spreads, chewing gum, chocolate products, confectionary, cordials, frostings and icings, peanut butter, sausage, and for numerous industrial products, including adhesives, dyes and inks, explosives, metal plating, plasticizers, polishes, textile finishes, and in leather tanning.

A process for converting starch to dextrose is shown in Fig. 4. The enzyme process shown overcomes flavor and color difficulties of the hydrochloric acid method. The enzyme is obtained by growing a mold (*Aspergillus phoenicis*, a member of the *Aspergillus niger* group). The mold yields the key glucoamylase as well as transglucosylase. The latter must be eliminated because it catalyzes the formation of undesirable glucosidic linkages. Through a special process, almost pure glucoamylase is obtained.

Purified starch slurry (30–40% solids), made from dent corn, is received in the converters from basic processing at the corn plant. A preliminary conversion using alpha-amylase enzyme or acid is carried out at 80–90°C (176–194°F), during which 15–25% of the starch is converted into dextrose. This thins the starch slurry, allowing easier addition of the glucoamylase enzyme. It also prevents formation of unhydrolyzable gelatinous material during the main conversion, and results in increased dextrose yields of from 3–4%. Thinning also reduces evaporation costs because starch concentrations of 30–40% can be handled compared with the 12–20% limit for the acid process. Before the main conversion, the starch-dextrose slurry is centrifuged to remove oil and protein by-products, which are processed for animal feed.

The slurry then goes to a 25,000-gallon (946-hectoliter) enzyme tank where, at pH 4.0–4.5 and 60°C, the major reaction with the glucoamylase takes place. It is a batch operation requiring about 72 hours. When conversion is complete, the batch (97–98.5% dextrose on a dry basis) is passed through a preliminary decolorizing filter of powdered carbon and then pumped on to the first of three evaporators. The remaining operations are evaporation and crystallization, followed by centrifuging, and rotary drying for dextrose crystals; and by a remelting and filtering process for handling of outsize crystals, the resulting liquid being returned to the third effect evaporator for reprocessing.

STAR CONNECTION. The connection of the various phases of an ac machine or circuit in which one end of all phases is connected to a common point, the other end of each phase going to a line. The *Y* is the three-phase star and is the most common example of this type connection.

STARK EFFECT. In 1913, Stark showed that every line in the Balmer series of hydrogen, when excited in a strong electric field of 100 kilovolts per square centimeter or more, is split into several components. If the spectrum is observed perpendicular to the field, some members of the line pattern are plane-polarized with the electric vector parallel to the field (*p*-components) and the others are polarized with the electric vector normal to the field (*s*-components). When the spectrum is observed parallel to the field, only unpolarized *s*-components are observed. A similar splitting of lines is noted in the cathode dark space of a discharge tube. The Stark effect is similar, in many respects, to the Zeeman effect but it is generally more difficult to study experimentally because of the high potential gradients needed in the light source. Its theory is quite different, and the observed spectral pattern varies markedly in character and in number of components as the field intensity increases.

STARLING (*Aves, Passeriformes*). A bird of any of several species native to Europe, Asia, and northern Africa. The African glossy starlings belong to a family containing also the Asiatic grackles of hill mynas, while the true starlings are placed in a closely related family.

Starling.

The one species which is common in North America is the common European starling, *Sturnus vulgaris*, which was introduced into New York in 1890 and has since spread more than a third of the way across the continent. The males are black with green and blue iridescence and light tips on many of the smaller feathers, and the females are brownish-gray. The beak is rather large. In Europe the starlings are valued as destroyers of insects and for their song. They are able mimics. The same good qualities are worthy of consideration in America but the birds are also destructive of native species and in the fall they become a nuisance by gathering in great flocks to roost in countless thousands in the trees of residential districts. The net verdict is against them, and they are commonly regarded as pests.

STARVATION. A state of existence without food or with inadequate food. When animals are completely deprived of food their only source of energy for the essential processes of metabolism is the material already present in the body. The carbohydrate stored as glycogen is quickly used up, leaving only the stored fat, the circulating protein (see **Amino Acids**), and the protein of the tissues. Studies of the progress of starvation in mammals have shown that the fat is almost completely used up and that the greatest loss of tissue proteins is from the muscles, due to their great bulk. In percentages, however, the liver, spleen, and gonads lose more of their bulk than other parts of the body, and the heart and central nervous system are maintained at the expense of the other parts with very little loss of substance. Extensive observations of the details of metabolism and bodily changes during starvation allowed to continue to the death of the animal have been recorded.

A remarkable result of starvation in some of the lower invertebrates is a progressive shrinkage of the body as a whole, in contrast with the emaciation which results in vertebrates. Observations on the flatworm, *Planaria*, have shown that ultimately it even retraces its development to assume an embryonic form.

STATE. 1. In its fundamental connotation, this term refers to the condition of a substance, as its state of aggregation, which may be solid, liquid, or gaseous—compact or dispersed.

2. As extended to a particle, the state may denote its condition of oxidation, as the state of oxidation of an atom, or the energy level, as the orbital of an electron, or in fact, the energy level of any particle.

3. In quantum mechanics, the word state is used in its most general context to refer to the condition of a system which is described by a wave function satisfying the Schrödinger equation for the system when this wave function is simultaneously an eigenfunction of one or more quantum mechanical operators corresponding to one or more dynamical variables. If this set of operators includes all those which will commute with the ones in the set, then the state of the system is as completely specified as the Heisenberg uncertainty principle (see **Uncertainty Principle**) allows and is characterized by the eigenvalues of these operators. These eigenvalues are the results which will always be found if measurements of the corresponding dynamical variables are made. In its more limited sense, the word state is used to refer to the condition of the system when its wave function is simultaneously an eigenfunction of the Hamiltonian operator. In this case the system is characterized by a definite value of the energy, i.e., the eigenvalue of the Hamiltonian operator, and is said to be an an *energy eigenstate*, a *stationary state*, or a definite *energy state*.

STATIC AND NOISE. Names commonly applied to all the various random electrical disturbances which are picked up by a radio receiver. These can be divided into two general classes, natural and artificially generated (as from machines) static. The first is caused by various types of natural electrical discharges, the most pronounced being those of lightning. However, a static-producing discharge is not necessarily, or even usually, a visible lightning discharge. Various static charges are often continually building up and discharging in the atmosphere and hence inducing disturbances in the receiver. Cosmic radiations are also responsible for static. These types of natural static are often called atmospherics. The types of artificial static are almost as numerous as the electrical machines which people have developed. Any sparking contact or poor electrical connection will produce static which will be picked up by nearby receivers. Unfortunately many types of this interference may be fed back along the power lines and directly into the receiver. X-ray and diathermy machines are also sources of interference but cannot be properly classed as static.

The elimination of static presents a particularly difficult problem since the frequencies in the static pulse cover a wide band, certain types being more prevalent in some frequency ranges than others. Artificial static is best eliminated by correcting the fault at the source although a filter in the power line often helps if the disturbance is coming into the set through the line. Natural static can be minimized, but not eliminated. For amplitude modulation systems limiting the frequency band to which the receiver responds will reduce the noise. Various types of limiters will also reduce the effect since the static signal is frequently greater than the desired one. Frequency modulation is inherently less susceptible to interfering noise and offers almost noise-free reception.

Noise in Industrial Measurement Systems.[1] Signals entering a data acquisition and control system include unwanted noise. Whether this noise is troublesome depends on the signal-to-noise ratio and the specific application. In general it is desirable to minimize noise to achieve high accuracy. Digital signals are relatively immune to noise because of their discrete (and high-level) nature. In contrast, analog signals are directly influenced by relatively low-level disturbances. The major noise-transfer mechanisms include conductive, inductive (magnetic), and capacitive coupling. Examples include the following:

- Switching of high-current loads in nearby wiring can induce noise signals by magnetic coupling (transformer action).
- Signal wires running close to ac power cables can pick up 50- or 60-Hz noise by capacitive coupling.
- Allowing more than one power or signal return path can produce ground loops that inject errors by conduction.

Conductance involves current flowing through ohmic paths (direct contact), as opposed to inductance or capacitance.

Interference via capacitive or magnetic mechanisms usually requires that the disturbing source be close to the affected circuit. At high frequencies, however, radiated emissions (electromagnetic signals) can be propagated over long distances.

In all cases, the induced noise level will depend on several user-influenced factors:

- Signal source output impedance.
- Signal source load impedance (input impedance to the data-acquisition system).
- Lead-wire length, shielding, and grounding.
- Proximity to noise source or sources.
- Signal and noise amplitude.

Transducers that can be modeled by a current source are inherently less sensitive to magnetically induced noise pickup than are voltage-driven devices. An error voltage coupled magnetically into the connecting wires appears in series with the signal source. This has the effect of modulating the voltage across the transducer. However, if the transducer approaches ideal current-source characteristics, no signifi-

[1]Information furnished by H. L. Skolnik, *Intelligent Instrumentation, Inc.*, Tucson, Arizona.

cant change in the signal current will result. When the transducer appears as a voltage source (regardless of impedance), the magnetically induced errors add directly to the signal source without attenuation.

Errors also are caused by capacitive coupling in both current and voltage transducer circuits. With capacitive coupling, a voltage divider is formed by the coupling capacitor and the load impedance. The error signal induced is proportional to $2\pi fRC$, where R is the load resistor, C is the coupling capacitance, and f is the interfering frequency. Clearly, the smaller the capacitance (or frequency), the smaller the induced error voltage. However, reducing the resistance only improves voltage-type transducer circuits.

Example. Assume that the interfering signal is a 110-volt ac 60-Hz power line, the equivalent coupling capacitance is 100 pF, and the terminating resistance is 250 ohms (typical for a 4- to 20-mA current loop). The resulting induced error voltage will be about 1 mV, which is less than 1 LSB in a 12-bit 10-volt system.

If the load impedance were 100 kΩ, as it could be in a voltage input application, the induced error could be much larger. The equivalent R seen by the interfering source depends on not only the load impedance, but also the source impedance and the distributed nature of the connecting wires. Under worst-case conditions, where the wire inductance separates the load and source impedances, the induced error could be as large as 0.4 volt. This represents about an 8% full-scale error.

Even though current-type signals are usually converted to a voltage at the input to the data-acquisition system, with a low-value resistor this does not improve noise performance. This is because both the noise and the transducer signals are proportional to the same load impedance.

It should be pointed out that this example does not take advantage of or benefit from shielding, grounding, and filtering techniques.

STATICS. Statics is that branch of mechanics which deals with particles or bodies in equilibrium under the action of forces or of torques. It embraces the composition and resolution of forces, the equilibrium of bodies under balanced forces, and such properties of bodies as center of gravity and moment of inertia.

Statics is the oldest branch of mechanics, some of its principles having been used by the ancient Egyptians and Babylonians in their constructions of temples and pyramids. As a science it was established by Archytas of Taras (ca. 380 B.C.) and primarily by Archimedes (287–212 B.C.). It was further developed by medieval writers on the "science of weights," such as Jordanus de Nemore (13th Century) and Blasius of Parma (14th Century). In the 16th century, it was revived by Leonardo da Vinci, Guido Ubaldi, and particularly by Simon Stevin (1548–1620) who laid the foundations of modern statics (inclined plane, equilibrium of pulleys, parallelogram of forces, etc.).

Although the laws of statics can in principle be derived from those of dynamics as a limiting case for vanished velocities or accelerations, statics has been developed, since the end of the 18th century, independently of dynamics. Its fundamental theme, like that of dynamics, is the concept of force, representing the action of one body on another and characterized by its point of application, its magnitude, and its direction (line of action), or briefly by a vector.

The following four principles may serve as the basic posulates for statics: (1) The principle of composition (addition) of forces; (2) the principle of transmissibility of force; (3) the principle of equilibrium; and (4) the principle of action and reaction.

Statistical analysis of framed structures or trusses, collections of straight members pinned or jointed together at the ends, is based on these principles.

STATICS (Graphical). The equilibrium of forces is often treated graphically in such practical problems as the stresses in the members of a framed structure. If three concurrent forces are in equilibrium, the three vectors drawn to a common scale to represent them may be made to form a closed triangle (Fig. 1); or if more than three, a closed polygon (Fig. 2). The principle may be extended and is much used in the

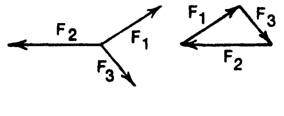


Fig. 1. Three forces in equilibrium.

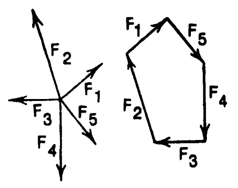


Fig. 2. Five forces in equilibrium.

calculation of the forces in a truss by means of the so-called stress diagram. A simple example is shown in Fig. 3, which represents a small roof-truss with equal loads resting on it at the joints A, B, C, D, E, and supported by the upward reactions of the walls at A and E. The several compartments of the figure are numbered, and both the external forces and the forces acting along the members between these compartments are represented, both in magnitude and direction, by the lines joining the corresponding numbers in the stress diagram. For example, the compressive force in the strut BF is represented by the line 8–9, while the tension in the vertical rod CF is given by 9–10. The closed figure 5–4–10–9–5 in the stress diagram indicates the equilibrium of the forces acting at the joint C. This method of analysis is attributed to Maxwell (see **Bow's Notation**).

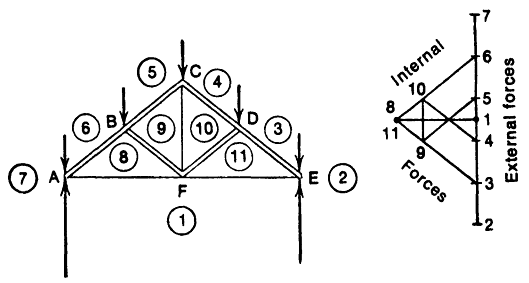


Fig. 3. Elevation of truss with corresponding stress diagram.

In the graphical solution of some types of trusses, it is found, on reaching a particular joint, that the arrangement of members is such that there are more than two unknowns. It is then necessary to replace the unknowns by a substitute system which reduces the number of unknowns at the joint to two. The substitute system consists of a single member inserted in such a way that the truss remains stable and determinate (see **Determinate Structure**). This member is called a substitute, fictitious or phantom member. When the solution reaches a joint where the stress in the members is unaffected by the substitution, the substitute arrangement is replaced by the original arrangement. The graphic procedure is reversed in direction to find the stress in the original members.

STATIONARY STATE. 1. In a quantum-mechanical system, a state corresponding to a definite value of the energy.

2. A thermodynamic system is in a stationary state when its intensive variables are time-independent. The definition includes two classes of situations;

(1) *Equilibrium states*: the system is in thermodynamic equilibrium; no irreversible processes occur in it and there are no exchanges with the outside world. The entropy production vanishes.

(2) *Stationary nonequilibrium states*: in this case irreversible processes proceed in the system. The time-independence of the intensive properties is due to a compensation of the effect of the irreversible processes by exchanges with the outside world.

An isolated piece of metal initially at a nonuniform temperature will, of course, reach a state of thermodynamic equilibrium. But if we cool it at one end and heat it at the other, it will reach a stationary non-equilibrium state. Another example is afforded by a system which receives a component M from the outside environment and transforms it through a certain number of intermediate states into a final product F which is returned to the external environment. After some time a stationary state arises when the concentration of the intermediate components no longer vary with time.

One may say that stationary non-equilibrium states are reached when there exist some constraints which prevent the system from reaching true thermodynamic equilibrium. In the first example it is the difference of temperature at both ends, in the second, it is the value of the concentrations of the initial and final products in the external environment.

STATISTIC. A statistic is a function of the observations in a sample designed to estimate a parameter of the population from which the sample was drawn, or to carry out a test of significance of a hypothesis.

STATISTICAL MECHANICS. One major problem of physics involves the prediction of the macroscopic properties of matter in terms of the properties of the molecules of which it is composed. According to the ideas of classical physics, this could have been accomplished by a determination of the detailed motion of each molecule and by a subsequent superposition or summation of their effects. The Heisenberg indeterminacy principle now indicates that this process is impossible, since we cannot acquire sufficient information about the initial state of the molecules. Even if this were not so, the problem would be practically insoluble because of the extremely large numbers of molecules involved in nearly all observations. Many successful predictions can be made, however, by considering only the average, or most probable, behavior of the molecules, rather than the behavior of individuals. This is the method used in statistical mechanics.

In the general approach to classical statistical mechanics, each particle is considered to occupy a point in phase space, i.e., to have a definite position and momentum, at a given instant. The probability that the point corresponding to a particle will fall in any small volume of the phase space is taken proportional to the volume, and the probability of a specific arrangement of points is proportional to the number of ways that the total ensemble of molecules could be permuted to achieve the arrangement. When this is done and it is further required that the number of molecules and their total energy remain constant, one can obtain a description of the most probable distribution of the molecules in phase space. The Maxwell-Boltzmann distribution law results.

When the ideas of symmetry and of microscopic reversibility are combined with those of probability, statistical mechanics can deal with many stationary state nonequilibrium problems as well as with equilibrium distributions. Equations for such properties as viscosity, thermal conductivity, diffusion, and others are derived in this way.

The development of quantum theory, particularly of quantum mechanics, forced certain changes in statistical mechanics. In the development of the resulting quantum statistics, the phase space is divided into cells of volume h^f, where h is the Planck constant and f is the number of degrees of freedom. In considering the permutations of the molecules, it is recognized that the interchange of two identical particles does not lead to a new state. With these two new ideas, one arrives at the Bose-Einstein statistics. These statistics must be further modified for particles, such as electrons, to which the Pauli exclusion principle applies, and the Fermi-Dirac statistics follow.

It is often possible to obtain similar or identical results from statistical mechanics and from thermodynamics, and the assumption that a system will be in a state of maximal probability in equilibrium is equivalent to the law of entropy. The major difference between the two approaches is that thermodynamics starts with macroscopic laws of great generality and its results are independent of any particular molecular model of the system, while statistical methods always depend on some such model.

STATOR. In machinery, a part or assembly that remains stationary with respect to a rotating or moving part or assembly such as the field frame of an electric motor or generator, or the stationary casing and blades surrounding an axial-flow compressor rotor or turbine wheel; a stator blade.

STAUROLITE. The mineral staurolite is a complex silicate of iron and aluminum corresponding to the formula $(Fe,Mg,Zn)_2Al_9Si_4O_{23}(OH)$ but somewhat varying and may carry magnesium or zinc. It is orthorhombic, prismatic, twins common, often producing cruciform crystals. It is a brittle mineral; fracture, sub-conchoidal; hardness, 7–7.5; specific gravity, 3.65–3.83; luster, subvitreous to resinous; color, dark brown, sometimes reddish to nearly black; grayish streak; tranlucent to opaque. Staurolite is a metamorphic mineral usually the result of regional rather than contact metamorphism, and is common in schists, phyllites and gneisses together with garnet, kyanite, and tourmaline.

Well-known European localities are in Switzerland and Brittany; and in the United States this mineral is common in the schists of New England, and those of the southern Alleghenies. Frequently the crystals are found loose in the soil after the disintegration of the country rock. The name staurolite is derived from the Greek meaning a cross, in reference to the twin crystals, the more nearly perfect crosses being somewhat in demand as curios.

STAYBOLT. The surfaces of pressure vessels, such as boiler drums and tanks, which are not of a natural bulged shape, such as the cylinder or the sphere, must be stayed against bulging by special tension rods called staybolts.

STEADY STATE. A characteristic of a condition, such as value, rate, periodicity, or amplitude, exhibiting only negligible change over an arbitrary long period of time. It may describe a condition in which some characteristics are static; others dynamic. The *steady-state deviation* of a system may be defined as the system deviation after transients have expired.

In medical parlance, a relatively steady condition of processes within the body sometimes is referred to as *homeostasis*.

STEAM CYCLES (Diagram Factor). This is a factor relating particularly to piston and cylinder engines. Although it has a meaning applied to internal combustion engines, diagram factor is principally a dimension of the steam engine. Certain analyses of the steam engine, especially those concerned with predicting the performance of a given unit under stated steam conditions, are most easily made with the use of the diagram factor. This factor is defined as the ratio of the actual mean effective pressure to the theoretical effective pressure; also as the actual work to the theoretical work. The theoretical case is that of a steam engine having no compression, no wire drawing, and no clearance.

The ratio of the area of this cycle to that of the actual engine operating between the same pressure limits, is typical for any one class of engine. The following table gives some values of diagram factor for steam engines. Knowing the steam conditions, and the type of engine, the probable pressure and horsepower realizable can be closely estimated by calculating the theoretical quantities and multiplying them by the diagram factor. In this light, diagram factor is a means of modifying theoretical calculations to bring them in line with actual experience.

DIAGRAM FACTORS

High-speed, simple automatic	0.70–0.85
Low-speed, releasing gear	0.80–0.90
Uniflows	
Full compression, condensing	0.75–0.85
Full compression, non-condensing	0.70–0.80
Controlled compression, condensing	0.85–0.90
Controlled compression, non-condensing	0.80–0.85

$$P_a = P_t \times f$$

$$IHP = \frac{2P_aLAN}{33{,}000}$$

P_a = actual mean effective pressure, in pounds per square inch
P_t = theoretical mean effective pressure, in pounds per square inch
f = diagram factor
IHP = indicated horsepower (steam engine)
L = stroke in feet
A = piston area, in square inches
N = rotative speed, in revolutions per minute

STEAM ENGINE. A positive displacement piston and cylinder machine, which, when supplied with steam at a pressure above its exhaust pressure, uses that steam expansively for the production of power which it makes available as a rotating torque at a crankshaft or flywheel. The steam engine is, with few exceptions, double acting. The steam engine is characterized by moderate or low speeds (100–500 rpm), the use of atmospheric exhaust (or a moderate vacuum), high starting torque, and ease of conversion to reversible operation if desired. The excellence of the steam turbine when a large amount of power is to be generated, especially where the exhaust is at a high vacuum, coupled with the high efficiency of the diesel engine as a prime mover, has restricted the steam engine as a prime mover. However, where a boiler must be supplied anyway, as for the generation of heating steam, the steam engine is usually superior to the diesel engine as a source of power; and where exhaust pressures are high (often the case in industry, where the exhaust is process steam), the steam engine offers advantages which are not seriously challenged by the other types of prime movers.

The principal parts of a steam engine are:

1. The frame or bedplate. In a multi-cylindered engine, this takes the form of a crankcase to which the cylinders are attached, but in a single-cylindered engine, the cylinder is often integral with the frame.
2. Cylinder, with valve chest.
3. Piston, piston rod, crosshead, connecting rod. This mechanical linkage receives a push from steam pressure at one end, and delivers it as a torque force on the crank.
4. Crankshaft, bearings, and flywheel. This part of the engine accomplishes the conversion of reciprocating to rotary motion, supports the shaft for power offtake, and steadies the speed.
5. Valves and valve gear. The device for admission and release of the steam to and from the cylinder, together with the means for actuating it from the crankshaft.
6. Governor. Stationary engines are automatically regulated for constant speed by means of a governor.
7. Lubrication. The piston and cylinder are lubricated by oil mixed with the steam.

A steam engine converts from 5 to 15% of the heat supplied to it into work, depending on the state of the steam supplied, and on the exhaust pressure. The heat unconverted is composed: first, of heat remaining in the exhaust steam; second, initial condensation; third, incomplete expansion; fourth, wire drawing; fifth, friction; and sixth, radiation (negligible). The first of these is the largest, and is reducible only within certain limits. Incomplete expansion results from the release of the steam at the end of the stroke at a pressure higher than the exhaust. By using a longer stroke, this could be eliminated, but there is a point beyond which the increased cost of the engine more than offsets the gain derived by eliminating this loss.

The cycle upon which the engine operates is briefly described as follows (see accompanying diagram): Slightly before the piston reaches the dead-center position corresponding to minimum cylinder volume, the valve connects the cylinder with the steam line so that as the piston starts on its outward travel, the full steam pressure is acting on it. The beginning of this action is known as the event of *admission*. When some 20 to 30% of the stroke has been completed, the valve closes the port on the event known as *cutoff*, and during the remainder of the stroke, the steam is expanded adiabatically to the accompaniment of decreasing pressure. Near the end of its stroke, the valve again opens the port, this time connecting the cylinder with the exhaust line. This event is known as *release*. The cylinder remains connected with the exhaust during the return stroke of the piston, and the steam is expelled until approximately $\frac{2}{3}$ of the return stroke has been completed. The valve then closes the port, and the remaining steam is trapped in the cylinder and compressed. The beginning of this process is known as the event of *compression*. The four events just described govern the form of the steam engine cycle. The engine using it will be able to develop a horsepower hour using from 10 to 25 pounds of steam, depending upon the expansion permitted by the terminal conditions of the steam.

The steam engine may be mechanically controlled to give variable output so that when it is connected to a load which varies, it maintains nearly constant speed. There are two methods of accomplishing this result. In one, called cutoff governing, the percent of the stroke during which the valve connects the cylinder with the boiler is varied, and in this way, different amounts of steam are admitted to the cylinder at one pressure. The mechanism to effect this type of control is incorporated in the valve drive. The other method, called throttling governing, consists of interposing an artificial resistance to create a pressure drop between the boiler and the engine, so that although the same volume is admitted on each stroke (the cutoff being constant), the weight of steam admitted will vary because of the variation in density created by throttling. The governor, in this case, operates on a throttle valve located at the steam inlet.

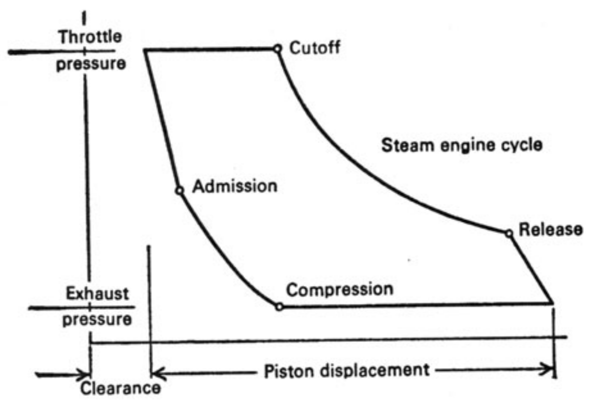
STEARIC ACID AND STEARATES. Stearic acid $H{\cdot}C_{18}H_{35}O_2$ or $C_{17}H_{35}{\cdot}COOH$ or $CH_3(CH_2)_{16}{\cdot}COOH$ is a white solid, melting point 69°C, boiling point 383°C, insoluble in water, slightly soluble in alcohol, soluble in ether. Stearic acid may be obtained from glyceryl tristearate, present in many solid fats such as tallow, and in smaller percentage in semi-solid fats (lard) and liquid vegetable oils (cottonseed oil, corn oil), by hydrolysis. The crude stearic acid, after separation of the water solution of glycerol, is cooled to fractionally crystallize the stearic and palmitic acids, which are then separated by filtration (oleic acid in the liquid), and fractional distillation under diminished pressure. With sodium hydroxide, stearic acid forms sodium stearate, a soap. Most soaps are mixtures of sodium stearate, palmitate and oleate.

The following are representative esters of stearic acid: Methyl stearate $C_{17}H_{35}COOCH_3$, melting point 38°C, boiling point 215°C at 15 millimeters pressure; ethyl stearate $C_{17}H_{35}COOC_2H_5$, melting point 35°C, boiling point 200°C at 10 millimeters pressure; glyceryl tristearate [tristearin $C_3H_5(COOC_{17}H_{35})_3$], melting point 70°C approximately.

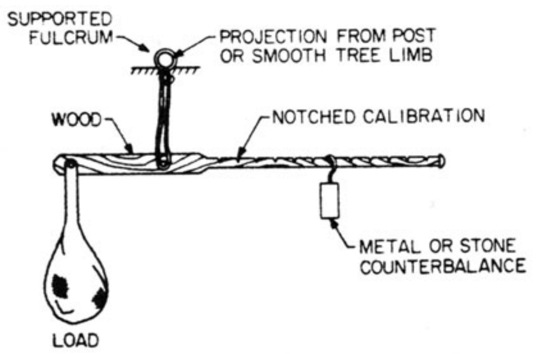
Stearic acid is used (1) in the preparation of metallic stearates, such as aluminum stearate for thickening lubricating oils, for waterproofing materials, and for varnish driers, (2) in the manufacture of "stearin" candles, and is added in small amounts to paraffin wax candles. As the glyceryl ester, stearic acid is one of the constituents of many vegetable and animal oils and fats.

See also **Rubber (Natural).**

STEELYARD. An early type of scale that employs an unequal firstclass lever arm constructed so that the fulcrum point can be suspended from above, such as by a hook on the end of a chain, the other end of the chain being fastened to a ceiling support. The load is sus-



Steam engine cycle.



Steelyard used for weighing since ancient times.

pended from the end of the short lever arm. A sliding poise on the long end of the lever arm is positioned to balance the load. Although largely replaced by spring scales, the steelyard, because of its portability and low cost, still finds limited, use, particularly in underdeveloped nations.

STELE. The vascular tissue of the axis of a plant is called the stele or central cylinder. The principal tissues composing it are the xylem and the phloem. Others are pith, pith rays, cambium, and pericycle. The most primitive type is the protostele, which consists of a solid central mass of xylem surrounded by a cylinder of phloem. There is no pith in a protostele. Protosteles are found in the roots of all plants and in the stems of some of the ferns. Many would set off the stele of the root as a separate type known as a radial stele, since the central mass of xylem is not a cylinder but has several arms projecting outward from its surface, with the phloem concentrated between these arms. In stems, only those of certain Lycopsida like *Lycopodium*, have radial steles.

STELLAR HEAT INDEX. As applied to a star, the difference between the visual and radiometric magnitudes: i.e.,

$$\text{heat index} = m_v - m_r$$

where m_r is the radiometric magnitude and m_v is the visual magnitude. A radiometric magnitude is obtained by measuring the overall energy brightness of a star by means of a thermocouple or bolometer, and by reducing the readings to outside atmosphere after correcting for losses in the optical system.

STELLAR LUMINOSITY. The intrinsic brightness of a star may be expressed either in terms of its absolute magnitude or in terms of the sun's brightness as unity. The luminosity of a star is defined as its intrinsic brightness in terms of the brightness of the sun as unity. That is, if our sun were replaced by a star of luminosity 100, the light received by the earth would be 100 times as great.

STELLAR MAGNITUDE. In the first star catalogues issued by Hipparchus and Ptolemy, the relative apparent brightness of the stars was designated by a system of six numbers referred to as the magnitudes of the stars. Twenty of the brightest stars were referred to as first magnitude, and those at the limit of visibility were called sixth magnitude. The stars with brightness intermediate between the two extremes were assigned to a magnitude number with the numbers increasing with faintness of the stars. With the application of the telescope to astronomy, many faint stars were discovered and the need for additional magnitude numbers became evident. Unfortunately for modern astronomers, the attempt was made to amplify the ancient magnitude system not only to include the fainter stars, but also to indicate finer gradations of brightness by a decimal system. The result is that astronomers are now using a system that was started about 2,000 years ago and has all the clumsiness and inconvenience for modern observers that is characteristic of so many of the ancient scientific instruments.

There is no definite evidence that Hipparchus or Ptolemy had any idea in mind at the time they first used the magnitude system other than to provide a rough descriptive term for the stars. In the early part of the nineteenth century, Sir John Herschel found that the apparent brightness of a first magnitude star is about 100 times that of a sixth magnitude star. In 1850 Pogson proposed a fixed scale of stellar magnitudes based upon the original scale of Hipparchus and Ptolemy, but so adjusted that it would agree at the sixth magnitude with the system employed by Argelander in his famous Bonner Durchmusterung. Adopting the announcement of Herschel that the ratio of brightness of a first and sixth magnitude star is approximately 100, Pogson proposed that the ratio between successive magnitudes should be $\sqrt[5]{100}$ or approximately 2.512. This leads to an analytical expression for the magnitude scale as follows:

Let I_1 be the intensity of a star of apparent magnitude m_1, and I_2 the intensity of a star of apparent magnitude m_2; then

$$\frac{I_1}{I_2} = 2.512^{(m_2 - m_1)}$$

or

$$\log I_1 - \log I_2 = 0.4(m_2 - m_1)$$

Since the magnitude scale is a scale of relative brightness, it is necessary to establish a system of standards. For this purpose, a group of stars in the immediate vicinity of the north celestial pole has been selected. The magnitudes of the stars in this "north polar sequence" have been very carefully determined and agreed upon by the International Astronomical Union. All magnitudes determined at the present time should be referred, either directly or indirectly, to this standard sequence. More recently, Johnson and Morgan have set up a series of "photometric standard stars" whose magnitudes are strictly given on a specified system.

The magnitude scale as originally established referred to the relative apparent visual brightnesses of the stars. With the application of photography to astronomy, difficulty with the magnitude scale immediately became evident. If we have two stars of the same visual magnitude, one of them blue and the other red, the photographic image of the blue star will be much stronger than the photographic image of the red star. The colors of the stars in the sky vary with the different spectral types, and the visual magnitude differences between a number of stars of different spectral types will differ considerably from the magnitude differences obtained by photographic means. Furthermore, the photographic magnitudes, so-called, will be different, depending upon the type of plate used and the characteristics of different telescopes, and it becomes necessary to be very explicit in defining the particular range of wavelengths of spectral energy that are to be used in any magnitude scale. The difference between the photographic and visual magnitudes of a star is known as the color index of the star, the term arising from the fact that the color is the determining factor in the magnitude scale difference.

With the application of various other types of radiation measuring instruments, such as bolometers and radiometers, to the measurement of the apparent brightness of the stars, the necessity has arisen for various different magnitude scales, such as bolometric magnitude, radiometric magnitude, etc. The problem of the intercorrelation of the different systems is at present in a very confused state and much research is being carried on in this important field. It is hoped that in the future some system of expressing the apparent brightnesses of the stars may be devised that will replace the present complicated inverse logarithmic scale of magnitudes.

For the purpose of expressing the intrinsic brightness of a star, independent of the distance of the star from the earth, a system of absolute magnitudes has been devised.

The apparent brightness of a star, or any other luminous object, depends both upon the intrinsic brightness of the object and also upon its distance from the observer. In the case of the stars, the apparent brightness, expressed as stellar magnitude, may be determined by any one of the standard methods of stellar photometry. In case the distance of the star is known, the intrinsic brightness may be immediately calculated. Conversely, if there is any method available for determining the intrinsic brightness of a star independently of a knowledge of the distance, this distance may be computed from the ratio between the apparent and intrinsic brightness.

The absolute magnitude of a star is the apparent brightness, expressed on the magnitude scale, that a star would have if it were situated at a distance of ten parsecs from the sun or, in other words, if the stellar parallax of the star were $\frac{1}{10}$ of a second. Analytically, the absolute magnitude, M, of a star is connected with the apparent magnitude, m, and the stellar parallax, π, by:

$$M = m + 5 + 5 \log \pi$$

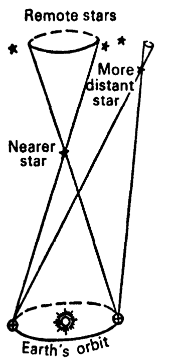
On this scale, we find the sun, with apparent magnitude -26.72 and parallax 206265″, to have an absolute magnitude of 4.85. Antares with parallax 0″.009 and apparent magnitude 1.22 is found to have an absolute magnitude of -4.0. On the basis of these absolute magnitudes and the defining relation of the magnitude scale, we find that the brightness ratio of Antares to the sun is 3470, or that the star Antares is actually 3470 times brighter than the sun.

The zero point of the absolute magnitude scale is arbitrary. By convention, a normal A0 main sequence star has an absolute magnitude of zero and a color index of zero, as well.

STELLAR PARALLAX. A term used by astronomers as a means for expressing the distance of a given star. Technically defined, stellar parallax is the angle that would be subtended by the mean distance of the earth from the sun (one astronomical unit) at the distance of the star from the sun. The symbol π is used, and as the angle is always small, it is always given in seconds of arc; this measure is understood unless explicitly stated otherwise.

From the earliest days of the Pythagoreans, any theory of the structure of the universe which postulated that the earth might move about the sun was objected to on the ground that such motion should produce an apparent motion of the stars. Copernicus, in proposing his heliocentric theory, met this objection by postulating that the distances of the stars were so incomparably greater than the distance of the earth from the sun that no instrumental methods would be capable of detecting the motion even if it did exist. The attempts to test the Copernican doctrine by searching for this so-called stellar parallax gave a tremendous impetus to the design of accurate instruments, but even with the improvement in instrumental equipment, the effect was not observed, and the Copernican theory lost ground. It was not until 1838 that Bessel was able to definitely prove that the effect is present.

The type of effect to be looked for is illustrated in the figure. The type of curve that the stars should apparently follow due to the earth's motion about the sun varies from an eclipse with eccentricity equal to that of the earth's orbit for stars at the pole of the ecliptic, to oscillations back and forth along a straight line for stars in the plane of the ecliptic.



Parallaxes of the stars. Owing to Earth's revolution, the nearer stars describe parallax orbits annually with respect to the remote stars.

The problem of determination of stellar parallax is theoretically very simple. All that is necessary is to make a series of observations of the positions of a star on any system of spherical coordinates (e.g., right ascension and declination) and from the observed changes in position throughout the year determine the stellar parallax. This so-called absolute method was attempted many times but failed to reveal any definite value because of the fact that the instrumental errors were larger than the effect sought. This is not surprising when we consider that the largest stellar parallax ever found (for the star Proxima Centauri) has a value of 0″.783, equivalent to the angle subtended by a ten-cent piece at a distance of approximately 3 miles. (See **Spherical Polar Coordinates.**)

With the failure of the absolute method to yield values for the parallaxes of the stars, Bessel and Struve decided upon an indirect or relative method for determining the desired quantity. This method is based upon the assumption that certain stars are at such a great distance that their parallaxes are too small for detection, but that there are other stars closer to the sun which should show motion relative to the distant background. Bessel selected the star 61 Cygni, which was assumed to be relatively close to the earth from a large proper motion, while Struve selected the star Vega, which has an appreciable proper motion and is also so bright as to imply closeness to the earth. Proceeding by different methods, both Bessel and Struve, in 1838, were able to show that the stars they had selected showed parallactic motion relative to the background of stars.

Until the application of photography to astronomy, the problem of determination of stellar parallaxes was very tedious and laborious, and up to 1880, the distances of fewer than 25 stars had been determined. With the application of photography, the progress of parallax determination became much more rapid, and at present, a number of long programs of observations in both the northern and southern hemispheres have been completed.

The photographic method consists in first selecting stars that are suspected, either from proper motion, spectral type, or other characteristics, to be relatively close to the sun. Plates are taken of these stars, great care being exercised in the guiding, and the brightness of the "parallax star" is reduced until its photographic image compares favorably with the images of the fainter "background stars." The plates are all taken at the same hour angle, either east or west, to eliminate so far as possible atmospheric effects, and the dates on which the plates are taken are separated as much as they can be to make the effect of the earth's motion as large as possible. Twenty or thirty plates, extending over several years, are taken, and the position of the parallax star is carefully measured with reference to half-a-dozen background stars. A least squares solution will yield the motion of the star relative to the background. This motion will consist both of the proper motion and the parallactic shift. The former may be separated from the latter, because proper motion is linear in character, whereas the parallactic shift is periodic. For stars within 5 million times the sun's distance from the earth (parallax 0″.04), the mean of two or three determinations will be correct within 20 percent. For twice this distance, the results are only accurate enough for statistical purposes, and beyond this distance, the trigonometric method is practically valueless. Occasionally, due to an unfortunate choice of parallax star or comparison stars, the value of the stellar parallax comes out to be a negative quantity. Such a "negative parallax" simply means that the star under observation is more distant than those selected for comparison purposes. For the more distant stars beyond the range of the trigonometric method, certain other methods are available, such as: parallaxes of members of moving clusters; mean parallaxes; dynamical parallaxes of double stars; and spectroscopic parallaxes.

See also **Dynamical Parallax;** and **Star.**

STEM (Plant). The stem of a plant is that part which bears the leaves and flowers and later fruits. Commonly it grows erect, lifting these various organs up above the ground. It is readily distinguished from the root by being separated into joints or nodes and internodes, and by bearing leaves or by having on its surface scars left by the falling of leaves.

Stems may be classified in several ways. If one considers the duration of the stem, it becomes an annual lasting but one year, a biennial living 2 years, or a perennial stem which grows for several years. If one considers the internal structure of the stem, he finds it to be herbaceous or woody. An herbaceous stem is one which is largely made up of parenchymatous cells, without a great mass of woody tissue. In temperate regions such stems last but a single year, at the end of which they die. In annuals the entire plant dies, while in herbaceous perennials the top dies, but the basal portion including the root and the lower stem lives on. Woody plants are those in which the stem is predominantly composed of vascular tissues. Such plants are either trees or shrubs or vines. Trees are commonly distinguished by the existence of a single stem or trunk which does not branch at its base, whereas in shrubs no single trunk exists, but several of equal size result from basal branchings. Branching is either excurrent or deliquescent. When excurrent, the trunk is distinctly recognizable throughout its length, the branches coming from it being much smaller, as in many conifers such as spruce or fir. Usually such trees have a conical shape, due to the progressively smaller branches from bottom to top of the tree. In deliquescent branching, the main trunk branches into several large branches which in turn divide, as in the elm. Vines are distinguished by their long relatively slender stems which usually require external support. Another classification separates erect stems from those which are procumbent or trailing on the ground, from stems which clamber over supports, and from twining stems, which wind tightly around any supporting object.

In many plants the stem is very much reduced in size, appearing as a small often flattened ball, as in the common Cyclamen or in certain Cactus plants. Other plants are said to be stemless, the stem existing only as a small object at the top of the root, the leaves arising from it seeming to come directly from the top of the root. A familiar example is the dandelion. The first year of growth in many biennials results in a similarly stemless plant; carrots, beets, and parsnips are common examples.

As previously noted, one of the functions of the stem is to elevate the leaves into a position where they may function most efficiently and the flowers to a position where they may become more conspicuous and where the resulting fruits may be better scattered. Not only does the stem perform this function, but it also permits a great increase in the number of leaves and flowers which may be borne. The stem is the organ through which sap ascends from the root to the leaves and through which organic materials elaborated in the leaves pass to the place of storage. The stem itself may be the place in which materials are stored.

The stem develops from a bud. In the seeds, the embryo has a terminal rudimentary bud, the plumule, from which the first stem develops. The tip of this stem bears a terminal bud, from which further increase in the stem is developed. In the axils of the leaves lateral buds are formed which develop into branches. Elongation of the stem takes place only in the tip, extending downward therefrom a few inches, and is caused by the cellular changes like those that occur also in the elongating root. In some plants, as in Grasses, intercalary growth also occurs. This is growth in the region of the older nodes of the stem. Increase in diameter of the stem results from the divisions of special cells called cambium cells.

When the extreme variations of stem are considered, with their range from tiny plants less than half an inch in height to forest giants towering 300 feet (91 meters) and over, and from vines with slender wiry stem less than $\frac{1}{16}$ inch in diameter through succulent herbs to sturdy trunks 40 feet or more through, it is not surprising that stem structure should be very variable and often complex. Yet they are all composed of the same types of cells and are all arranged on two fundamental patterns, one found in dicotyledonous plants, the other characterizing the monocotyledons.

In dicotyledons, the growing tip of the young stem has cells which are all alike, having a dense cytoplasm, and large nuclei, which, if the stem is growing, will be dividing frequently. These cells comprise the promeristem, the region where active cell division occurs, but little change in cell form.

As these cells increase in number, some are carried ahead, while others remain unchanged in position. The latter gradually show very evident changes in size and shape, and in the nature of their walls. The outermost layer, called the dermatogen or protoderm, is made up of somewhat flattened cells which will become the epidermis, a protective covering against entrance of disease-producing organisms and against excessive loss of water. Within the body of the stem tip certain strands of cells become distinct by their elongate shape and dense protoplasmic content. These procambium strands are the beginnings of the vascular tissue presently to appear. In cross sections of the stem the procambium appears as a ring of separate masses of cells. The remaining cells of the growing tip are parenchymatous cells, called the ground meristem, changed but little from the promeristem condition.

As the procambial cells grow older, they gradually change in form. Those cells which are nearest the center of the stem become xylem cells, those towards the circumference of the stem become phloem, with cambium cells separating the two types. In some stems the differentiation of cells continues until the procambial strands have united to form a continuous cylinder which gives place to concentric rings of xylem and phloem cells separated by a band of cambium cells. In other stems, the strands remain distinct, forming separate vascular bundles. The ground meristem or parenchyma in the center of the stem becomes the pith, that surrounding the strands becomes the cortex and pericycle, while the radiating masses of parenchyma cells between the separate strands make up the pith rays. All these tissues, derived indirectly from the differentiation of the promeristem cells, form the primary body, composed of primary tissues.

The epidermal cells are often somewhat elongated in the direction of the length of the stem. The outer wall of an epidermal cell is frequently much thickened and cutinized, so that it becomes impervious to water. Many stomata are found in the epidermis. As the diameter of the stem increases, the epidermal cells gradually become stretched until they finally break apart and are lost.

The cortex inside the epidermis is comprised of several kinds of cells. See Fig. 1. Those nearest the surface are the collenchyma cells.

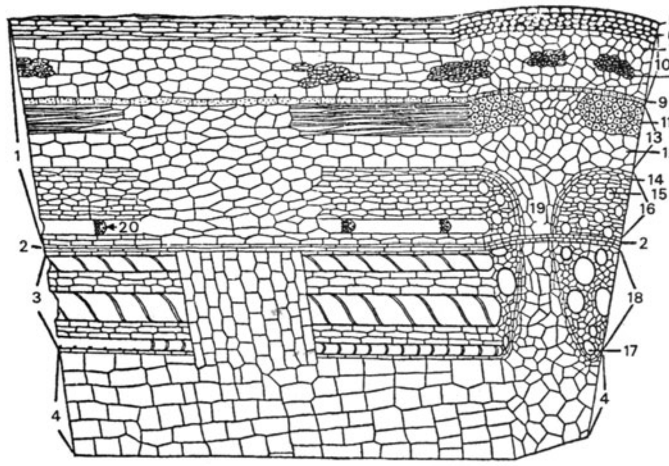


Fig. 1. Diagram showing the tissues derived from the primary meristems: (1) bark; (2) cambium; (3) wood; (4) pith; (5) epidermis; (6) collenchyma; (7) parenchyma of cortex; (8) stone cells; (9) starch sheath; (10) primary cortex; (11) pericyclic fibers; (12) pericyclic parenchyma; (13) pericycle; (14) phloem parenchyma; (15) sieve tube; (16) phloem; (17) vessel; (18) xylem; (19) pith ray; (20) sieve plate. (*After Stevens.*)

These are modified parenchyma cells, the walls of which are thickened in their angles. These cells usually contain chloroplasts. Because of their thickened walls, collenchyma cells give support to the stem, while at the same time they manufacture food. The parenchyma cells of the cortex are thin-walled, and either rounded in shape or, through mutual pressure, more or less angular. Those near the surface of the stem often contain chloroplasts, and carry on photosynthesis. They give rigidity to the stem because of their turgor pressure, and also serve as storage tissue. In some stems there are also found in the cortex thick-walled sclerenchyma cells. These may be either long slender fibers or short stone cells. In a few plants the innermost cells of the cortex form a definite endodermis, the walls of each cell being much thickened.

The tissues inside the cortex include the pericycle, the vascular bundles, pith rays, and the pith. This is the stele.

The cells of the pericycle are very similar to those of the cortex, so much so that it is often very difficult to distinguish one from the other. In the pericycle of stems, which usually is much thicker than that in roots, sclerenchyma cells, both fibers and stone cells, may be found along with the parenchyma, just as in the cortex.

The cells comprising the vascular system—xylem and phloem—are very much modified. The procambial cells towards the center of the stem first elongate greatly, without appreciably increasing in diameter. Soon changes appear in the wall, secondary deposits of cellulose being laid down against the primary wall. The manner in which the wall is thickened varies in different cells. In some the thick deposits are in rings: these are formed while the cell is still elongating and so are gradually separated. Such cells are called annular cells. In other cells of the first-formed xylem, which is called protoxylem, the wall thickenings are in the form of spirals, producing spiral cells, which allow a certain amount of growth even when the thick wall is formed. In cells which differentiate later, when elongation has been completed, the thickening of the wall will be much more extensive, only irregularly distributed, narrow slits being left unthickened. These slits extend transversely in the wall of the cell. Such cells are called scalariform or reticulate cells. Differentiation of the xylem cells continues until most of the inner part of the procambial strand has been changed to xylem cells. Those which form after the narrow protoxylem cells are called metaxylem cells. Protoxylem and metaxylem together make up the primary xylem. It should be understood that, as the xylem cells develop, the end walls are cut away forming long tubes, called vessels or tracheae. Water moves rapidly through such tubes.

The cells on the outside of the procambial strand become the primary phloem cells. These are the sieve tubes and companion cells. Early stages in the formation of these cells are much like those of xylem cells, elongation first occurring and then changes in the cell wall. In the formation of phloem cells, a single cell divides into two which become very unequal; the larger one continues to increase in size and loses its nucleus; the smaller one frequently divides again. The walls of these cells remain comparatively thin, with characteristic perforated thin places, called sieve plates, in the end walls of the larger cells, which form the sieve tubes. The smaller cells, called companion cells, are characterized by a dense cytoplasmic content, small vacuoles and prominent nucleus. They are connected with the sieve tubes by numerous small thin places, called simple pits, in their walls. Phloem cells are channels in which food passes through the stem. In addition to these various cells, the phloem contains parenchyma cells which are used for storage of materials and in some plants long thick-walled fibers, called phloem or bast fibers.

Parenchyma and fibers also occur in the xylem. Between the xylem and phloem elements there is a band of cells which remain unmodified and become an important tissue in many stems. This is the cambium, which by its divisions gives rise to the secondary tissues which compose the bulk of the stems of woody plants. These secondary tissues are the secondary xylem and phloem, and differ only slightly from the primary xylem and phloem, being unlike mainly in their origin. Secondary rays, which provide for lateral translocation of food, are also produced by the cambium.

In the center of the stem is the pith, composed of large, thin-walled cells arranged in irregular fashion. They function principally as places of storage of food.

In nearly all monocotyledons, no cambium is formed, therefore the monocot stem is composed entirely of primary tissues. The arrangement of these tissues is vastly different from that of the stems of dicotyledons. The vascular tissues occur in the form of separate small bundles which are scattered throughout the stem. It is impossible to distinguish any limit separating cortex from pericycle and pith. In many monocots the central portion of the stem is entirely free from bundles and recognized as a pith. Often the pith breaks up, forming a hollow stem.

All gymnosperms have woody stems. The development and structure of these are quite similar to that of dicotyledons, but, except in a few uncommon species, the xylem is composed entirely of tracheids, distinguished from the tracheae of the angiosperms by the fact that they are elongated cells instead of long tubes formed from many cells, and no companion cells are formed in the phloem.

In most plants the function of the stem is to display the leaves and reproductive organs in the most favorable position and to carry materials from one part of the plant to another. In many plants the stem is a highly specialized structure with different functions. Often these specialized stems take over the function of one of the other organs of the plant.

The outer tissues of the young stems of nearly all plants are green. Therefore some photosynthesis takes place in these tissues. There are many plants in which the stem is the principal, if not the only, place where photosynthesis occurs. In many cases, as in some of the cacti, the appearance of the stem differs very little from that of any other plant; but leaves are very much reduced or entirely lacking, all photosynthesis occurring in the stem. In other species of cactus, such as the Prickly Pear and the widely cultivated crab or Christmas cactus, the stem is very much flattened, but still distinctly recognizable as a stem. In some plants, however, the modification has become extreme. The ultimate branches of the stem have become very much flattened and have a shape which gives them every appearance of a leaf. Only their position in the axil of a tiny scale, the real leaf, betrays their true nature. The dainty Smilax, *Asparagus asparagoides*, of the florist, has branches of this kind. So also does the Butcher's Broom, a marsh plant of Europe, which is widely cultivated and appears during the Christmas season, stained a brilliant scarlet. The inconspicuous greenish-white flowers of this plant occur in the center of that part which is commonly assumed to be a leaf. The tiny needle-like "leaves" of the garden asparagus are really branches.

In a few plants the stem becomes a very important reproductive part. See Fig. 2. Runners, long slender branches from the base of the stem, grow out horizontally over the surface of the ground, and root at their tip. There a new plant is formed. With death and disintegration of the connecting stem, the young plant becomes separate from its parent.

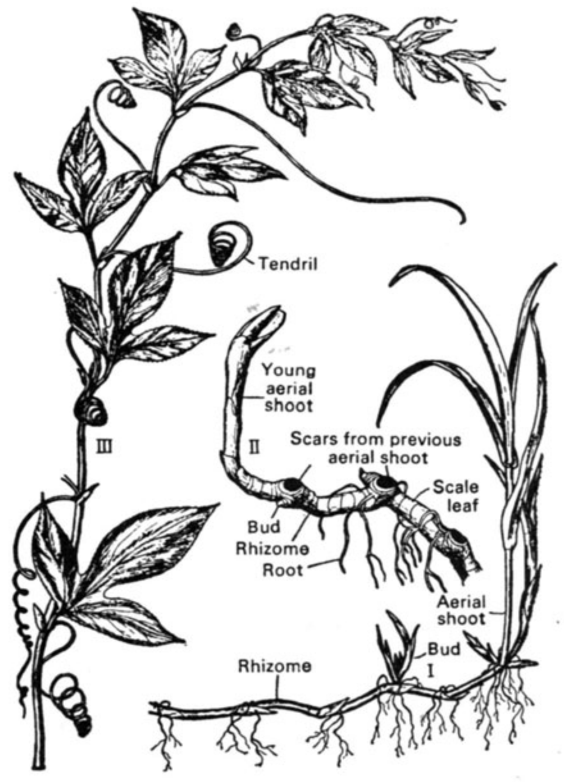


Fig. 2. Types of modified stems: (1) rhizome or quack grass (*Agropyron repens*); (II) rhizome of Solomon's Seal (*Polygonatum commutatum*); (III) stem of passion flower (*Passiflora incarnata*) with tendrils that are modified stems.

This is a form of vegetative reproduction. A stolon differs very little from a runner; it is a prostrate branch which regularly roots at its nodes, and sends up new plants not only at its tip but also from the nodes. Rootstocks or rhizomes are spreading underground branches which produce adventitious roots at their nodes and often spread widely.

Many rhizomes become very fleshy because of an accumulation in them of food materials. Their principal function then is storage. Storage occurs in the stems of many different plants. If only a portion of the rhizome becomes enlarged, it is called a tuber. The common white potato is a very familiar tuber, which is formed at the tip of a slender rhizome. See Fig. 3. The "eyes" of the potato are really the nodes of the stem; from them buds will give rise to branches when the potato grows. More modified than the tuber is the corm. This is a short thick erect rootstock. Often it is much broader than it is long. Buds are formed on the upper surface of the corm. Each of these buds grows into a new plant, exhausting the substance of the old crom. New croms form at the base of the old one. Corms known to all are those of *gladiolus* and *crocus*. These are usually incorrectly called bulbs. See Fig. 4. A bulb is a fleshy bud, composed of a short thick stem and many fleshy or scaly leaves or leaf bases. Onions form true bulbs, as do tulips, hyacinths and many lilies.

The materials stored in the stems so far considered are mainly reserve food. Other stems become swollen with stored water. Many cacti and Euphorbias have stems of this sort.

Thorns are usually stems or branches which have become stiff and pointed and serve to protect the plant. Tendrils are organs which support a plant as it grows up through other plants. Not all tendrils are stems.

Fig. 3. Potato tuber with scale-like leaf at "eye."

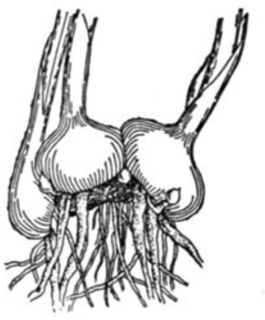


Fig. 4. Corms of *Gladiolus*, consisting chiefly of fleshy stems.

Some, like those of the grape, are definitely so; others are leaves or parts of leaves. Usually plants which have tendrils have slender stems which lack sufficient strength to support themselves. Plants of this type are vines. There are several types of vines, or climbing plants. One of them supports itself by twining tightly around any available support. The direction of twining is very constant for any species, some invariably turning in a clockwise direction and others counterclockwise.

Other climbers support themselves by adventitious roots which form in abundance and cling tightly to any support. Other climbing plants are supported solely by the presence of many spines or prickles, often hooked, or pointed backwards so that the stem does not easily slide off any object on which it rests. Climbing roses illustrate this type of climber. But many tropical vines are much better illustrations. Often these grow to great lengths, hanging in long festoons from the tops of tall trees, or growing in tangled masses over low shrubby plants. In these tropical climbers, which are commonly called lianas, the stems often assume curious flattened or fluted or irregular shapes. In diameter they vary from a fraction of an inch to many inches; they may attain a length of 400 or 500 feet (122 or 152 meters). They are one of the most characteristic and annoying features of the tropical rain-forest.

STEPHANITE. The mineral stephanite, silver antimony sulfide, Ag_5SbS_4, is found in short prismatic or tabular orthorhombic crystals. It is a brittle mineral; hardness, 2–2.5; specific gravity, 6.25; metallic luster; color, black; streak, black; opaque.

Stephanite occurs associated with other silver minerals and is believed to be primary in character. Localities are in the Czech Republic and Slovakia, Saxony, the Harz Mountains, Sardinia; Cornwall, England; Chile and Mexico. In the United States it is found in Nevada, where it is an important silver ore. It was named for the Archduke Stephan of Austria, mining director of that country at the time this mineral was first described.

STEPPE. An extensive, treeless grassland area in southeastern Europe and Asia developing in the semiarid mid-latitudes of that region. They are generally considered drier than the prairie which develops in the subhumid mid-latitudes of the United States.

STEPPER MOTORS. The importance and usage of the electric stepper motor has increased immensely during the past two decades. Stepper motors, for example, are widely used in modern electronic typewriters, word processors, and other computerized products. They are important components of industrial robotic systems, large and small, and in other machines where repeatable positioning control is required. Stepper motors for most applications must be designed to work reliably over hundreds of thousands to millions of cycles.

The advantages of stepper motors are low cost, ruggedness, construction simplicity, high reliability, no maintenance, wide acceptance, no "tweaking" to stabilize, and no feedback components needed; they are inherently fail-safe and tolerate most environments. Steppers are simple to drive and control in an open-loop configuration. They provide excellent torque at low speed, up to 5 times the continuous torque of a brush motor of the same frame size or double the torque of the equivalent brushless motor. Frequently, a gearbox can be eliminated. A stepper-driven system is inherently stiff, with known limits to the dynamic position error.

There are three main types of stepper motors.

Permanent-Magnet Motors. The tin-can, or "canstack," motor shown in Fig. 1 is perhaps the most widely used type in commercial, nonindustrial applications. It is essentially a low-cost, low-torque, low-speed device ideally suited for use in computer peripherals, for example. The motor construction results in relatively large step angles, but the overall simplicity favors high-volume, low-cost production. The axial-air gap or disk motor is a variant of the permanent-magnet design, which achieves higher performance mainly because of its very low rotor inertia. This does restrict the applications of the motor to those situations involving little inertia, such as positioning the print wheel in a daisy-wheel printer.

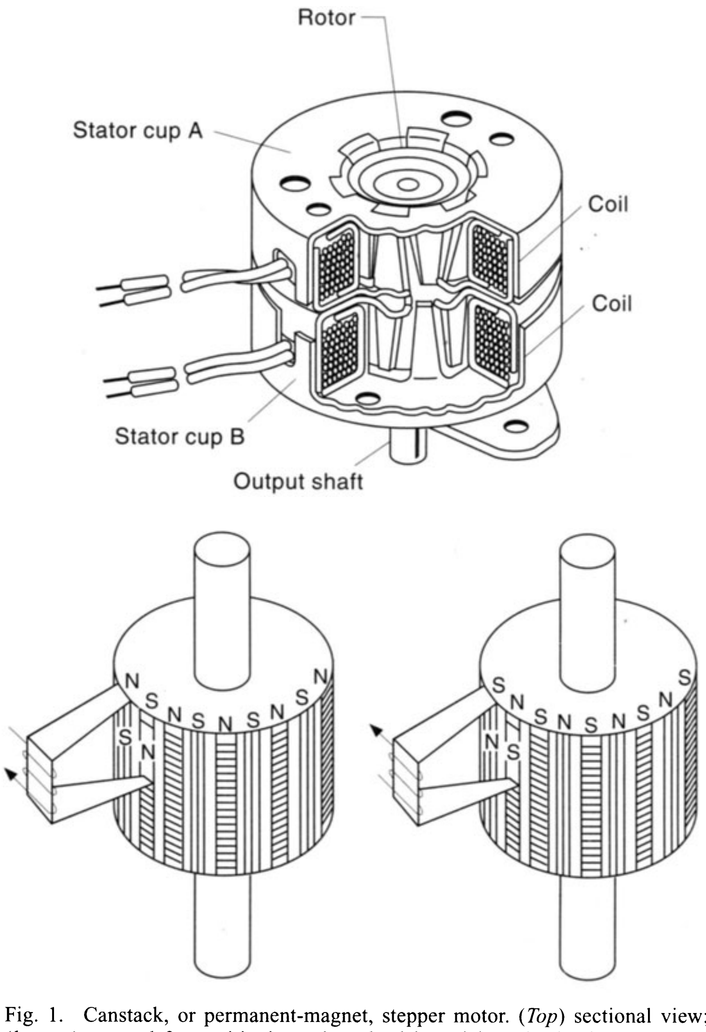


Fig. 1. Canstack, or permanent-magnet, stepper motor. (*Top*) sectional view; (*bottom*) as used for positioning print wheel in a daisy-wheel printer. (*Airpax Corp., U.S.*).

Disadvantages of the stepper motor include resonance effects, relatively long settling times, and rough performance unless a microstepper is used. Undetected position loss may result in open-loop systems. Steppers consume current regardless of load conditions, and they tend to run hot. Steppers tend to be noisy, especially when operated at high speeds. Some of the foregoing limitations can be overcome by use of a closed-loop system.

Variable-Reluctance Motors. There is no permanent magnet in a variable-reluctance motor. Thus the rotor spins freely without detent torque. Torque output for a given frame size is restricted, although the torque-to-inertia ratio is good. This type of motor is frequently used in small sizes for applications such as micropositioning tables. Variable-reluctance motors are seldom used in industrial applications. Having no permanent magnet, these motors are not sensitive to current polarity and thus require a different driving arrangement compared to other types (Fig. 2).

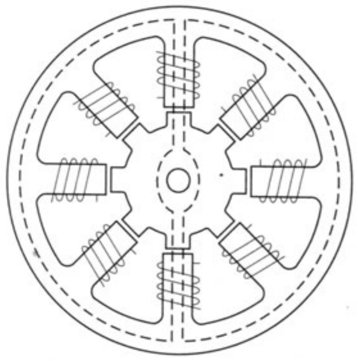


Fig. 2. Variable-reluctance motor.

Hybrid Stepper Motors. The hybrid motor is the most widely used stepper motor in industrial applications. Most hybrid motors are two-phase, although five-phase designs are available. A recent development is the enhanced hybrid, which uses flux-focusing magnets to give a significant improvement in performance, but at extra cost.

The rotor of the "model" hybrid stepper illustrated in Fig. 3 consists of two pole pieces with three teeth on each. Between the pole pieces is a permanent magnet that is magnetized along the axis of the motor, making one end a north pole, and the other a south pole. The teeth are offset at the north and south ends, as shown in the diagram.

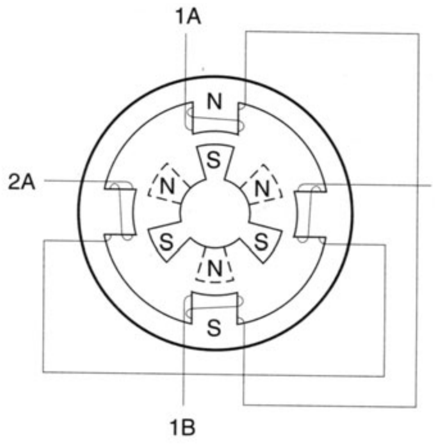


Fig. 3. Simple 12-step-per-revolution hybrid motor.

The stator consists of a shell having four teeth that run the full length of the rotor. Coils are wound on the stator teeth and are connected together in pairs.

With no current flowing in any of the motor windings, the rotor will tend to take up one of the five positions shown in the diagram. This is because the permanent magnet in the rotor attempts to minimize the reluctance, or magnetic resistance, of the flux path from one end to the other. This occurs when a pair of north- and south-pole rotor teeth are aligned with two of the stator poles. The torque, tending to hold the motor in one of these positions, is usually small and called the detent torque. The motor shown has 12 possible detent positions.

If current is passed through one pair of stator windings, as shown in Fig. 4(a), the resulting north and south stator poles will attract teeth of the opposite polarity on each end of the rotor. Thus there are only three stable positions for the rotor, the same as the number of rotor teeth. The torque required to deflect the rotor from its stable position is thus much greater and is referred to as the holding torque.

By changing the current flow from the first to the second set of stator windings [Fig. 4(b)], the stator field rotates through 90° and attracts a new pair of rotor poles. This results in the rotor turning through 30°, corresponding to one full step. Reverting to the first set of stator windings, but energizing them in the opposite direction, the stator field will be rotated through another 90° and the rotor will take another 30° step, as shown in Fig. 4(c). Finally, the second set of windings is energized in the opposite direction [Fig. 4(d)] to give a third step position. Returning to the first condition [Fig. 4(a)] and after these four steps, the rotor will have moved through one tooth pitch. This simple motor, therefore, performs 12 steps per revolution. Obviously, if the coils are energized in the reverse sequence, the motor will change direction.

If the two coils are energized simultaneously (Fig. 5), the rotor takes up an intermediate position, since it is equally attracted to two stator poles. Greater torque is produced under these conditions because all the stator poles are influencing the motor. The motor can be made to take a full step simply by reversing the current in one set of windings. This causes a 90° rotation of the stator field, as before. In fact, this would be the normal way of driving the motor in the full-step mode, always keeping two windings energized and reversing the current in each winding alternately.

By alternately energizing one winding and then two (Fig. 6), the rotor moves through only 15° at each stage, and the number of steps per revo-

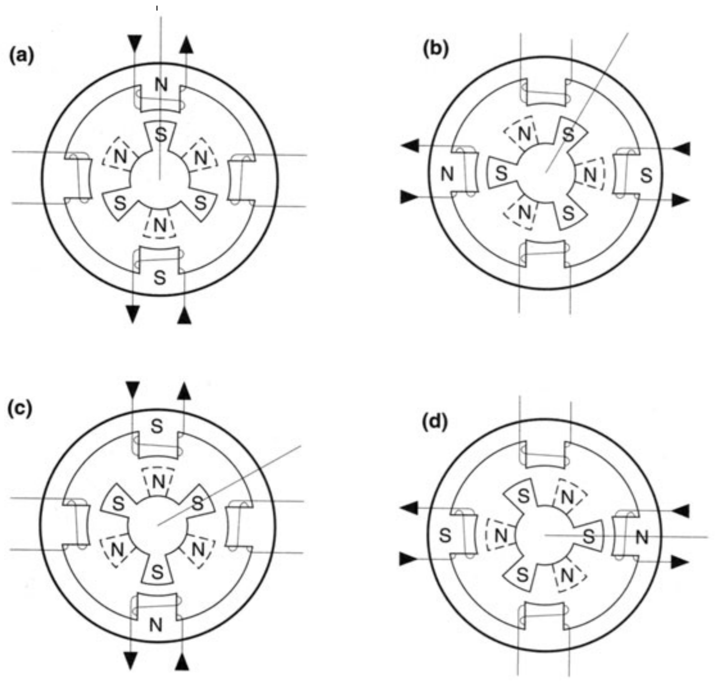


Fig. 4. Full stepping, one phase on. (*Parker-Hannifin Corp., Compumotor Div.*)

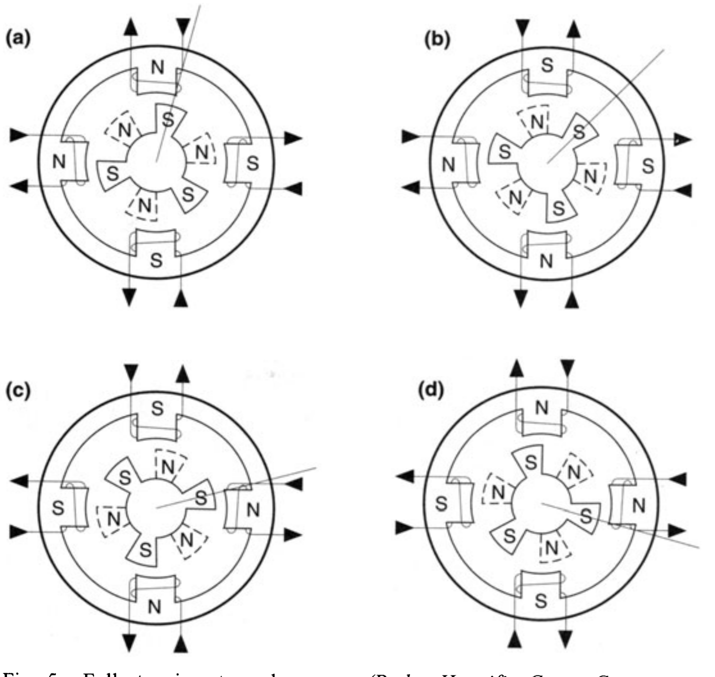


Fig. 5. Full stepping, two phases on. (*Parker-Hannifin Corp., Compumotor Div.*)

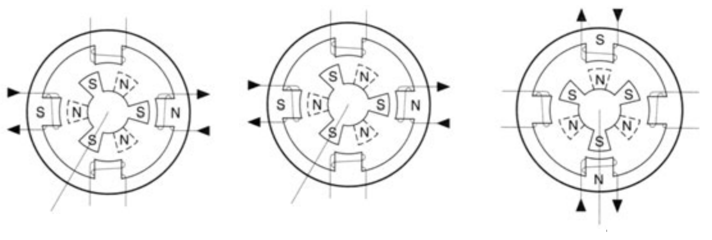


Fig. 6. Half stepping. (*Parker Hannifin Corp., Compumotor Div.*)

lution will be doubled. This is called half-stepping. Most industrial applications make use of this stepping mode. Although sometimes there is a slight loss of torque, this mode results in much better smoothness at low speeds, and less overshoot and ringing occur at the end of each step.

Current Patterns in Motor Winding. When the motor is driven in its full-step mode, energizing two windings, or phases, at a time (Fig. 7), the torque available on each step will be the same (subject to very small variations in the motor and drive characteristics). In the half-step mode, two phases are alternately energized and then only one, as shown in Fig. 8. Assuming the drive delivers the same winding current in each case, this will cause greater torque to be produced when there are two windings energized—that is, alternate steps will be strong and weak. Although the available torque obviously is limited by the weaker step, there is a improvement in low-speed smoothness over the full-step mode.

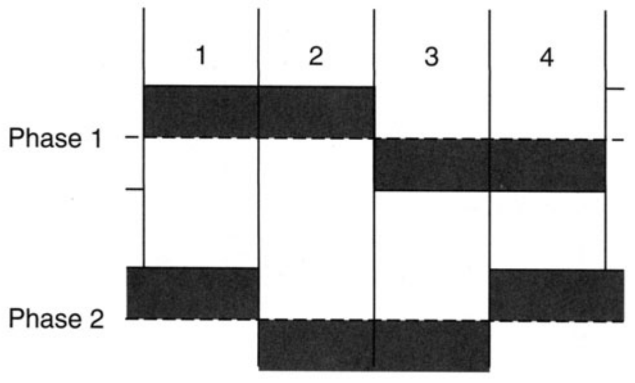


Fig. 7. Full step current, two phases on. (*Parker-Hannifin Corp., Compumotor Div.*)

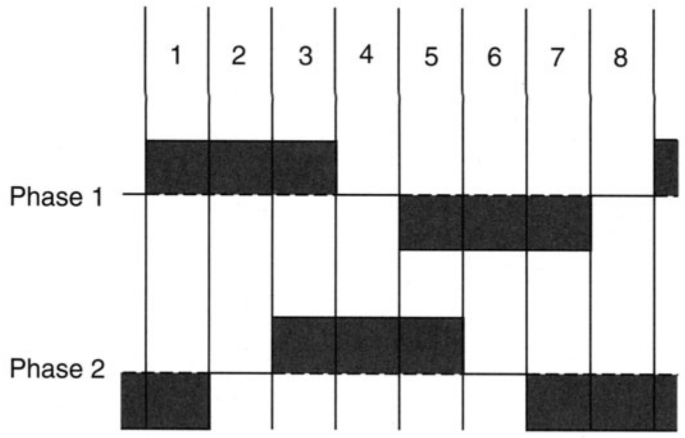


Fig. 8. Half-step current (*Parker-Hannifan Corp., Compumotor Div.*)

The motor designer would like to produce approximately equal torque on every step and to have this torque be at the level of the stronger step. This goal can be achieved by using a higher current level when there is only one winding energized. This does not overly dissipate the motor because the manufacturer's current rating will assume two phases to be energized. (The current rating is based on the allowable case temperature.) With only one phase energized, the same total

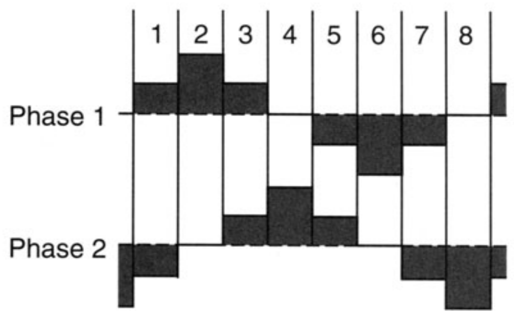


Fig. 9. Half-step current, profiled. (*Parker-Hannifin Corp., Compumotor Div.*)

power will be dissipated if the current is increased by 40%. Using the higher current in the one-phase-on state produces approximately equal torque on alternate steps, as indicated in Fig. 9.

Microstepping. It will be noted from the prior discussion that energizing both phases with equal currents produces an intermediate step position halfway between the one-phase-on positions. If the two phase currents are unequal, the rotor position will be shifted toward the stronger pole. This effect is utilized in the microstepping drive, which subdivides the basic motor step by proportioning the current in the two windings. In this way the step size is reduced and the low-speed smoothness is improved dramatically. High-resolution microstep drives divide the full motor step into as many as 500 microsteps, giving 100,000 steps per revolution. In this situation the current pattern in the windings closely resembles two sine waves with a 90° phase shift between them (Fig. 10). Thus the motor is driven very much as though it were a conventional ac synchronous motor. In fact, the stepper motor can be driven in this manner from a 60-Hz (U.S.) or 50-Hz (Europe) sine-wave source by including a capacitor in series with one phase. It will rotate at 60 r/min.

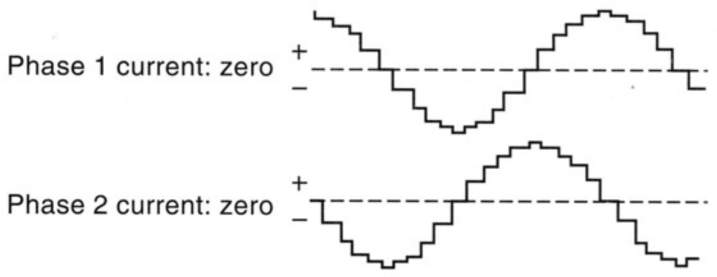


Fig. 10. Phase currents in microstep mode. (*Parker-Hannifin Corp., Compumotor Div.*)

Standard 200-Step Hybrid Motor. The standard stepper motor operates in the manner just described as a model, but it has a greater number of teeth on the rotor and stator, giving a smaller basic step size. The rotor is in two sections, as described previously, but has 50 teeth on each section. The half-tooth displacement between the two sections is retained. The stator has eight poles, each with five teeth, making a total of 40 teeth (Fig. 11).

Visualize that a tooth is placed in each of the gaps between the stator poles, in which case there would be a total of 48 teeth, two less than the number of rotor teeth. If rotor and stator teeth were aligned at 12 o'clock, they would also be aligned at 6 o'clock. But at 3 and 9 o'clock the teeth would be misaligned. However, due to the displacement between the sets of rotor teeth, alignment will occur at 3 o'clock and 9 o'clock at the other end of the rotor.

In practice, the windings are arranged in sets of four and wound such that diametrically opposite poles are the same. Thus, referring to Fig. 11, the north poles at 12 and 6 o'clock attract the south-pole teeth at the front of the rotor and the south poles at 3 and 9 o'clock attract the north-pole teeth at the back. By switching current to the second set of coils, the stator field pattern rotates through 45°, but to align with this new field, the rotor only has to turn through 1.8°. This is equivalent to one-quarter of a tooth pitch on the rotor, giving 200 full steps per revolution.

Note that there are as many detent positions as there are full steps per revolution, namely, 200. The detent positions correspond with rotor

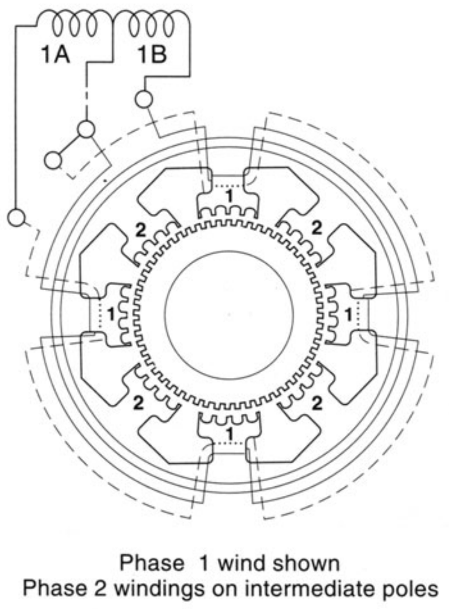


Fig. 11. 200-step hybrid motor. (*Parker-Hannifin, Compumotor Div.*)

teeth being fully aligned with stator teeth. When power is applied to a stepper drive, it is usual for it to energize in the zero-phase state in which there is current in both sets of windings. The resulting rotor position does not correspond with a natural detent position, so an unloaded motor will always move by at least one-half step at power on. Of course, if the system were turned off other than in the zero-phase state, or the motor is moved in the meantime, a greater movement may be seen at power-up.

For a given current pattern in the windings there are as many stable positions as there are rotor teeth (50 for a 200-step motor). If a motor is desynchronized, the resulting position error will always be a whole number of rotor teeth, or a multiple of 7.2°. A motor cannot "miss" individual steps. Position errors of one or two steps may be due to noise, spurious step pulses, or a controller fault.

Bifilar Windings. Most motors are described as being bifilar wound, which means there are two identical sets of windings on each pole. Two lengths of wire are wound together as though they were a single coil. This produces two windings that are electrically and magnetically almost identical. If one coil were wound on top of the other, even with the same number of turns, the magnetic characteristics would be different.

The origin of the bifilar winding goes back to the unipolar drive. Rather than reversing the current in one winding, the field may be reversed by transferring current to a second coil wound in the opposite direction. (Although the two coils are wound the same way, interchanging the ends has the same effect.) Thus, with a bifilar-wound motor, the drive can be kept simple. However, this requirement has now largely disappeared with the widespread availability of the more efficient bipolar drive. Nevertheless, the two sets of windings do provide additional flexibility.

If all the coils in a bifilar-wound motor are brought out separately, there will be a total of eight leads (Fig. 12). This is becoming the most common configuration, since it gives the greatest flexibility. However, there are still a number of motors produced with only six leads, one lead serving as a common connection to each winding in a bifilar pair. This arrangement limits the range of applications of the motor, since the windings cannot be connected in parallel. Some motors are made with only four leads. These are not bifilar-wound and cannot be used with a unipolar drive. There is obviously no alternative connection method with a four-lead motor, but in many applications this is not a drawback and the problem of insulating unused leads is avoided. Occasionally a five-lead motor may be encountered. These should be avoided inasmuch as they cannot be used with conventional bipolar drives requiring electrical isolation between the phases.

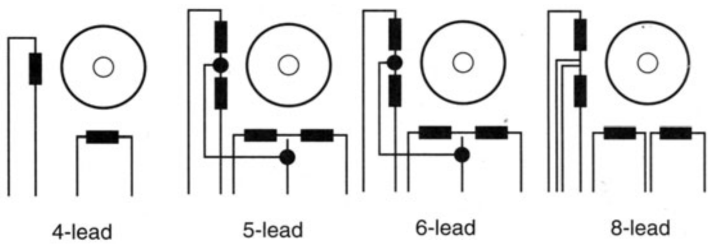


Fig. 12. Motor lead configurations.

Linear Stepper Motors

The linear stepper is essentially a conventional rotary stepper that has been "unwrapped" so that it operates in a straight line. The moving component is referred to as the forcer, and it travels along a fixed element, or platen. For operational purposes the platen is equivalent to the rotor in a normal stepper, although it is an entirely passive device and has no permanent magnet. The magnet is incorporated in the moving forcer together with the coils (Fig. 13).

The forcer is equipped with four pole pieces, each having three teeth. The teeth are staggered in pitch with respect to those on the platens so that switching the current in the coils will bring the next set of teeth into

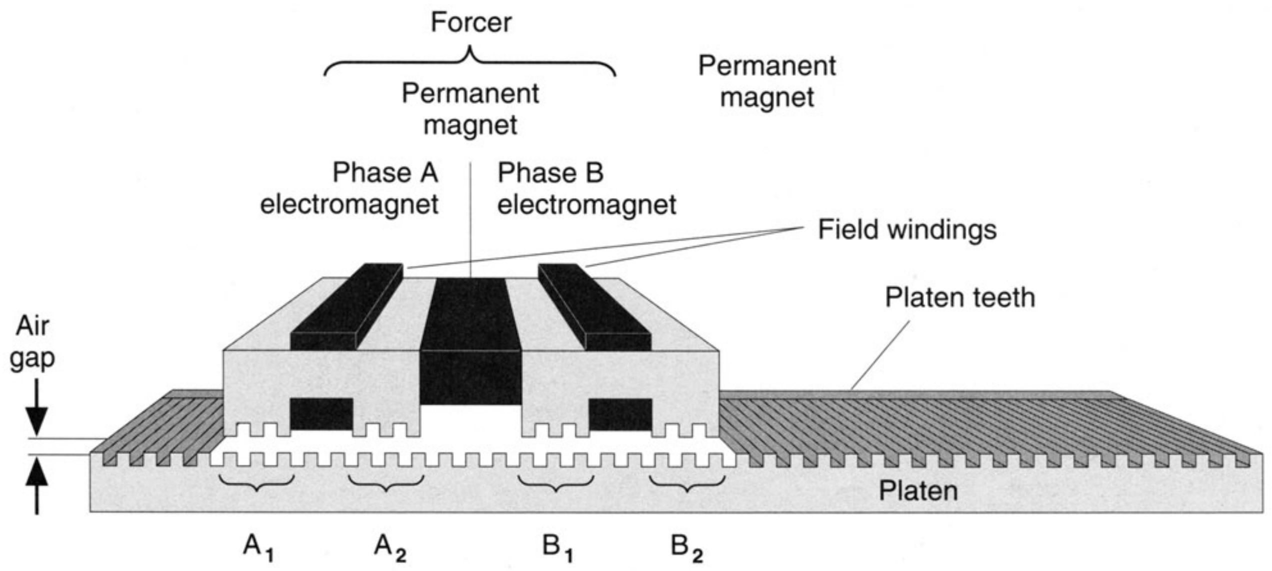


Fig. 13. Principle of linear stepping motor.

alignment. A complete switching cycle (four full steps) is equivalent to one tooth pitch on the platen. Like the rotary stepper, the linear motor can be driven from a microstep drive. In this case a typical linear resolution will be 12,500 steps per inch (4921 steps/cm).

The linear motor finds favor in applications involving a low mass to be moved at very high speed. In a lead-screw-driven system the predominant inertia usually is the lead screw rather than the load to be moved. Hence most of the motor torque goes to accelerate the lead screw, and this problem becomes more severe the longer the travel required. In using a linear motor, all the developed force is applied directly to the load and the performance achieved is independent of the length of the move. A screw-driven system can develop greater linear force and better stiffness. However, the maximum speed may be as much as 10 times higher with the equivalent linear motor.

With further reference to Fig. 13, the forcer consists of two electromagnets *A* and *B* and a strong rare-earth permanent magnet. The two pole faces of each electromagnet are toothed to concentrate the magnetic flux. Four sets of teeth on the forcer are spaced in quadrature so that only one set at a time can be aligned with the platen teeth.

The magnetic flux passing between the forcer and the platen gives rise to a very strong force of attraction between the two pieces. The attractive force can be up to 10 times the peak holding force of the motor, requiring a bearing arrangement to maintain precise clearance between the pole faces and the platen teeth. Either mechanical roller bearings or air bearings are used to maintain the required clearance.

When current is established in a field winding, the resulting magnetic field tends to reinforce permanent magnetic flux at one pole face and cancel it at the other. By reversing the current, the reinforcement and cancellation are exchanged. Removing current divides the permanent magnetic flux equally between the pole faces. By selectively applying current to phases *A* and *B* it is possible to concentrate the flux at any of the forcer's four pole faces. The face receiving the highest flux concentration will attempt to align its teeth with the platen. Figure 14 shows the four primary states or full steps of the forcer. The four steps result in motion of one tooth interval to the right. Reversing the sequence moves the forcer to the left.

Repeating the sequence in the example will cause the forcer to continue its movement. When the sequence is stopped, the forcer stops, with the appropriate tooth set aligned. At rest, the forcer develops a holding force that opposes any attempt to displace it. As the resting motor is displaced from equilibrium, the restoring force increases until the displacement reaches one-quarter of a tooth interval (Fig. 15). Beyond this point the restoring force drops. If the motor is pushed over the crest of its holding force, it slips or jumps rather sharply and comes to rest at an integral number of tooth intervals away from its original location. If this occurs while the forcer is traveling along the platen, it is referred to as a stall condition.

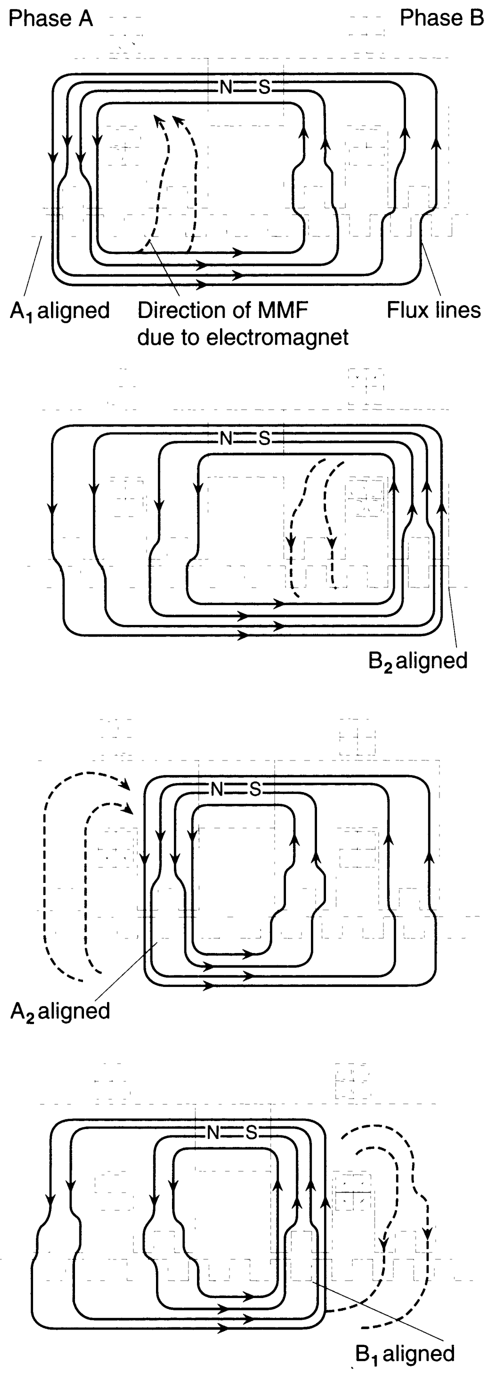


Fig. 14. Four cardinal states of full steps of the force. (*Parker-Hannifin Corp., Compumotor Div.*)

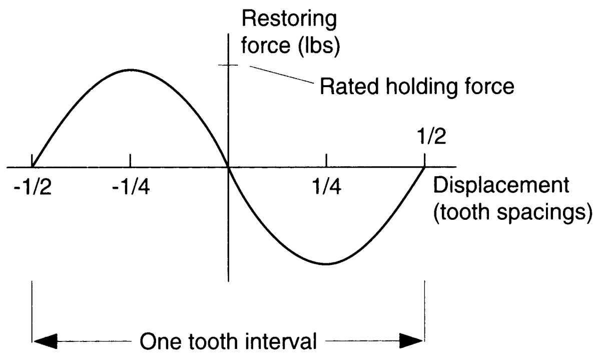


Fig. 15. Force versus displacement (linear stepper). (*Parker-Hannifin Corp., Compumotor Div.*)

Linear Step Motor Characteristics. These include velocity ripple, platen mounting, environment, life expectancy, yaw (plus pitch and roll), and accuracy. To summarize, the worstcase accuracy of a linear step motor can be given by

$$\text{Accuracy} = A + B + C + D + E + F$$

where A = cyclic error due to motor magnetics, which recurs once every pole pitch as measured on motor body
B = unidirectional repeatability—error measured by repeated moves to the same point from different distances in the same direction
C = hysteresis—backlash of motor when changing direction due to magnetic nonlinearity and mechanical friction
D = cumulative platen error—linear error of platen as measured on motor body
E = random platen error—nonlinear errors remaining in platen after linear error is disregarded
F = thermal expansion error—error caused by change in temperature, expanding or contracting the platen

STEP RESPONSE TIME. Of a system or an element, the time required for an output to make the change from an initial value to a large specified percentage of the final steady-state value either before or in the absence of overshoot, as a result of a step change to the input. See figure that accompanies entry on **Response (Instrument).** This time is usually stated for 90, 95 or 99% change.

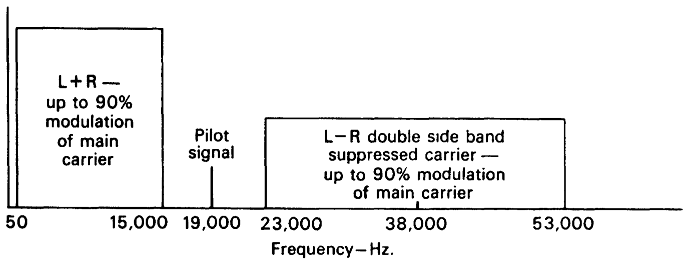
Time Constant. The value T in an exponential response term $A \exp(-t/T)$ or in one of the transform factors $1 + sT$, $1 + j\omega T$, $1/(1 + sT)$, $1/(1 + j\omega T)$, where s = complex variable; t = time, seconds; T = time constant; $j = \sqrt{-1}$; ω = frequency, radian/second.

For the output of a first-order (lag or lead) system forced by a step or an impulse, T is the time required to complete 63.2% of the total rise or decay; at any instant during the process, T is the quotient of the instantaneous rate of change divided into the change to be completed. In higher-order systems, there is a time constant for each of the first-order components of the process. In a Bode diagram, breakpoints occur at $\omega = 1/T$.

STEPTOE. A hill or mountain whose top projects above a lava flow which has surrounded its lower flanks.

STEREO BROADCASTING. A system of radio broadcasting using frequency modulation in which the modulating signal contains information obtained from a stereo microphone which following its modulation, transmission, and reception can subsequently be reproduced by a stereo loudspeaker. The complete communication process involves generating electric signals which correspond to variations in sound intensity at two different points in the physical location where the program originates, transmitting a combination of these two signals by means of a frequency modulated radio transmitter, separating the signals in the listener's receiver and supplying the components to the loudspeaker system.

In the United States, transmission is in accordance with standards established by the Federal Communications Commission. The electrical signals obtained from the two microphones are arbitrarily designated Left (L) and Right (R) signals. They are combined to produce a sum signal (L + R) and difference signal (L − R). According to the standards, each of the audio signals may contain frequency components in the range of 50–15,000 Hz. A frequency translation of the L − R signal spectrum is effected by amplitude modulation on a subcarrier of 38 kHz which is then suppressed. To the combination of the L + R spectrum and the translated L − R spectrum is added a 19 kHz pilot signal, which is to be used to aid in the restoration of the L and R signals after demodulation of the transmitted signal in the listener's receiver. The distribution of the composite modulating signal, which is used for the frequency modulation of the transmitter, is depicted in the figure. Radio transmission takes place in the frequency range 88.1 to 107.9 Hz.



Frequency distribution of composite modulating signal for stereo broadcasting.

Upon reception, after demodulation which recovers the composite modulating signal, the signal requires processing to recover the L and R audio signals which were originally developed by the two microphones. The inverse of the frequency translation which occurred prior to the frequency modulation is now performed on the L − R signal. The composite signal is passed through a bandpass filter which passes only frequencies in the range from 23 to 53 kHz. The 19 kHz pilot signal is also extracted using a filter and is then passed through a doubler-amplifier circuit from which a signal of 38 kHz is obtained. This signal and the output of the bandpass filter are applied to a detector in which *heterodyne* action occurs. The resultant output from the detector has the waveform of the L − R audio signal at the transmitter, together with some extraneous signals which were introduced by the modulation and demodulation processes. The L + R and L − R signals are now fed to a circuit which separates them into the L and R signals in a manner inverse to that in which they were obtained at the transmitter. After further amplification, the L and R signals are furnished to an appropriate stereophonic loudspeaker system. It should be noted that the L + R signal not only provides one portion of the stereo signal but also serves as the compatible monophonic signal for those listeners whose receivers are not equipped for stereo reception and reproduction.

See also **Radio Communication.**

STEREOGRAPHIC PROJECTION. 1. The stereographic projection of a sphere on a plane is defined as follows: For a given point P, called the pole, on the surface of a sphere S, and for a given plane M not passing through P, and perpendicular to a diameter through P, the line joining P with a variable point Q and M intersects S in a second point R. This mapping of the points R of the sphere S on the points Q of M is called a stereographic projection of S on M.

2. This type of map projection is used to some extent by navigators, but more commonly for maps of the celestial sphere, particularly in constructing the basic map for star finders. A stereographic projection has a valuable property in that circles on the sphere appear as circles on the projected map. Because of this property, several attempts have been made to use the projection for drawing lines of position obtained from celestial objects. In the proposed methods, the subastral point would be plotted on the chart and the circle of position drawn about this point with a radius equal to the observed zenith distance of the object. This method is proposed for use by aviators, where speed in drawing lines of position and obtaining a fix is particularly desirable, and where extreme accuracy is never possible because of errors inherent in measuring altitude with a bubble sextant in a moving plane.

STEREO-POWER. For prism binoculars or similar stereo systems, the ratio of the distance between the objective axes to the distance between eyepiece axes multiplied by the magnifying power. A measure of the stereoscopic radius.

STEREOSCOPE. The sensation of depth in an object is due to binocular vision; that is, to the fact that two eyes do not each see exactly the same view. By taking two pictures with a camera moved a few inches—or with a double stereoscopic camera—two slightly different pictures are obtained. A stereoscope is a device by which each eye sees only one of these pictures, and the same sensation of depth is obtained as with direct binocular vision.

STEREOSPECTROGRAM. A method of representing spectral data in which the three variables, concentration of solute, optical density, and wavelength of light, are plotted in three dimensions to produce a three-dimensional figure; or else in two dimensions by choosing an oblique axis in addition to the customary x-axis and y-axis.

STEREO SYSTEM. An acoustical system in which a plurality of microphones (or other transducers), transmission channels and reproducers are arranged so as to provide a sensation of spatial distribution of the original sound sources to the listener.

STERLING SILVER. Silver alloy, usually with copper, containing at least 92.5% silver.

STERN-GERLACH EXPERIMENT. An experimental test by O. Stern and W. Gerlach (Germany, 1924) of the magnetic moment of atoms. A stream of metallic atoms, issuing from a vaporizing furnace through a narrow slit, entered a strong magnetic field. The magnetic intensity was perpendicular to the atom stream, and had a strong gradient in its own direction. If magnetic moments of atoms are due to revolving electrons, the atoms should, according to classical theory, begin to precess at all angles about the field direction, and the atomic beam should simply broaden into a band. According to the quantum theory,

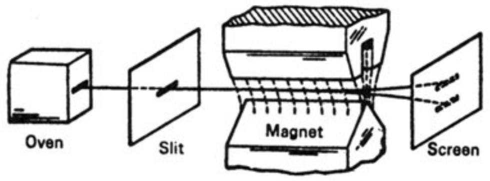
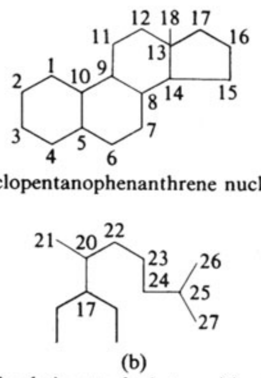


Diagram of Stern-Gerlach experiment for determining magnetic moments of gaseous atoms. Beam of vaporized metal is split by strong inhomogeneous magnetic field.

they should precess at certain angles only, and the original stream should be divided into distinct streams. Since the beam was split into $2J + 1$ different beams, the experiment showed that in a magnetic field not all orientations to the field, but only $2J + 1$ discrete directions, are possible. See **Magnetism; Precession.**



Cyclopentanophenanthrene nucleus.

(b)
Side chain attached at position 17.

Fig. 1. Steroid molecule nucleus and side chain.

STEROIDS. Organic compounds characterized from a structural standpoint by the cyclopentanophenanthrene nucleus as shown in Fig. 1. Biochemically, the steroids are closely related to the terpenes. Steroids occur widely in nature, both in animals and plants. Many steroids are hormones, such as estrogens and cortisone, which are produced by the body's endocrine system and which are of great importance in the regulation of numerous body processes, such as growth and metabolism. Similarly, steroid hormones are important to several physiological processes within plants. Auxin, for example, is a plant growth hormone that regulates longitudinal cell structure so as to permit bending of the stalk or stem in phototropic response. The most common animal steroid, a steroid alcohol (or *sterol*) is *cholesterol*, which is the precursor of bile acids, steroid hormones, and provitamin D3. It should be stressed, however, that not all steroids are hormones; and not all hormones are steroids.

Over the last 35 to 40 years, steroid therapy, that is, the medical augmentation of steroid hormones that are insufficient in the body (as well as treating diseases and elements which result from an overabundance of certain steroid hormones in the body), is one of the major chapters in the history of medical progress. Similarly, the understanding of steroid chemistry has contributed markedly to plant biology, notably to plant breeding and the development of plant growth regulators. Commencing with the isolation of steroid hormones from natural sources, techniques were later developed to synthesize a number of hormones. Then, a further step involved the application of steroid hormones, developed by way of synthesis, but for which there are no known counterparts in nature. Some of the steroid hormones, including steroidal synthetics, of importance medically and to scientific investigations, are listed and described briefly in the accompanying table.

The early history of steroid chemistry commenced with the observations of Mauthner, Windlaus, Wieland, Jacobs, Diels, and other organic chemists and biologists in their early observations on the products of oxidation, aromatization, and other reactions of cholesterol, bile acids, and plant glycosides. The interrelationship between sterols and bile acids was recognized early during these investigations. Shortly after the steroid character of the female and male sex hormones had been established, x-ray crystallography demonstrated errors in early theories concerning molecular structure. Rosenheim and King, working with monomolecular layers and x-ray evidence gained by Bernal, formulated the concept of the cyclopentanophenanthrene nucleus.

REPRESENTATIVE STEROID HORMONES AND STEROIDAL SYNTHETICS[1]

Antiinflammatory, Antiallergic, and Antirheumatic Agents (Adrenal Corticosteroids)

Betamethasone (9-fluoro-16β-methylprednisolone; 16β-methyl-11β,17α,21-trihydroxy-9α-fluoro-1,4-pregnadiene-3,20-dione). $C_{22}H_{29}O_5F$, mw = 329.5. Also, the *betamethasone acetate*, $C_{24}H_{31}O_6F$, mw = 434.5; and *betamethasone disodium phosphate*, $C_{22}H_{28}O_8FNa_2P$, mw = 516.4. Both of the latter compounds are used for treating carpal tunnel syndrome, the most common of the entrapment neuropathies. The median nerve is subjected to compression and possibly ischemia in the confined space between the carpal bones and the flexor retinaculum of the wrist.

Chloroprednisone acetate (6α-chloroprednisone acetate; 6α-chloro-$\Delta^{1,4}$-pregnidien-17β,21-diol-3,11,20-trione-21-acetate). Mw = 436.6. Multiple uses.

Corticosterone (11,21-dihydroxyprogesterone; Δ^4-pregnene-11β,21-diol-3,20-dione; 11β,21-dihydroxy-4-pregnene-3,20-dione). $C_{21}H_{30}O_4$, mw = 346.4. Multiple uses.

Cortisone (17-hydroxy-11-dehydrocorticosterone; 17α,21-dihydroxy-4-pregnene-3,11,20-trione; Δ^4-pregnene-17α,21-diol-3,11,20- trione; Kendall compound; Wintersteiner compound F). $C_{21}H_{28}O_5$, mw = 360.4. Multiple uses.

Desoxycorticosterone (deoxycorticosterone; 11-desoxycorticosterone; 21-hydroxyprogesterone; 4-pregnen-21-ol-3,20-dione; Kendall desoxy compound B; Reichstein substance Q). $C_{21}H_{30}O_3$, mw = 330.2. Also, the *desoxycorticosterone acetate* (DCA). $C_{23}H_{32}O_4$, mw = 372.4; and *desoxycorticosterone pivalate*, $C_{26}H_{38}O_4$, mw = 414.6. Multiple uses.

Dexamethasone (hexadecadrol; 9α-fluoro-16α-21-trihydroxy-16α-methyl-1,4-pregnadiene-3,20-dione). $C_{22}H_{29}FO_5$, mw = 392.4. Widely used in the treatment of benign intracranial hypertension, brain abscess, brain metastases, brain tumor, cerebral thrombosis, Cushing's syndrome, encephalitis, hypertensive encephalopathy, lumbar disk disease, meningococcal cerebral edema, shock, superior vena cava obstruction in cancer patients, ulcerative colitis.

Dichlorisone acetate (9α,11β-dichloro-1,4-pregnadiene-17α,21-diol-3,20-dione-21-acetate). $C_{23}H_{28}O_5Cl$, mw = 455.3.

Fluocinolone acetonide (6α,9α-difluoro-16α hydroxyprednisolone-16,17-acetonide). $C_{24}H_{30}O_6F_2$, mw = 452.50.

Fluorohydrocortisone (fluorocortisone; 9α-fluoro-11β,17α,21-trihydroxy-4-pregnene-3,20-dione). $C_{21}H_{29}O_5$, mw = 380.4. Used in treating Shy-Drager syndrome (parenchymatous degeneration of the central nervous system); also in treating orthostatic hypotension (a cause of temporary loss of consciousness when a person rises to an erect position). Also *fluorometholone* (9α-fluoro-11β,17α-dihydroxy-6α-methyl-1,4-pregnadiene-3,20-dione). $C_{22}H_{24}FO_4$, mw = 376.4; and *fluprednisolone* (6α-fluoroprednisolone), $C_{21}H_{27}FO_3$, mw = 378.4; and *flurandrenolone* (6-fluoro-16α-hydroxyhydrocortisone-16,17-acetonide), $C_{24}H_{33}O_6F$, mw = 436.5.

Hydrocortisone (cortisol; 11β,17α,21-trihydroxy-4-pregnene-3,20-dione), $C_{21}H_{30}O_5$, mw = 362.5. Used in treating adrenal insufficiency, notably in cancer patients, contact dermatitis, panhypopituitarism, psoriasis, shock, and urticaria. Also *hydrocortisone acetate* (cortisol acetate), $C_{23}H_{32}O_6$, mw = 404.5, used in treating rheumatoid arthritis; and *hydrocortisone sodium succinate* (11β,17α,21-trihydroxy-4-pregnene-3,20-dione-21-hydrogen succinate, sodium salt), $C_{25}H_{33}O_8Na$, mw = 484.5. The latter compound is used in treating ulcerative colitis.

Methylprednisolone (Δ^1-6α-methylhydrocortisone). $C_{22}H_{30}O_5$, mw = 374.5. Used in treating thrombocytopenia with intracranial hemorrhage, Gram-negative bacteremia, posttransfusion purpura, and shock. Also *methylprednisolone sodium succinate, $C_{26}H_{33}O_8$*Na, mw = 496.5.

Paramethasone (6α-fluoro-16α-methylprednisolone). $C_{22}H_{30}O_5$, mw = 392.45. Also *paramethasone acetate*, $C_{24}H_{31}O_6F$, mw = 434.5.

(continued)

REPRESENTATIVE STEROID HORMONES AND STEROIDAL SYNTHETICS *(continued)*

Antiinflammatory, Antiallergic, and Antirheumatic Agents (Adrenal Corticosteroids)

Prednisolone (methacortandralone; 1,4-pregnadiene-3,20-dione-11β,17α,21-triol). $C_{21}H_{28}O_5$, mw = 360.4. Also *prednisolone phosphate sodium* (disodium prednisolone-21-phosphate), $C_{21}H_{27}Na_2O_8P$, mw = 484.4, used in treating ulcerative colitis. Also *prednisolone pivalate* (prednisolone trimethylacetate), $C_{26}H_{36}O_6$, mw = 444.6.

Prednisone (metacortandricin; 17α,21-dihydroxy-1,4-pregnadiene-3,11,20-trione). $C_{21}H_{26}O_5$, mw = 358.4. Used in the treatment of scores of ailments and diseases. To mention a few: acute erythroleukemia, acute gouty arthritis, acute pericarditis, aspiration pneumonitis, autoimmune hemolytic anemia, breast cancer, bronchial asthma, chronic hepatitis, dermatomyositis, desquamative interstitial pneumonia, Hodgkin's disease, hypercalcemia, immune neutropenia, lymphocytic leukemia, osteoporosis, hemoglobinuria, prostate cancer, psoriasis, radiation enteritis, rheumatoid arthritis, trichinosis, ulcerative colitis, usual interstitial pneumonia, viral anthropathies.

Triamcinolone (9α-fluoro-16α-hydroxyprednisolone). $C_{21}H_{27}FO_6$, mw = 394.4. Used in treating acute gouty arthritis and uremic pericarditis. Also *triamcinolone acetonide* (9α-fluoro-11β,21-dihydroxy-16α,17α-isopropylidenedioxy-1,4-pregnadiene-3,20-dione), $C_{24}H_{31}FO_6$, mw = 434.4. Used in treating acne vulgaris. Also *triamcinolone diacetate* (9α-fluoro-16α-hydroxyprednisolone-16,21-diacetate), $C_{25}H_{31}FO_8$, mw = 478.49.

Androgens and Anabolic Agents

Androsterone (3α-hydroxy-17-androstenone). $C_{19}H_{30}2$, mw = 290.4. Also *fluoxymesterone* (9α-fluoro-11β, 17β-dihydroxy-17α-methyl-4-androsten-3-one), $C_{20}H_{29}FO_3$, mw = 336.4. Used in treating paroxysmal nocturnal hemoglobinuria. Also *aldosterone* (electrocortin; 18-formyl-11β,21-dihydroxy-4-pregnene-3,20-dione) $C_{21}H_{28}O_5$, mw = 360.4.

Hydroxydione sodium (21-hydroxypregnane-3,20-dione-21-sodium hemisuccinate). $C_{25}H_{35}O_6Na$, mw = 454.5.

Spironolactone (3-(30-oxo-7α-acethylthio-17β-hydroxy-4-androsten-17α-yl)-propionic acid γ-lactone), $C_{24}H_{32}O_4S$, mw = 416.5. Used in treating congestive heart failure, hypertension, hypokalemia.

Methandrostenolone (17α-methyl-17β-hydroxy-1,4-androstadien-3-one). $C_{20}H_{28}O_2$, mw = 300.4.

Methylandrostenediol (MAD; methandriol; 17α-methyl-5-androsten-3β,17β-diol). $C_{20}H_{32}O_2$, mw = 304.4.

Methyl testosterone (17α-methyl-Δ^4-androsten-17-β-0 1-3-one). $C_{20}H_{30}O_2$, mw = 302.4.

Norethandrolone (17α-ethyl-19-nortestosterone). $C_{20}H_{30}O_2$, mw = 302.4. Also *oxandroline* (17β-hydroxy-17α-methyl-2-oxa-5α-androstane-3-one), $C_{19}H_{30}O_3$, mw = 306.4.

Oxymetholone (2-hydroxymethylene-17-α-methyldihydrotestosterone). $C_{21}H_{32}O_3$, mw = 332.4. Used in treating agnogenic myeloid metaplasia and hereditary angioedema. Also *promethologne* (2α-methyl-dihydro-testosterone propionate), mw = 360.5.

Testosterone (trans-testosterone; 17β-hydroxy-4-androsten-3-one). $C_{19}H_{28}O_2$, mw = 288.4. Used in treating acne vulgaris, impotence, polycystic ovary syndrome, male hypogonadism. Also *testosterone cypionate*, $C_{27}H_{40}O_3$, mw = 412.6; *testosterone enanthate*, $C_{26}H_{40}O_3$, mw = 400.6; *testosterone phenylacetate*, $C_{27}H_{34}O_3$, mw = 406.5; *testosterone propionate*, $C_{22}H_{32}O_3$, mw = 344.4.

Estrogens

Equilenin (1,3,5-10,6,8-estrapentaen-3-ol-17-one). $C_{18}H_{18}O_2$, mw = 266.3. Also *equilin* (1,3,5,7-estratetraen-3-ol-17-one), $C_{18}H_{20}O_2$, mw = 268.3.

Estradiol (β-estradiol; dihydrofolliculin, dihydroxyestrin; 3,17-ephidhydroxyestratriene). $C_{18}H_{24}O_2$, mw = 272.3. Also *estradiol benzoate*, $C_{25}H_{28}O_3$, mw = 376.4; *estradiol cypionate*, $C_{26}H_{36}O_2$, mw = 396.6; *estradiol diprionate*, $C_{24}H_{32}O_4$, mw = 384.5.

Estriol (trihydroxyestrin; 1,3,5-estratriene-3,16α,17β-triol). $C_{18}H_{24}O_3$, mw = 288.3.

Estrone (folliculin; ketohydroxyestrin; 1,3,5-estratriene-3-ol-17-one). $C_{18}H_{22}O_2$, mw = 270.3. Also *estrone benzoate*, $C_{25}H_{26}O_3$, mw = 374.4.

Ethynyl estradiol (17-ethinyl estradiol; 17α-ethynyl-1,3,5-estratriene-3,17β-diol). $C_{20}H_{24}O_2$, mw = 296.4. Used to treat acne vulgaris, osteoporosis.

Mestranol (ethylestradiol-3-methylether; 3-methoxy-19-nor-17α-pregna-1,3,5-trien-20-yn-17-ol). $C_{21}H_{26}O_2$, mw = 310.4. Used in treating acne vulgaris.

Progestogens and Progestins

Acetoxypregnenolone (21-acetoxypregnenolone; 3-hydroxy-21-acetoxy-5-pregnen-20-one). $C_{23}H_{34}O_4$, mw = 374.5.

Anagestone acetate (6α-methyl-4-pregnen-17α-ol-20-one). $C_{24}H_{36}O_3$, mw = 372.6.

Chlormadinone acetate (6-chloro-$\Delta^{4,6}$-pregnadiene-17α-ol-3,20-dione acetate). $C_{23}H_{29}ClO_4$, mw = 414.9.

Dimethisterone (17β-hydroxy-6α-methyl-17α-(prop-1-nyl)-androst-4-ene-3-one). $C_{23}H_{32}O_2 \cdot H_2O$, mw = 358.5. Also *ethisterone*, $C_{21}H_{28}O_2$, mw = 312.4.

Ethynodiol diacetate (19-nor-17α-pregn-4-en-20-yne-3β,17-diol diacetate). $C_{24}H_{32}O_4$, mw = 384.5. Used in treating acne vulgaris.

Flurogestone acetate (17α-acetoxy-9α-fluoro-11β-hydroxy-4-pregnene-3,20-dione). $C_{23}H_{31}O_5F$, mw = 406.5.

Hydroxymethylprogesterone (medroxyprogesterone; 17α-hydroxy-6α-methyl-4-pregnene-3,20-dione). $C_{22}H_{23}O_3$, mw = 344.5. Used for treating menopausal symptoms, secondary amenorrhea. Also *hydroxymethylprogesterone acetate*, $C_{24}H_{24}O_3$, mw = 386.5. Used in treating hypogonadal females. Also *hydroxyprogesterone* (4-pregen-17α-ol-3,20-dione), $C_{21}H_{30}O_3$, mw = 330.4. Also *hydroxyprogesterone caproate* (17α-hydroxy-4-pregnene-3,20-dione caproate), $C_{27}H_{40}O_4$, mw = 428.6.

Melengestrol acetate (MGA; 6-dehydro-17-hydroxy-6-methyl-16-methylene-progesterone acetate), $C_{25}H_{32}O_4$, mw = 396.51.

Norethindrone (norethisterone; 17α-ethynyl-17-hydroxy-19-nor-17α-4-en-20-yn-3-one). $C_{20}H_{26}O_2$, mw = 298.4. Used in treating acne vulgaris. Also *norethindrone acetate*, $C_{22}H_{28}O_3$, mw = 340.4. Also norethynodrel, $C_{20}H_{26}O_2$, mw = 298.4. Also *normethisterone*, $C_{19}H_{28}O_2$, mw = 288.4.

Pregnenolone (Δ^5-pregnen-3β-ol-20-one). $C_{21}H_{32}O_2$, mw = 308.4. Important in the synthesis of adrenal hormones.

Progesterone (progestin; progestone; Δ^4-pregnene-3,20-dione). $C_{21}H_{30}O_2$, mw = 314.4. Used in treating excessive uterine bleeding, hypogonadal females, menopausal symptoms, polycystic ovary syndrome, secondary amenorrhea.

Diuretic, Antiduretic and Anesthetic Agents

Aldosterone and spironolactone are described earlier in this list.

Hydroxydione sodium (21-hydroxypregnane-3,20-dione-21-sodium hemisuccinate). $C_{25}H_{35}O_6Na$, mw = 454.5

[1]This is an abridged list of steroid hormones. Some are much more important and widely used than others. Some are relatively recent to steroid therapy; others have been used more widely in the past than presently. See also **Hormones.**

Once the skeleton of the steroids was established, there remained the task of understanding the steric relationships of the molecules. With reference to Fig. 1, it will be noted that there are nine asymmetric carbon atoms in the steroid skeleton—C_5, C_{10}, C_9, C_8, C_{14}, and C_{13} in the ring system. There are also two asymmetric carbons in the side chain attached at C_{17}. These are C_{20} and C_{25}. With reference to Figs. 1 and 2, the relative configuration of C_5 and C_{10}, of C_9 and C_8, and of C_{14} and C_{13}, determines whether the junctions between rings a/b, b/c, and c/d, respectively, are *trans* or *cis*. According to an arbitrary convention, one designates the substituent groups α or β depending upon whether they are situated below the plane of the molecule, when depicted in a certain way. Usually, in a structural diagram of a steroid molecule, a dotted line connection will be used between atoms to designate an alpha position; a regular solid line for a beta position.

Figure 2 illustrates one of the two most important configurations of steroid skeletons, among 63 other possibilities, as they are found in nature. The side chain is usually attached in β-position to C_{17}. The configurations on C_{20} and C_{24} have likewise been determined and are known to produce steric isomerisms.

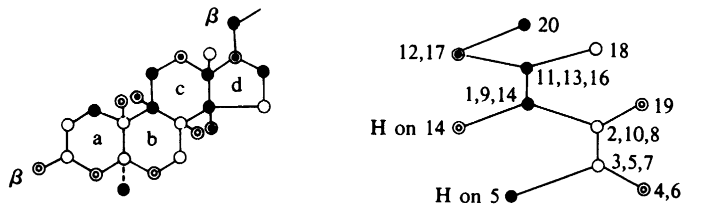


Fig. 2. Constellation of a saturated beta-sterol; ● = atoms in bottom plane; ◎ = atoms in second plane; ○ = atoms in third plane; ⊙ = atoms in top plane. Lateral view of molecule is shown at right.

The sterols, from which the name of the entire group is derived, are monovalent alcohols with a secondary hydroxyl group on C_3 usually in β-position. The best known representative is *cholesterol*. See Fig. 3. This compound forms esters with a great variety of acids. Both the free and esterified sterols accompany the neutral fat and the phosphatides in most animal and plant fat. Upon alkaline hydrolysis, the other lipid constituents form fatty acid soaps; the fraction which remains insoluble in aqueous alkaline solution is called "unsaponifiable" and consists primarily of sterols. The variations in the cholesterol content of blood of animals, particularly in humans, are of significance for the diagnosis of various diseases. See **Cholesterol.** Cholesterol is the principal sterol of all vertebrate animals. It is also found in some mollusks and in crustaceans, where it may be of alimentary origin.

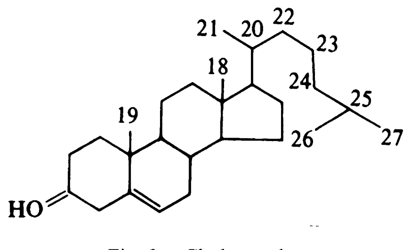


Fig. 3. Cholesterol.

Sterol Biogenesis. The finding that ingestion of radioactively labeled acetic acid leads to the synthesis of radioactive cholesterol was the first step in the elucidation of sterol biosynthesis. A growth factor for *Lactobacilli*, replaceable by acetic acid, was found to have the structure $HOCH_2 \cdot CH_2 \cdot C(CH_3)(OH) \cdot CH_2 \cdot CO_2H$. This compound was termed *mevalonic acid*. Its close relationship to a trimer of acetic acid is evident. Six molecules of the C_6 acid polymerize, losing their carboxyl groups, to the linear isoprenol *squalene*, a hydrocarbon occurring in nature. Twelve of the carbon atoms shown as circles in Fig. 4, originate from the carboxy groups of the original acetic acid, the remaining 18 from the methyl groups. Squalene folds in the manner indicated in

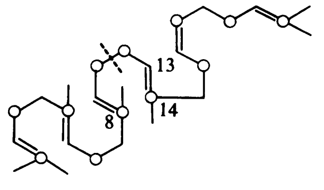


Fig. 4. Squalene.

the formula and yields (with a two-step rearrangement of the methyl group from C_8 and C_{13}) *lanosterol*, a "protosterol" found in wool and fat. This protosterol loses three methyl groups in positions 4, 4, and 14 in the course of biosynthesis, yielding *zymosterol*, found in yeast, which is convertible to cholesterol. The gradual oxidative degradation of the side chain in cholesterol to bile acids and subsequently to the various steroid hormones in animals is well established and has been confirmed by C-14 tracer studies. Many of the enzymes operative during these hormone syntheses in the insertion of hydroxyl groups on individual carbon atoms have been separated and localized in various cell constituents. Major steps in sterol biogenesis are shown in Fig. 5.

Classification of Medically Important and Useful Steroid Hormones

In addition to the medical uses of steroid hormones for alleviating conditions brought about by insufficiencies or overabundance of any particular hormone of this class within the body and thus returning the desirable hormone balance, numerous therapies do not fall directly into these two categories. Rather, steroid hormones are used in connection with some ailments and diseases because of positive clinical results even though much remains to be learned concerning the details of their function.

Steroid hormones are difficult to classify because some of them serve large numbers of uses. The conventional approach places them into four categories.

Androgens and Anabolic Agents. Androgen is the male sex hormone. The androgenic hormones are synthesized in the body by the testis, the cortex of the adrenal gland, and, to a slight extent, by the ovary. The androgens have a number of sexually related functions. Androgens also serve as anabolic agents, i.e., in nutrition of muscle and bone in both male and female persons. A number of androgens have been synthesized. Some of these reduce or eliminate the production of male secondary sex characteristics when administered to females (growth of facial hair, lowering of the voice, etc.)

Estrogens. Estrogen is a general term for female sex hormones. They are responsible for the development of the female secondary sex characteristics, such as the deposition of fat and the development of the breasts. Estrogens are produced by the ovary, and, to a lesser degree, by the adrenal cortex and testis. Some synthetic *nonsteroid* compounds, such as diethylstilbestrol and hexestrol, have estrogenic activity.

Progestogens and Progestins. Progesterone (Δ^4-pregnene-3,20-dione), $C_{21}H_{30}O_2$, is the female sex hormone secreted in the body by the corpus luteum, by the adrenal cortex, or by the placenta during pregnancy. It is important in the preparation of the uterus for pregnancy, and for the maintenance of pregnancy. Progesterone is believed to be the precursor of the adrenal steroid hormones.

Adrenal Corticosteroids. Among these hormones are compounds which have been found to be antiinflammatory, antiallergic, and antirheumatic agents and consequently are very important in steroid therapy. Cortisone was first applied in 1949 and became a major drug for treating rheumatoid arthritis, among other ailments, soon thereafter. Hydrocortisone followed and a bit later several synthetic analogues (not found in the body) were developed. Well known among these is prednisone, found particularly effective in cases of diseases of collagen tissue. But its use is much more widespread. Several examples are given in the accompanying table. See also **Hormones.** Steroid therapy throughout the years has had to cope with production of numerous side effects. Problems like this provide an incentive to continued vigorous research for new compounds.

In addition to the foregoing four major classifications (by application), there are also steroid hormones which are effective diuretic, an-

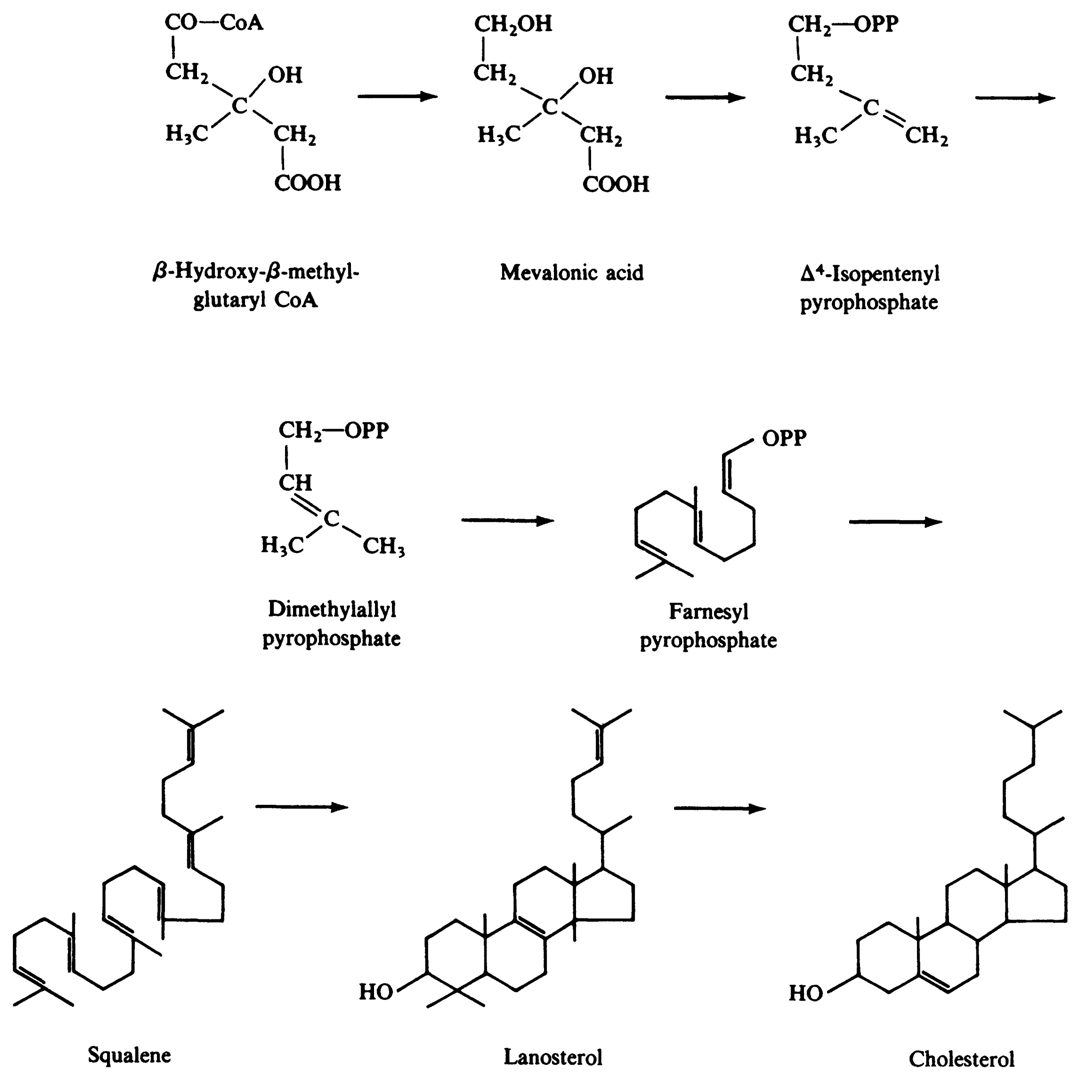


Fig. 5. Structures of key compounds involved in biogenesis of sterols.

tiduretic, and local anesthetic agents. Among these are aldosterone, spironolactone, and hydroxydione sodium.

Bile Acids

The bile acids are monocarboxylic acids of the steroid group with 24 carbon atoms and 1–3 secondary hydroxyl groups. They occur in the bile of all vertebrates from the teleosts upward, mostly in peptidic conjugation with glycine and taurine. The bile acids are described in the entry on **Bile.**

STETHOSCOPE. An instrument for listening to sounds originating within the body, especially of the heart and lungs. The stethoscope is a low-pass filter the exact acoustical characteristics of which vary significantly from one manufacturer to another. Electronic stethoscopes are available.

STIBNITE. The mineral stibnite, antimony sulfide, Sb_2S_3, is found in radiated groups of acicular orthorhombic crystals or in other sorts of aggregates, as well as blades, also as columnar or granular masses. It shows a highly perfect pinacoidal cleavage; conchoidal fracture; hardness, 2; specific gravity, 4.63–4.66; luster, metallic and very brilliant on cleavage faces or freshly fractured surfaces. Its color is a steely gray; the streak very similar in color, may be covered with a black, sometimes iridescent tarnish.

Stibnite is the most common antimony mineral known and is the chief ore of that metal. It is a primary ore mineral and occurs with other antimony minerals and galena, sphalerite, and silver ores. It is found in Germany, Rumania, the Balkans, Italy, Borneo, Peru, Japan, China, Mexico; and in the United States in California and Nevada.

The name stibnite is derived from the Latin word for antimony, *stibium.*

STIFFNESS. In general, the ability of a system to resist a prescribed deviation. In the case of a deformable elastic medium, stiffness is the ratio of a steady force to the elastic displacement produced by it, e.g., for a spring the force required to produce unit stretch. The term is applied most often to an elastic system vibrating about a position of equilibrium. Acoustic stiffness is the quantity which, when divided by 2π times the frequency, gives the acoustic reactance associated with the potential energy of the medium or its boundaries. The unit commonly used is dyne/centimeter. Mechanical stiffness is expressed in terms of the various elastic constants and moduli.

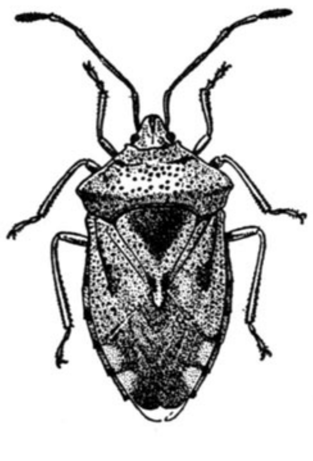
STIGMA. 1. A secondary sexual mark of insects. In many species of butterflies, it consists of a patch on the wing of the male bearing modified scales on a more or less modified area of the wing membrane. 2. A term used by some entomologists in place of spiracle. 3. A pigmented spot or "eye spot" sensitive to light in some flagellate protozoa.

STIGMATIC. Two uses of this term in optics are: (1) For a bundle of rays, homocentric. (2) For an optical system, having equal focal power in all meridians.

STILBITE. The mineral stilbite, $NaCa_2(Al_5Si_{13})O_{36}\cdot 14H_2O$, is a zeolite, the compound monoclinic crystals of which are usually grouped in approximately parallel positions, forming sheaflike aggregates, which have a soft pearly luster, whence the name stilbite from the Greek, meaning luster. The less commonly used term desmine is likewise from the Greek, meaning a bundle. Stilbite has one perfect cleavage; uneven fracture; is brittle; hardness, 3.5–4; specific gravity, 2–2.2; luster, vitreous to pearly; color, usually white but may be brownish, yellowish, red or pink. Its streak is white, and it is transparent to translucent. Like the other zeolites stilbite occurs in cavities in basalts and traps, rarely in granites and gneisses. Of the many localities may be mentioned Trentino, Italy; the Harz Mountains; Valais, Switzerland; Arendal, Norway; the Ghats Mountains of India; and Mexico. The Triassic traps of New Jersey and Pennsylvania furnish specimens as do also rocks of the same age in Nova Scotia. This mineral sometimes is called *desmine*.

STIMULUS (Nerve). Physical agent used to set up a nerve impulse. Since the nerve fiber is excitable by electric currents, a common stimulus is a brief current pulse whose amplitude and duration and rate of delivery can be controlled with precision. Other physical agents can be used, depending upon the input under examination, e.g., sudden stretch in the case of stretch receptor, sound wave for auditory input, and light in the case of the eye.

STINK BUG (*Insecta, Hemiptera*). A flattened bug of generally ovate form, in many species with an angular outline. Most species are moderately large, reaching a length of about half an inch. The many species, constituting the family *Pentatomidae*, are also characterized by the fetid odor of the secretion discharged from glands opening on the lower surface of the body.


Stink bug.

One species, the harlequin cabbage bug or calico-back, is a troublesome pest. It is best controlled by clean cultivation of fields, hand picking of bugs and their eggs, and the use of trap crops, planted early to attract the insects.

Some members of the group eat other insects and are probably beneficial in destroying pests, but unfortunately even the harmless species may contaminate berries with their unpleasant odors.

STOCHASTIC. The adjective "stochastic" implies the presence of a random variable; e.g., stochastic variation is variation in which at least one of the elements is a random variable and a stochastic process is one wherein the system incorporates an element of randomness as opposed to a deterministic system.

The word derives from Greek στόχος, a target, and a stochastiches was a person who forecast a future event in the sense of aiming at the truth. In this sense it occurs in sixteenth-century English writings. Bernoulli in the *Ars Conjectandi* (1719) refers to the "ars conjectandi sive stochastice." The word passed out of usage until revived in the twentieth century.

STOCHASTIC PROCESS. A family of variates (x_t) where t assumes values in a certain range T. In most practical cases x_t is the observation at time t and T is a time-range, but t may also refer to distribution in space and may be considered for discontinuous or continuous values.

A stochastic process (x_t) is said to be stochastically continuous if, for values $t, t + h_1, t = h_2, \ldots$ with h_n tending to zero as n tends to infinity

$$\lim_{n\to\infty} x_t + h_n$$

exists in the sense of stochastic convergence and is equal to x. Likewise, if

$$\lim \frac{x_{t+h_n} - x_t}{h_n}$$

exists in the sense of stochastic convergence, the process is said to be stochastically differentiable. And if the process exists in $a \leq t \leq b$ and the Riemann integral

$$\int_b^a x_t dt$$

exists in the sense of stochastic convergence, the process is said to be stochastically integrable.

STOKES FLOW. Flow of a viscous fluid at a very small Reynolds number when inertial, acceleration forces are negligible and the Navier-Stokes equations reduce to

$$\mu\nabla^2\mathbf{v} - \text{grad}\, p = 0$$

The approximation may not be possible in all parts of the flow even at very low speeds.

STOKES LAW FOR VISCOSITY. A solid sphere moving with velocity V through a fluid of viscosity μ experiences a resistance to motion

$$F = 6\pi\mu a\mathrm{V}$$

where a is the radius of the sphere. The law is accurate only if the flow Reynolds number $\rho aV/\mu$ is less than 0.1.

STOKES THEOREM. The surface integral of the curl of a vector function equals the line-integral of that function around a closed curve bounding the surface.

$$\int_S \nabla \times \nabla \cdot d\mathbf{S} = \mathbf{V} \cdot d\mathbf{S}$$

If the components of **V** in rectangular Cartesian coordinates are u, v, w and the direction cosines of the normal to $d\mathbf{S}$ are λ, μ, ν the theorem may also be given as

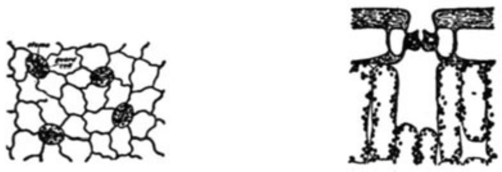
$$\int_S \left[\lambda\left(\frac{\partial w}{\partial y} - \frac{\partial v}{\partial z}\right) + \mu\left(\frac{\partial u}{\partial z} - \frac{\partial w}{\partial x}\right) + \nu\left(\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}\right)\right] dS = \int_C (u\,dx + v\,dy + w\,dz)$$

See also **Curl;** and **Vector.**

STOLON. 1. In botany, a branch that grows out horizontally from the base of the stem, takes root, and gives rise to a new plant at the nodes or at the tip.

2. In zoology, a shoot growing from the base of an animal or a colony, from which other individuals arise by budding.

STOMATE (or Stoma). The name applied to the minute pores which occur abundantly in the epidermis of leaves and, less abundantly, in the epidermis of young stems, flower parts and fruits. Each stomate is located between two distinctive epidermal cells called *guard cells.* The size and shape of the guard cells varies considerably from one species to another. Unlike other epidermal cells the guard cells contain chloroplasts. In many species of plants stomates occur in both the upper and lower epidermis of the leaf. In many other species, especially of woody plants, stomates occur only in the lower epidermis. Even in those species in which the stomates are present in both epidermises there are commonly, although not invariably, more per unit area in the lower epidermis. In floating leaves, such as those of the water lily, stomates are present only in the upper epidermis. In many kinds of plants, the stomates are restricted to grooves or furrows in the leaf.


(a) Portion of the lower epidermis of a geranium leaf; (b) sunken stoma of carnation.

The number of stomates per unit area varies with the kind of plant and also, within limits, with the conditions under which the plant has developed. The range is from a few thousand to about a hundred thousand per square centimeter of leaf surface. A single corn plant has been estimated to bear from 140 to 240 million stomates and the number on a large tree could be represented only by a figure of astronomical dimensions. The size of the individual stomates also varies greatly from species to species, their dimensions being expressed in microns. In some species the fully open stomates may be as large as 8 to 10 × 30 to 40 micromillimeters, as measured along the two axes of the elliptical pore, but in most species they are smaller. Species in which the stomates are relatively small usually have more per unit area than those which have relatively large stomates.

The structure of the cell wall of a guard cell is quite complex, varying considerably in thickness and elasticity from one part of the cell to another. This wall structure is such that when the guard cells increase in turgidity their inner walls—those bounding the pore—bow away from each other causing a widening of the stomate. In general, therefore, when the guard cells are turgid the stomate is open; when the guard cells are flaccid the stomach is closed.

Shifts in the turgidity of the guard cells causing opening and closing of stomates are conditioned by a number of factors among which the most important are light, temperature, and the internal water supply of the leaf. In general stomates are open in the light and closed in the dark although there are many exceptions to this statement. In general, low temperatures are unfavorable to stomatal opening and when the temperature falls below the optimum for stomatal opening for a given species the stomates will remain closed or open only incompletely even if light conditions and water supply are favorable. Similarly drought conditions, by resulting in a reduction in the water content of the leaves, are usually unfavorable to stomatal opening even if light and temperature conditions are favorable. During prolonged droughts the stomates of many species remain nearly or completely closed most of the time. Night opening of the stomates is of regular occurrence in some species such as certain cacti, in which they are not usually open in the daytime, and may occur in some other species under certain conditions.

When the stomates are open they serve as the principal pathways through which gases diffuse into or out of the leaf; when the stomates are closed all gaseous exchanges between a leaf and its environment are greatly retarded. The gases of greatest physiological importance which enter or depart from a leaf principally through the stomates are oxygen, carbon dioxide, and water vapor. Loss of water vapor in the process of transpiration occurs principally through the stomates. Similarly the inward diffusion of carbon dioxide and outward diffusion of oxygen, the gaseous exchanges accompanying photosynthesis, occur principally through the stomates.

STONE AGE. An archeological term to designate a cultural level that is characterized by the use of stone implements. Classically it is divided into the Eolithic, Paleolithic, and Neolithic Periods. The Stone Age is the first of the so-called three-age system (Bronze Age and Iron Age following).

STONE ROLLER (*Osteichthyes*). Of the group *Cypriniformes*, family *Cyprinidae*, the stone roller is a bottom-feeding fish of moderate size found in small streams from Wyoming to New York and south to the Gulf. It is herbivorous.

STOPPING POWER. A measure of the effect of a substance upon the kinetic energy of a charged particle passing through it. The linear stopping power S_l is the energy loss per unit distance and is given by $S_l = -dE/dx$, where E is the kinetic energy of the particle and x is the distance traversed in the medium. The *mass stopping power* S_m is the energy loss per unit surface density traversed, and is given by $S_m = S_l/\rho$, where ρ is the density of the substance. The *atomic stopping power* S_a of an element is the energy loss per atom, per unit area normal to the particle's motion, and is given by $S_a = S_l/n = S_mA/N$, where n is the number of atoms per unit volume, N is the Avogadro number, and A is the atomic weight. The *molecular stopping power* of a compound is similarly defined in terms of molecules; it is very nearly if not exactly equal to the sum of the atomic stopping powers of the constituent atoms. The *relative stopping power* is the ratio of the stopping power of a given substance to that of a standard substance, commonly aluminum, oxygen or air. The *stopping equivalent* for a given thickness of a substance is that thickness of a standard substance capable of producing the same energy loss. The *air equivalent* is the stopping equivalent in terms of air at 15°C and 1 atmosphere as the standard substance. The term equivalent stopping power is not clearly defined, but sometimes is used synonymously with relative stopping power and sometimes with stopping equivalent.

STORAGE BATTERY. See **Battery.**

STORAGE (Computer). Any medium which is capable of storing information. As generally defined, however, a storage unit is a device on or in which data can be stored, read, and erased. The major classifications of storage devices associated with computer systems are: (1) *immediate-access*; (2) *random-access*; and (3) *sequential-access*. As a general rule, the cost per bit of information is greater for immediate-access storage devices, but the access time is considerably faster than for the other two types.

The various physical means to effect storage are described under (**Memory (Electronic)**).

Immediate-Access Storage. In these devices, information can be read in a microsecond or less. Usually an array of storage elements can be directly addressed and thus all information in the array requires the same amount of time to be read. Specific storage configurations in this class include core storage, and monolithic storage.

Random-Access Storage. Storage devices in which the time required to obtain information is independent of the location of the information most recently obtained. This strict definition must be qualified by the observation that what is meant is *relatively* random. Thus, magnetic drums are relatively nonrandom access when compared with monolithic storage, but are relatively random access when compared with magnetic tapes for file storage. Disk-storage and drum-storage units usually are referred to as random-access storage devices. The time required to read or write information on these units generally is in the 10- to 200-milli-second range, but is dependent upon where the information is recorded with respect to the read/write head at the time the data are addressed.

Sequential-Access Storage. Storage devices in which the items of information stored become available only in a one-after-the-other sequence, whether or not all the information or only some of it is desired. Storage on magnetic tape is an example.

Some other computer storage configurations are defined by:

Auxiliary Storage. A storage device in addition to the main storage of a computer; e.g., magnetic tape, disk, diskette, or magnetic drum. Auxiliary storage usually holds much larger amounts of information than the main storage, and the information is accessible less rapidly.

Buffer Storage. (1) A synchronizing element between two different forms of storage, usually between internal and external. (2) An input device in which information is assembled from external or secondary storage and stored ready for transfer to internal storage. (3) An output device into which information is copied from internal storage and held for transfer to secondary or external storage. Computation continues while transfers between buffer storage and secondary or internal storage or vice versa take place. (4) Any device which stores information temporarily during data transfers.

Circulating Storage. A device or unit which stores information in a train or pattern of pulses, where the pattern of pulses issuing at the final end are sensed, amplified, re-shaped and re-inserted into the device at the beginning end.

External Storage. (1) The storage of data on a device which is not an integral part of a computer, but in a form prescribed for use by the computer. (2) A facility or device, not an integral part of a computer, on which data usable by a computer is stored such as off-line magnetic tape units, or punch card devices.

Internal Storage. (1) The storage of data on a device which is an integral part of a computer. (2) The storage facilities forming an integral physical part of the computer and directly controlled by the computer. In such facilities all data are automatically accessible to the computer; e.g., magnetic core, and magnetic tape on-line.

Main Storage. Usually the fastest storage device of a computer and the one from which instructions are executed.

Program Storage. A portion of the internal storage reserved for the storage of programs, routines, and subroutines. In many systems protection devices are used to prevent inadvertent alteration of the contents of the program storage.

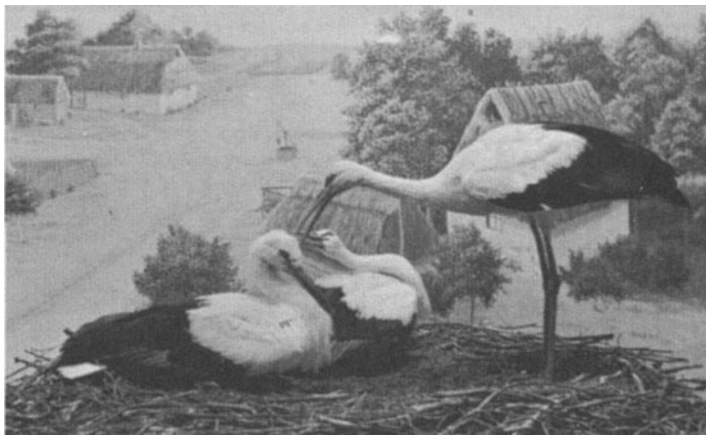
Serial Storage. A storage technique in which time is one of the factors used to locate any given bit, character, word, or groups of words appearing one after the other in time sequence, and in which access time includes a variable latency or waiting time of from zero to many word times. A storage is said to be serial by word when the individual bits comprising a word appear serially in time; or a storage is serial by character when the characters representing coded decimal or other nonbinary numbers appear serially in time; e.g., magnetic drums are usually serial by word but may be serial by bit, or parallel by bit, or serial by character and parallel by bit.

Working Storage. A portion of the internal storage reserved for the data upon which operations are being performed. Synonymous with working space and temporary storage and contrasted with program storage.

Thomas J. Harrison, International Business Machines Corporation, Boca Raton, Florida.

STORKS (*Aves, Ciconniformes*). Large birds of the Old World and South America. They range from Argentina northward to Mexico, but not into the United States. They are found widely in Europe; Asia, Africa, and in parts of Australia. Their size ranges from about 42 to 60 inches (107 to 152 centimeters) tall when standing and ranges from 4 to 6.5 feet (1.2 to 2 meters) in length. The color varies with species, ranging from white through dull gray, sometimes having a somewhat greenish sheen. There are numerous species, of which at least 28 tropical species are known. Most storks are migrating birds, flying with their neck stretched and legs slanted backward in a V-shape. The migration pattern is from colder to warmer areas in the fall and returning to the colder climates for spring and summer. Some species prefer chimney tops as nesting locations, while other species like marshes near water and in tall grasses. There are usually 3 to 5 blunt, oval, white eggs, the eggs being laid at intervals of about 2 days. Both parents incubate the eggs for 30 to 38 days. Some species fly in large flocks, particularly those migrating from Africa to Europe and Asia in early spring. The return flight is usually made some time in August. Numerous bird-watching groups have banded the legs of the baby storks to keep track of these migrations. After banding, a baby will often "play dead" for quite a period. The young storks are called siblings.

The common white stork of Europe, *Ciconia alba*, which nests on the tops of chimneys is probably the best-known species. See accompanying illustration. The jabiru (*Jabiru mycteria*) is a white stork with green head and green coloring on the partly-naked neck. The upper portion of the neck also may have dark blue and orange coloring. This bird may


Storks. (*A. M. Winchester.*)

achieve a length of some 55 inches (140 centimeters) and is one of the larger flying birds. Nesting preference is pine trees. Eggs are light-green in color. This stork is known to prey on grass fires, catching small animals as they are flushed out by the heat. Other dietary items include small fish and a variety of insects. Some varieties of jabiru have a bronze coloration on their wings and a bit of green color in the tail—with coral color legs.

Adjutant is a name applied to storklike birds of several species. These birds occur in Africa and the Oriental region where they are valuable as scavengers. From at least one species (*Leptoptilus crumenifer*) the soft downy feathers known as marabou are secured. Birds as tall as 6–7 feet (1.8–2.1 meters) have been reported in India. These birds feed on frogs, carrion, and fish. They are known for their success in chasing vultures away from carrion.

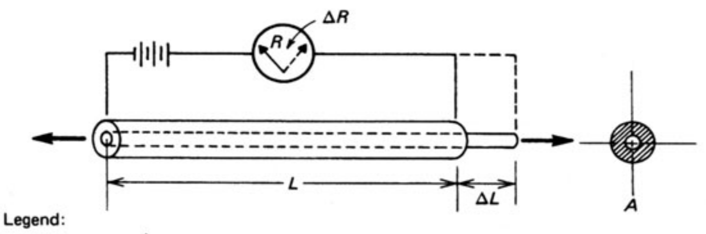
The name stork sometimes is confused with that of the related herons, as in the case of the whale-headed or shoe-billed species of the White Nile. This species, called both heron and stork, has an enormous and powerful beak, both broad and deep and provided with a strong hook at the tip. See also **Ciconiiformes**.

STRAIGHT EDGE. One of three surfaces, any two of which when placed together, coincide throughout their length. A straight-edge also is the name of a hand device in the form of a piece of material whose edge is "perfectly" straight and which is used for testing plane surfaces and drafting straight lines. A ruler is a straight-edge.

STRAIN ENERGY. A term that usually denotes the elastic energy stored in a stressed body that can be recovered as work upon unloading.

STRAIN GAGE. An instrumental device used to measure the dimensional change within or on the surface of a specimen. The electrical-type strain gage may operate on the measurement of a capacitance, inductance, or a resistance change that is proportional to strain. The bonded resistance-type strain gage is the most widely used in many fields. The principle of operation was discovered in 1856 by Lord Kelvin. The fundamental relationships between resistance change and strain are shown by Fig. 1. When a conductor of length L and cross-sectional area A is elongated, the length increases and the area decreases by Poisson's effect to produce an increase in resistance. The resistance change $\Delta R/R$ then is related to the length change $\Delta L/L$, or strain ϵ by the strain sensitivity or gage factor. If the strain sensitivity were dependent upon dimensional change only, resulting from the usual Poisson ratio of 0.3, then all metallic conductors would have a theoretical value of 1.6 in the elastic range and 2.0 in the plastic range of the alloy used where the Poisson ratio becomes 0.5. However, the resistivity alters with strain in order to produce the gage-factor range of 2.0 to 4.5 experienced by the alloys most often used for metallic strain gages.

Early bonded strain gages were developed in the 1930s by Dr. Arthur C. Ruge at the Massachusetts Institute of Technology. The gage was made from fine strain-sensitive wire attached to a thin paper carrier



Legend:

$R = \rho \frac{L}{A}$ = Resistance, ohms
L = Conductor Length
A = Cross Section Area
ρ = Resistivity Constant
$S = \frac{\Delta R/R}{\Delta L/L}$ = Strain Sensitivity (gage factor when applied to a specific gage)
$\Delta R/R$ = Resistance Change
$\Delta L/L$ = Strain

Fig. 1. Basic relationships between resistance change and strain in resistance-type strain gage.

with nitrocellulose cement for dimensional stability and to provide electrical isolation from a metal specimen. The carrier was bonded to the specimen with the same cement so that the specimen surface strain would be reliably transmitted into the fine-wire grid. The grid shape was designed to provide maximum gage resistance in the smallest possible gage length and width. Foil also can be used as the strain element. In addition to material variations, a large number of grid shapes is available in modern strain gages, particularly in the foil construction to meet the needs of specific applications. Gage lengths range from 6 to $\frac{1}{64}$th inch and resistance values from 1,000 to 60 ohms.

Strain gage sensing element materials most commonly used are: (1) *Constantan* (copper-nickel alloy) used mainly for static strain measurement because of low and controllable temperature coefficient; (2) *Nichrome V* (nickel-chrome alloy) frequently used for high-temperature static and dynamic strain measurements; (3) *Dynaloy* and *IsoElastic* (nickel-iron alloy plus other proprietary ingredients), used for dynamic tests where the larger temperature coefficients of these materials are of no consequence; (4) *Stabiloy* and *Karma* (nickel-chrome alloys) containing other ingredients which provide wider temperature compensation range; and (5) platinum alloy (usually tungsten) which shows unusual stability and fatigue life at elevated temperatures. Additionally, semiconductor strain gages are used. These are similar to conventional metallic gages, the principal difference being the greater response of semiconductor gages to both strain and temperature. They have large and nonlinear resistance versus strain, arising primarily to the piezoresistive effect. They have found their main use in development of high-output transducers, such as load cells and pressure cells.

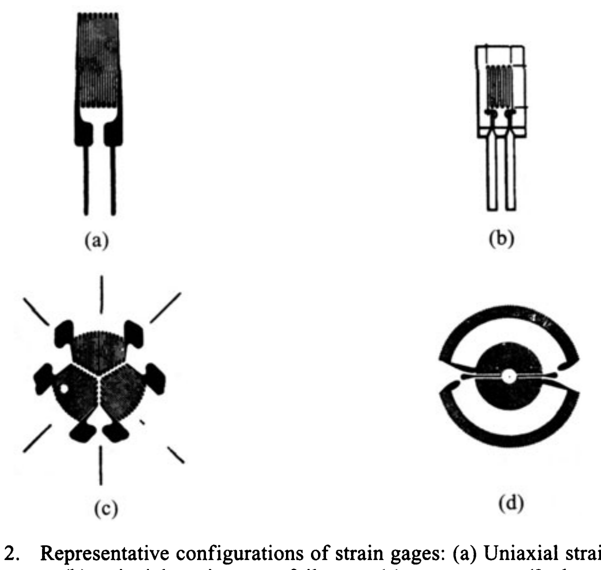


Fig. 2. Representative configurations of strain gages: (a) Uniaxial strain gage, wire type; (b) uniaxial strain gage, foil type; (c) rosette gage (3-element foil) rosette, 60 degrees planar; and (d) 4-element "diaphragm" gage which takes advantage of the tangential tensile strains developed at the center of the diaphragm and the compressive radial strains present at the edge.

The application of strain gages breaks down into two main areas: (1) applications where the gage measures strain as the primary objective of measurement (stress analysis of various structures—bridges, boat hulls, etc.); and (2) uses where the measurement of strain, in turn, is a measure of another variable, such as pressure, impact, acceleration, and other force-associated variables, i.e., the strain gage becomes a transducer. Because strain gages, when properly compensated for temperature effects, can achieve overall accuracies of plus-or-minus 0.10% or better and because of their great flexibility, relatively low cost, and availability of numerous configurations, they are widely used in numerous kinds of transducers. A few representative strain gage configurations are illustrated in Fig. 2.

STRAIN HARDENING EXPONENT. A measure of the rate of hardening with strain as in a tensile test, as expressed by the constant "n" in the equation $\sigma = \sigma_0 \delta^n$, where σ is the true stress, σ_0 the true stress at unit strain, and δ the true strain.

STRAIN THEORY. A theory first proposed by von Baeyer to explain the relative stability of various carbon compounds. It may be stated in the form: The regular tetrahedral-symmetric position is the most stable of all possible positions of neighboring carbon compounds; variations from this position produces increased energy content, and hence strain. Since the angle at the vertex of a regular tetrahedron is 109°28; this theory ascribes minimum strain to cyclopentane, of the polymethylenes. The theory is borne out by the lesser stability of cyclobutane and cyclopropane, but not to the degree that might be expected by the stability of some of the higher-membered rings. In that case, the lesser strain is often due to a spatial or three-dimensional structure.

Extensions of the strain theory have been made, with varying success, to other hydrocarbon ring structures, saturated and unsaturated, to ring compounds in which the hydrogen atoms have been variously substituted, and to rings containing atoms other than carbon, as well as to bicyclic and polycyclic systems.

STRATH. 1. In geography, a broad alluviated valley. 2. In geomorphology, a valley or confluent valleys which represent a local base level of erosion or a local and incipient peneplain.

STRATIFICATION (Ecology). As described under **Ecology,** the primary classification of ecological communities is by the broad nature of the environment—fresh-water, salt-water, terrestrial, etc. communities. Vertical stratification also occurs in many of these communities as, for example, certain fishes that prefer shallow water, others that inhabit deep water; or in a tropical forest, animals and birds that prefer the tree tops, others that prefer any one of several intervening levels between the surface and the top canopy. Further, some organisms prefer an underground habitat.

STRATIGRAPHY. The study of the origin and chronological successions of the observable rocks of the lithosphere, in which each lithologic unit is considered to be a formation. The term formation is usually confined to bedded or stratified rocks, including lava flows and volcanic ashes. The major principles involved in the correlation (dating) of formations are: (1) The law of superposition, or that the chronological sequence of any stratigraphic section depends upon the original order in which the formations were laid down; thus the fundamental basis of stratigraphy is structural geology. (2) Index fossils, or those species of fossils whose stratigraphic age are already known. (3) lithology. Igneous rocks may be dated by the age of the sedimentary rocks which they intrude, or overlie; or by radioactive minerals. See also **Paleontology; Structural Geology;** and **Radioactivity.**

STREAM FUNCTION. A parameter of two-dimensional, nondivergent flow, the value of which is constant along each streamline. For flow in the (x, y)-plane, the stream function ψ is related to the respective coordinate velocities u and v by the equations

$$u = -\frac{\partial\psi}{\partial y}, \quad v = \frac{\partial\psi}{\partial x}$$

Stokes's stream function (also called *current function*). If the flow is three-dimensional but is axisymmetric (i.e., the same in every plane containing the axis of symmetry), a Stokes's stream function ψ will exist such that

$$v_s = -\frac{1}{r}\frac{\partial\psi}{\partial n}$$

where v_s is the speed in an arbitrary direction s, r the distance from the axis of symmetry, and n is normal to the direction s, increasing to the left. Note that Stokes's stream function has dimensions L^3T^{-1}.

Stream functions can also be defined for more complex three-dimensional flows.

STREAMING (Molecular). Application of kinetic theory to the flow of gas through a tube at low pressures, such that the mean free path is large compared with the diameter of the tube. In this case, the streaming of the gas is due to the random motion of the molecules, and to the density gradient down the tube, so that the numbers of molecules traversing a given cross section in opposite directions is different. For a tube of circular cross section, the mass flowing per second is proportional to the pressure difference and the cube of the radius.

STRESS RELIEVING. A heat treatment to relieve residual stresses.

STRESS-RUPTURE TEST. A form of short time creep test in which a tensile specimen is deformed to rupture (fracture) under constant load and temperature.

STRESS-STRAIN CURVE. A graphical representation of the relation between unit stress and unit deformation in a stressed body as a gradually increasing load is applied.

STRESS (Structural). A quantitative expression of a condition within an elastic material due to deformation, or strain, brought about by external forces, inequalities of temperature, or otherwise. Its measure is always the ratio of a force to an area. By some, stress is interpreted as a force distributed over an area, and the above ratio is called the "unit stress." Central, torsional and bending loads cause stress. The total resisting force acting at any section of the body divided by the area of the section is the average stress, commonly expressed in pounds per square inch. The unit stress is the resisting force on a unit of area. The component of a stress which acts at right angles to a surface is known as the normal stress. If this stress is produced by a load whose resultant passes through the center of gravity of the area, it is called an axial or direct stress. A normal resisting force which causes the fibers to increase in length is a tensile stress, while one which shortens the fibers is a compressive stress. The latter is often called a bearing stress. The component of any stress which lies in the plane of the area is a shearing stress. See **Elasticity.**

The "conventional stress" as applied to tension or compression tests is the instantaneous value of the load divided by the original cross-section area. This type of stress is used by engineers for design purposes since their primary concern is how large a load or stress a structure is capable of carrying, as measured in terms of the structures' original dimensions. Frequently, however, for scientific purposes, it is important to know the true stress which is determined by dividing the force at a given instant by the area existing at that time.

Direct tensile or compressive stresses are known as primary stresses. The bending stress, resulting from deflection, is called a secondary stress. The stresses developed in a column due to the lateral deflections are of a secondary nature. The rigidity of the riveted or welded joints of a truss which has deflected due to the axial deformation of its members causes bending stresses in the members which are classified as secondary stresses. The resistance offered by a body to a combination of direct and bending loads is frequently called a combined stress. A normal stress which occurs at a point in a plane on which the shearing stress is zero is known as a principal stress. If this normal stress is tensile, it is often called a diagonal tension stress; if compressive it is known as a diagonal compression stress.

The internal resisting force which arises in a restrained body due to temperature changes is a thermal stress. The adhesive resistance which is developed in the concrete surrounding the steel reinforcing rods when a reinforced concrete member is subjected to load is known as bond stress. Safe unit resisting forces which are used in design are called working stresses. These are usually taken as a percentage of the ultimate stress or the elastic limit of the material.

The stress developed in bridge members as a result of traction between the wheels of the live load and the supporting surface is called a traction stress. The effect of these stresses is usually neglected in highway bridge design but must be considered in railway bridges.

STRIATION. Three uses of this term are: 1. A striped appearance of the positive column of a Crookes tube. 2. A defect of optical materials, such as optical glass, having the appearance of streaks through the material, and seriously affecting the material for use as lenses or windows. 3. The scratches on bedrock or on pebbles and boulders which are the result of glaciation. These are called glacial striae to distinguish them from the striae which occur on the surfaces of fault planes.

STRIGIFORMES (*Aves*). This order of birds is distinctly set off from other bird kinds. In physical structure they are remarkably homogeneous. They are recognized immediately by their large heads, their forward-directed eyes, the seeming absence of a neck, and their soft plumage. Their length is 15–80 centimeters (6–31 inches) and the weight is 55–4200 grams (2 ounces to 9 pounds). They have 11 primaries and 12 (rarely 10) tail feathers. The fourth toe is abducted and reversed diagonally. The beak is drawn down steeply, making it appear small. The upper mandible is hooked. The plumage is particularly well suited for nocturnal activity: gray, brown, black, and white shades arranged in patches, stripes, bands, and streaks effect a muted coloration. See accompanying illustration. They possess calls, mostly sounding an "oo" or an "ee," that serve to find and to keep contact with their conspecifics during their nocturnal activities. Their clutch consists of 1–12 white, roundish eggs. The 2 families are: 1. Barn Owls (*Tytonidae*); and 2. Owls in the narrower sense (*Strigidae*). Altogether there are 28 genera and 144 species.

Face of tawny owl (*Strix aluco.*)

Owls are to be distinguished from raptors not only by both external and internal morphology, but quite notably by their behavior. Owl young, unlike those of raptors, are born blind. Most owls are active at dusk or during the night, and some during the day. They feed chiefly on smaller vertebrates, especially rodents, but also on insects and worms. Some, such as the fishing owls, are adapted for a specialized diet. Some smaller species prefer to consume insects. Indigestible parts of the prey (fur, feathers, bones, and chitin) are regurgitated.

The notion that owls cannot see well in daylight is still widespread, but untrue. Owls are indeed far-sighted, yet their vision is excellent in daylight; in complete darkness they are just as helpless as humans. The particular structure of their eyes, however, enables them to orient themselves adequately in very dim light. The number of light-sensitive cells in the retina (rods) is greatly multiplied in those owls which are active at night, such as the tawny owl (*Strix aluco*). On the other hand, such species have very limited color discrimination, while owls that are active at twilight or during the day do recognize colors. The retina possesses not only the rods, but also more cones than that of nocturnal species. The owl's eyes, which are usually very large, are immovable, forcing the bird to turn its head for a change of visual field.

The bony ring supporting the sclera (the dense, fibrous, opaque white outer coat of the eyeball), which is peculiar to a bird's eye, has developed into a regular cylinder in owls; it unites the dioptric apparatus (lens, cornea, and iris) with the relatively restricted retina. This results in an eye whose shape is strongly reminiscent of the "telescopic eyes" of some deep sea fish and some nocturnal mammals. Such "telescopic eyes" are adjusted to seeing in the dimmest of lights. The angle of the corneal fenestra is very wide (approximately 160 degrees), and the light gathered from a large field of vision is reduced to a small image on the retina.

The sense of hearing is superbly developed in owls. The margins of the auditory meatus are remodeled into feathered flaps which can close over the ear completely and so protect the sensitive inner parts. When erected, the flaps become wide, movable sound-funnels which enable the owl to pick up the faintest sounds from various directions. The feathery tufts on the head of various owl species are not ears, but decoration.

All owls lay pure white, more or less round eggs which as a rule are incubated only by the female. During that period it is usually the male alone that takes care of the feeding; they often pile up a store of food near the brooding ground. Both parents share in the rearing of the young.

No owl builds a genuine nest. The closest to it is a more or less thick layer of usually dry plant material put together by the short-eared owl, which is a ground-brooder. Other species dig a hollow at their brooding place and tear up boluses or food leftovers to use as a kind of nest bedding. In most cases, however, the eggs are deposited directly on the ground of the chosen brood site, sometimes in an abandoned crow's nest or in the aerie of a predator.

When young owls hatch, they are blind and covered with a whitish down, and their eyes are closed. Often the several young of one brood hatch at intervals of several days, for the eggs are laid in intervals and most species begin incubation after the first egg is laid. After one week, on the average, the ears and eyes open and gradually the young begin to get a second dress which in some species is downy and in others resembles the dress of the adult birds.

The family of the barn owls (*Tytonidae*) stands well removed based on recent studies, phylogenetically, from the true owls. The facial veil is more or less heart-shaped; the talon of the middle toe has a "comb;" and the posterior edge of the sternum has either two notches or none at all. There are 2 genera with a total of 11 species: (1) Barn Owls (*Tyto*), with 9 species; and (2) Bay Owls (*Pholidus*), with 2 species.

The 9 species of barn owls (*Tyto*) are as follows: 1. The Barn Owl (*Tyto alba*) reaches a length of 34 centimeters (13 inches), with a wingspread of 95 centimeters (37 inches), and weighs about 300 grams ($10\frac{1}{2}$ ounces). The facial veil is heart-shaped. The eyes are relatively small and black-brown; the wings are long and rather pointed. Since the intertarsal joints are close to each other when the bird is perching, it gives a knock-kneed impression. 2. The Cape Grass Owl (*Tyto capensis*) is found in southern Africa; it is a ground-breeding bird. 3. The Grass Owl (*Tyto longimembris*) occurs in the grasslands of India and southern China, some of the Sunda Islands, and Australia. 4. The Celebes Barn Owl (*Tyto rosenbergii*) inhabits the rain forests of Celebes. 5. The Minnahassa Barn Owl (*Tyto inexpectata*) inhabits northern Celebes. 6. The Madagascar Grass Owl (*Tyto soumagnei*) is relatively small, has long wings, and inhabits woodland glades in the rain forests of Madagascar; it feeds mainly on amphibia. 7. The Masked Owl (*Tyto novaehollandiae*) is found in the jungles of Australia and some Sunda Islands. 8. The Sooty Owl (*Tyto tenebricosa*) is the only species without yellow-brown or yellow-orange shades in its plumage. Its habitat is humid parts of the jungle and rain forests in New Guinea and Australia. 9. The New Britain Barn Owl (*Tyto aurantia*) occurs only in New Pomerani (New Britain).

The 2 species of bay owls (*Pholidus*) are as follows: 1. The Bay Owl (*Pholidus badius*) reaches a length of 30 centimeters (12 inches). 2. The Congo Bay Owl (*Pholidus prigoginei*) was discovered only in 1951, in the highlands northwest of Lake Tanganyika.

Bays owls are strictly nocturnally active birds, inhabiting the jungle. Wide, natural cavities serve as brood sites; the 3–5 purely white, roundish eggs are deposited directly on the floor, which is usually covered with a layer of decayed wood. The exact breeding season is not known. Bay owls feed on small vertebrates, such as small mammals, birds, reptiles, and amphibia, and possibly also on fish, for they like to hunt near the water. They are known by a hollow-sounding "hoo."

The family Owls (*Strigidae*) comprises 24 genera with 133 species. The claw of the middle toe has no pecten (comb). The 2 subfamilies are: (1) True Owls (*Buboninae*) and (2) Long-Eared Owls and Disk-Eyed Owls (*Striginae*).

The 18 genera which comprise the subfamily *Buboninae* are as follows: 1. The Scops Owls (*Otus*); 2. *Jubula* (Maned Owl, *Jubula lettii*); 3. *Lophostrix* (Crested Owl, *Lophostrix cristata*); 4. the Eagle Owls (*Bubo*); 5. the Fish Owls (*Ketupa*); 6. the Fishing Owls (*Scotopelia*); 7. the Spectacled Owls (*Pulsatrix*); 8. *Nyctea* (Snowy Owl, *Nyctea scandiaca*); 9. *Surnia* (Hawk Owl, *Surnia ulula*); 10. the Pygmy Owl (*Glaucidium*); 11. *Micrathene* (Elf Owl, *Micrathene whitneyi*); 12. *Uroglaux* (New Guinea Hawk Owl, *Uroglaux dimorpha*); 13. the Hawk Owls (*Ninox*); 14. *Gymnoglaux* (Bare-legged Owl, *Gymnoglaux lawrencii*); 15. *Sceloglaux* (Laughing Owl, *Sceloglaux albifacies*); 16. *Athene* (Little Owl, *Athene noctua*); 17. *Speotyto* (Burrowing Owl, *Speotyto cunicularia*); 18. the Tropical Tawny Owls (*Ciccaba*).

The 6 genera which comprise the subfamily of Long-eared Owls and Disk-eyed Owls (*Striginae*) are as follows: 1. The Disk-eyed Owls (*Strix*); 2. *Rhinoptynx* (*Rhinoptynx clamator*); 3. the Long-eared Owls (*Asio*); 4. *Pseudoscops* (Jamaican Owl, *Pseudoscops grammicus*); 5. *Nesasio* (Fearful Owl, *Nesasio solomonensis*); 6. the Tengmalm's Owls (*Aegolius*). See also **Owls.**

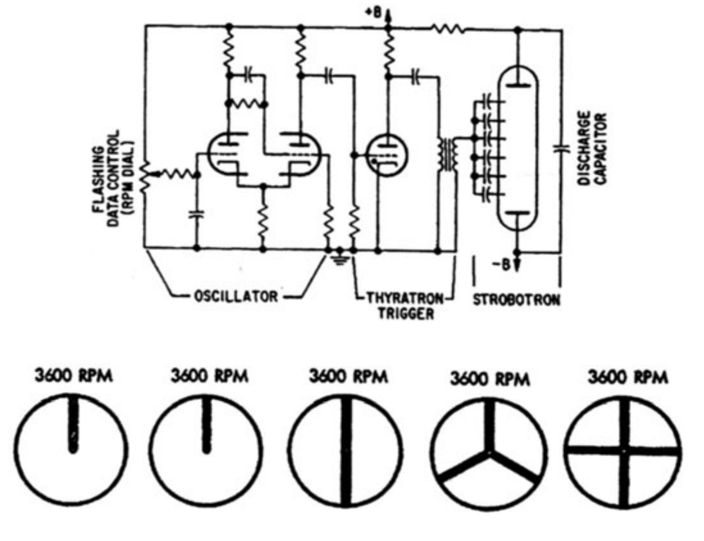
STRINGER. 1. A configuration of nonmetallic inclusions in wrought metals that is elongated in the direction of working. In steels, stringers are usually formed of oxides or sulfides. 2. One of the longitudinal beams in the floor system of a bridge.

STRIPPING. In nuclear reactions, an effect observed primarily in bombardment with deuterons, whereby only part of the incident particle merges with the target nucleus, and the remainder proceeds with most of its original momentum in a direction determined solely by its electromagnetic interactions with the target nucleus. The effect is strongly marked in deuteron bombardment, and in this case typically leads to directional neutron and proton beams that emerge from the target (if the latter is sufficiently thin) predominantly in the forward direction. The angular divergence of the beams decreases with increasing energy of the incident deuterons. Subsidiary peaks are sometimes observed at rather small angles relative to the forward direction. When the (d, p) type of stripping occurs with deuterons having energies smaller than or comparable with the Coulomb barrier of the target nucleus, it is often called the Oppenheimer-Phillips process.

For the use of the term stripping in chemical technology, see **Distillation.**

STROBOSCOPE. An instrument that permits intermittent observations of a cyclically moving object in such a way as to produce an optical illusion of stopped or slowed motion. This phenomenon is readily apparent, for example, when rewinding a tape at many revolutions per minute when the tape deck is located under a 60-Hz incandescent lamp. Patterns on the reel tend to slow and then appear to stop before reversing their direction. Of course, stroboscopic effects have been known for decades,[1] one of the first scientific applications being found in very high-speed photography. Intermittency of observation can be provided by mechanical interruption of the line of sight (as with a motion picture camera) or by intermittent illumination of the object being viewed. The industrial stroboscope basically is a lamp plus the electronic circuits required to turn the lamp on and off very rapidly, at rates as high as 150,000 flashes per minute and higher.

The schematic diagram of an electronic stroboscope is shown in the accompanying illustration. The device includes a strobotron tube with its associated discharge capacitors, a triggering tube to fire the strobotron, an oscillator to determine the flashing rate, and a power supply. With the use of harmonic techniques, speeds up to 1 million r/min can be measured. Accuracy is nominally $\pm 1\%$ of the dial reading after calibration.



Electronic stroboscope, (*a*) Schematic representation of circuit. (*b*) Images obtained at harmonic and subharmonic flashing rates of a stroboscope. Even with an asymmetrical object, the correct fundamental image is repeated when the stroboscope is flashing at one-half, one-third, and so on, the speed of the object. The proper setting for a fundamental speed measurement is the highest setting at which a single stationary image can be achieved. This does not hold, however, if the fundamental is beyond the flashing rate of the stroboscope. There are several ways to distinguish fundamental from submultiple images. The flashing rate can be decreased until another single image appears. If this occurs at half the first reading, the first reading was the actual speed of the device. If it occurs at some other value, then the first reading was a submultiple. Or the user can double the flashing rate and check for a double image. Or the user can flip the range switch to the next higher range. Because of the 6:1 relationship between ranges, a 6:1 pattern should appear. The 6:1 relationship between ranges also makes it convenient to convert speed readings from revolutions per minute into cycles per second. One simply flips to the next lower range and divides the new reading by 10.

Because of their portability and easy setup, stroboscopes find a variety of applications, principally in machine and vehicle research, development, and testing. (*GenRad.*)

To serve as a tachometer, a stroboscope must have its own flashing-rate control circuits and calibrated dial. Stroboscope tachometer test disks are available. These disks can be cut out and mounted on light cardboard or metal. The center must be carefully located and fitted onto the drive shaft. Although more automatic means are available to measure belt slippage, this was commonly accomplished by stroboscopes in earlier times.

[1]Invented independently by Stampfer of Vienna and Plateau of Ghent in 1832. Stampfer chose the name "stroboscope," which is derived from the Greek words meaning "whirling watcher."

STROMATOLITE. A term that has been generally applied to variously shaped (often domal), laminated, calcareous sedimentary structures formed in a shallow-water environment under the influence of a mat or assemblage of sediment-binding blue-green algae that trap fine (silty) detritus and precipitate calcium carbonate and that commonly develop colonies or irregular accumulations of a constant shape, but with little or no microstructure. It has a variety of gross forms, from near-horizontal to markedly convex, columnar, and subspherical. Stromatolites were originally considered animal fossils, and although they are still regarded as fossils because they are the products of organic growth, they are not fossils of any specific organism, but rather consist of associations of different genera and species of organisms that can no longer be recognized and named or that are without organic structures. An excellent treatise on stromatolites is "Stromatolites" (M. R. Walter, editor), Elsevier, New York, 1976.

STROMATOLITH. A term proposed by Foye, in 1916, for banded gneisses composed of alternate layers of igneous and metamorphic (schistose) rocks.

STRONTIANITE. The mineral strontianite is strontium carbonate, $SrCO_3$, usually occurring in whitish-yellow or whitish-green masses of radiated acicular crystals, or in fibrous or granular form. When distinctly crystallized it is obviously orthorhombic, but such crystals are rare. It has a nearly perfect prismatic cleavage; uneven fracture; brittle; hardness, 3.5; specific gravity, 3.785; luster, vitreous; color, as above, also green, gray and colorless; streak, white; transparent to translucent. Strontianite occurs in veins chiefly in limestones, occasionally in the crystalline rocks, and usually associated with calcite and celestite. It is found in the metalliferous veins in the Harz Mountains and Saxony. It is commercially important in Westphalia where it is mined for use in the beet sugar industry. In the United States, crystalline masses and geoges of strontianite are found in Schoharie County, New York, long a famous locality for this mineral.

STRONTIUM. Chemical element, symbol Sr, at. no. 38, at. wt. 87.62, periodic table group 2, mp 769°C, bp 1384°C, density 2.54 g/cm^3 (20°C). Below 215°C, elemental strontium has a face-centered cubic crystal structure; between 215–605°C, a hexagonal close-packed crystal structure; and above 605°C, a body-centered cubic crystal structure.

Strontium is a silver-white metal, soft as lead, malleable, ductile, oxidizes rapidly on exposure to air, burns when heated in air emitting a brilliant light and forming oxide and nitride, reacts with H_2O yielding strontium hydroxide and hydrogen gas. Discovered by Hope and by Klaproth in 1793, and isolated by Davy in 1808.

There are four stable isotopes, ^{84}Sr and ^{86}Sr through ^{88}Sr, and seven known radioactive isotopes, ^{82}Sr, ^{83}Sr, ^{85}Sr, and ^{89}Sr through ^{92}Sr, all with relatively short half-lives measurable in hours or days except ^{90}Sr which has a half-life of about 26 years. The latter isotope represents a hazard from nuclear blasting activities because of its long half-life, tendency to contaminate food products, such as milk, and retention in the body. See also **Radioactivity.** In terms of abundance, strontium is 21st among the elements occurring in the rocks of the earth's crust. In terms of the content of sea water, the element ranks 11th, with an estimated 38,000 tons of strontium per cubic mile (9120 tons/cubic kilometer) of seawater. First ionization potential 5.692 eV; second, 10.98 eV. Oxidation potentials $Sr \rightarrow Sr^{2+} + 2e^-$, 2.89 V; $Sr + 2OH^- + 8H_2O \rightarrow Sr(OH)_2{\cdot}8H_2O + 2e^-$, 2.99 V. Other important physical properties of strontium are given under **Chemical Elements.**

Occurrence and Characteristics. Strontium occurs chiefly as sulfate (celestite, $SrSO_4$) and carbonate (strontianite, $SrCO_3$) although widely distributed in small concentration. The commercially exploited deposits are mainly in England. The sulfate or carbonate is transformed into chloride, and the electrolysis of the fused chloride yields strontium metal.

As is to be expected from its high oxidation potential (2.89 V) strontium, like calcium and barium, reacts readily with all halogens, oxygen and sulfur to form halides, oxide and sulfide. See also **Celestite; Strontianite.** In all its compounds it is divalent. It reacts vigorously with H_2O to form the hydroxide, displacing hydrogen and it forms a hydride with hydrogen. Strontium hydroxide forms a peroxide on treatment with

H_2O_2 in the cold. Strontium exhibits little tendency to form complexes; the ammines formed with NH_3 are unstable, the β-diketones and alcoholates are not well characterized, and the chelates formed with ethylenediamine and related compounds are the only representatives of the type. Common compounds of strontium are the following:

Strontium acetate, $Sr(C_2H_3O_2)_2$, white crystals, soluble, formed by reaction of strontium carbonate or hydroxide and acetic acid.

Strontium carbide (acetylide), SrC_2, black solid, formed by reaction of strontium oxide and carbon at electric furnace temperature; the carbide reacts with water yielding acetylene gas and strontium hydroxide.

Strontium carbonate, $SrCO_3$, white solid, insoluble ($K_{sp} = 9.4 \times 10^{-10}$), formed (1) by reaction of strontium salt solution and sodium carbonate or bicarbonate solution, (2) by reaction of strontium hydroxide solution and CO_2. Strontium carbonate decomposes at 1,200°C to form strontium oxide and CO_2, and is dissolved by excess CO_2, forming strontium bicarbonate, $Sr(HCO_3)_2$, solution.

Strontium chloride, $SrCl_2 \cdot 6H_2O$, white crystals, soluble, formed by reaction of strontium carbonate or hydroxide and HCl. Anhydrous strontium chloride, $SrCl_2$, absorbs dry NH_3 gas.

Strontium chromate, $SrCrO_4$, yellow precipitate ($K_{sp} = 3.75 \times 10^{-5}$) formed by reaction of strontium salt solution and potassium chromate solution.

Strontium cyanamide, $SrCN_2$, formed with the cyanide, $Sr(CN)_2$, by heating strontium carbide at 1,200°C with nitrogen.

Strontium hydride, SrH_2, white solid, formed by heating strontium metal or amalgam in hydrogen gas at 250°C. Is reactive with H_2O, yielding strontium hydroxide and hydrogen gas.

Strontium nitrate, $Sr(NO_3)_2$, white crystals, soluble, formed by reaction of strontium carbonate or hydroxide and HNO_3.

Strontium oxide, SrO, white solid, mp about 2,400°C, reactive with H_2O to form strontium hydroxide ($K_{sp} = 3.2 \times 10^{-4}$); strontium peroxide, $SrO_2 \cdot 8H_2O$, white precipitate, by reaction of strontium salt solution and hydrogen or sodium peroxide, yields anhydrous strontium peroxide SrO_2, upon heating at 130°C in a current of dry air.

Strontium oxalate, SrC_2O_4, white precipitate ($K_{sp} = 5.6 \times 10^{-8}$) formed by reaction of strontium salt solution and ammonium oxalate solution.

Strontium sulfate, $SrSO_4$, white precipitate ($K_{sp} = 3.2 \times 10^{-7}$), formed by reaction of strontium salt solution and H_2SO_4 or sodium sulfate solution, insoluble in acids. On heating with carbon strontium sulfate yields strontium sulfide, SrS, while on boiling with sodium carbonate solution, $SrSO_4$ yields strontium carbonate.

Strontium sulfide, SrS, grayish-white solid (thermodynamic K_{sp} 500) reactive with water to form strontium hydrosulfide, $Sr(SH)_2$, solution. Strontium hydrosulfide is formed (1) by reaction of strontium sulfide and H_2O, (2) by saturation of strontium hydroxide solution with H_2S. Strontium polysulfides are formed by boiling strontium hydrosulfide with sulfur.

Stephen E. Hluchan, Business Manager, Calcium Metal Products, Minerals, Pigments & Metals Division, Pfizer Inc., Wallingford, Connecticut.

Editor's Note re Strontium Isotope Research At any given time, the Sr isotope composition in seawater is uniform throughout the ocean because the oceanic residence time of Sr (5 million years) exceeds the mixing time of the oceans (~ 1000 years). However, over geologic time, the $^{87}Sr/^{86}Sr$ ratio in seawater has varied as the result of fluxes of Sr to the oceans from various sources. These would include submarine hydrothermal activity, fluxes from rivers, and submarine recycling, the latter occurring by limestone recrystallization and erosion of ancient sedimentary carbonate. J. Hess and colleagues (University of Rhode Island) reported in 1986 that the seawater Sr isotope composition appears to be a smoothly varying function of time and can be useful for high-precision correlations of oceanic sediments for certain periods of time. These researchers prepared a detailed record of the Sr isotope ratio during the last 100 million years by measuring this ratio in well over a hundred foraminifera samples. Sample preservation was evaluated from scanning electron microscopy studies, measured Sr/Ca ratios, and pore water Sr isotope ratios. Results show that the marine Sr isotope composition can be used for correlating and dating well-preserved authigenic marine sediments throughout much of the Cenozoic to a precision of ± 1 mil years. See also **Condrite.**

In 1990, R. C. Capo and D. J. DePaolo (University of California, Los Angeles and Berkeley, respectively) reported that "marine carbonate samples indicate that during the past 2.5 million years the $^{87}Sr/^{86}Sr$ ratio of seawater has increased by 14×10^{-7}. The high average rate of increase of this ratio indicates that continental weathering rates were exceptionally high. Nonuniformity in the rate of increase suggests that weathering rates fluctuated by as much as ±30 percent of present-day values. Some of the observed shifts in weathering rate are contemporaneous with climatic changes inferred from records of oxygen isotopes and carbonate preservation in deep sea sediments."

Studies of Metamorphism. As reported by J. N. Christensen, J. L. Rosenfield, and D. J. DePaolo (University of California, Berkeley), "Measurement of the radial variation of the $^{87}Sr/^{86}Sr$ ratio in a single crystal from a metamorphic rock can be used to determine the crystal's growth rate. Such variation records the accumulation of ^{87}Sr from radioactive decay of ^{87}Rb (rubidium) in the rock matrix from which the crystal grew. This method can be used to study the rates of petrological processes associated with mountain building." This methodology has been applied by the researchers mentioned to the study of the rates of tectonometamorphic processes from rubidium and strontium isotopes in garnet."

Isotopic Tests for Upwelling Water. In studies of the Yucca Mountain, Nevada, area as a potential site for a high-level nuclear waste repository, the area has been aggressively scrutinized geologically for possible upwelling of deep-seated waters. Strontium and uranium isotopic compositions of hydrogenic materials were used by scientists J. S. Stuckless, Z. E. Peterman, and D. R. Muhs (U.S. Geological Survey, Denver, Colorado) to assist in confirming other geological methods. Their findings indicated in 1991 that the vein deposits are isotopically distinct from groundwater in the two aquifers that underline Yucca Mountain, thus indicating that the calcite could not have precipitated from groundwater and thus providing evidence against upwelling water at the site.

Additional Reading

Capo, R. C., and D. J. DePaolo: "Seawater Strontium Isotopic Variations from 2.5 Million Years Ago to the Present," *Science*, 51 (July 6, 1990).

Christensen, J. N., Rosenfeld, J. L., and D. J. DePaolo: "Rates of Tectonometamorphic Processes from Rubidium and Strontium Isotopes in Garnet," *Science*, 1405 (June 21, 1989).

Hess, J., Bender, M. L., and J.-G. Schilling: "Evolution of the Ratio of Strontium-87 to Strontium-86 in Seawater from Cretaceous to Present," *Science*, **231,** 979–983 (1986).

Macdougall, J. D.: "Seawater Strontium Isotopes, Acid Rain, and the Cretaceous-Tertiary Boundary," *Science*, 485 (January 29, 1988).

Meyers, R. A: "Handbook of Chemicals Production," McGraw-Hill, New York, 1986.

Sax, N. R., and R. J. Lewis, Sr.: "Dangerous Properties of Industrial Materials," 8th Edition, Van Nostrand Reinhold, New York, 1992.

Staff: "ASM Handbook—Properties and Selection: Nonferrous Alloys and Pure Metals," ASM International, Materials Park, Ohio, 1990.

Staff: "Handbook of Chemistry and Physics," CRC Press, Boca Raton, Florida, 1992–1993.

Stuckless, J. S., Peterman, Z. E., and D. R. Muhs: "U and Sr Isotopes in Ground Water and Calcite, Yucca Mountain, Nevada: Evidence Against Upwelling Water," *Science*, 551 (October 25, 1991).

STRUCTURAL GEOLOGY. That branch of geology which deals with the form, arrangement, and internal structure of the rocks, and especially with the description, representation, and analysis of structures, chiefly on a moderate to small scale. The subject is similar to tectonics, but the latter is generally used for the broader regional or historical phases. Structural petrology is the study of the internal structure or fabric of a rock, commonly with the aim of clarifying the rock's deformational history. (*American Geological Institute*)

STRUT. A structural member subjected to compression. Its conditions of loading and analysis are the same as for a column. If there is any difference between strut and column, it rests on the following points. A column is usually thought of as being a fairly large compression member, vertical in position. Small columns are frequently called struts; also, struts are compression members which are incorporated into structures in many positions besides the vertical.

STUFFING BOX. A device for preventing leakage or transfer of fluid between moving parts, usually consisting of a relatively soft packing compressed or confined by an adjustable member called a gland. Stuffing box packings differ from gaskets in that they are used in confined spaces, and do not of themselves withstand stresses due to fluid pressure. In the usual form, the stuffing box consists of a hollow cylinder surrounding the moving (reciprocating or rotating rod or shaft) member; the space between the hole and the rod or shaft is filled with packing compressed by the gland. Packings may consist of relatively plastic material, bonding such substances as cotton fabric, rope, or asbestos, but packings of rubber, leather, pressed graphite, or molded plastics are also used.

In high-speed machinery, such as turbines and rotary compressors, where no adequate cooling is available, and in high-pressure equipment where small clearances are required, and for which packings are inadequate, devices known as labyrinths may be used. A labyrinth consists of a series of projections on the rotating element, running in close contact with grooves on the stationary element. To enable a labyrinth to function as a seal, there must be some fluid flow; the fluid first passes a restriction and then expands into a chamber, which consumes a certain amount of energy. After a series of such expansions, a considerable pressure drop will exist between the initial and terminal points of the labyrinth. In order to render the device operative, there must always be some leakage at the terminal point. The effective operation of a labyrinth is less marked for liquids than for vapors.

STURGEONS (*Osteichthyes*). Of the order *Chondrostei*, family *Acipenseridae*, sturgeons are fishes of moderate to large size, found in the oceans and fresh waters of the northern hemisphere. Their chief external characteristic is the series of bony plates arranged in rows along the back and sides, separated by wide spaces containing only small hard elements. See accompanying diagram. The sturgeons are considered excellent food fishes and are the source of caviar. There are about fifteen European species and nine North American species.

Sturgeon.

Sturgeons have poor vision, but this is partly compensated by fleshy whiskers which trail in the sand and assist in locating food at the bottom. Rare among the fishes, the sturgeons have taste buds external of the mouth. It is believed that these also assist in locating a food supply. Also quite rare among fishes is the fact that sturgeons eat quite slowly. The diet includes crawfish, insect larvae, snails and some small fish.

The largest of the sturgeons is the *Huso huso* (giant beluga) found in the Black and Caspian seas and the Volga River. Records indicate the largest of these to weigh 2860 pounds (1297 kilograms) with a length of 28 feet (8.4 meters). The age was unknown. A 75-year old sturgeon was weighed at 2200 pounds (998 kilograms) length 13 feet (3.9 meters). The Eurasian species frequently weigh in excess of a half-ton. The marine sturgeon *Acipenser sturo* is found in the temperate waters of the Atlantic, usually weighs about 500 pounds (227 kilograms) as an adult and attains a length of about 10 feet (3 meters). Marine sturgeons also prefer the muddy bottom and diet on shrimp, clams, worms, crustaceans, and small fishes.

The largest of American fresh water fishes is the *Acipenser transmontanus* (Pacific coast white sturgeon). They usually weigh about 300 pounds (136 kilograms) but records indicate weights up to 1300 pounds (590 kilograms) and more. The *Acipenser medirostris* is also a sturgeon of the American Pacific and averages about 350 pounds (159 kilograms) length of about 7 feet (2.1 meters).

Sturgeons are probably best known for the caviar which is processed from the roe. Caviar production is no longer limited to Eastern Europe, but has been undertaken in other areas, including the United States. In the interest of conservation, eggs can be stripped from the living fish (somewhat similar to practices in trout hatcheries). Various means are used to free the eggs from the egg membranes. The eggs are then placed in a brine which extracts the liquid from the eggs. After draining the brine, the eggs can be packed and sold commercially.

STYLET. Small sharp structures used for piercing. The term applies to the calcareous spines of the proboscis of some nemertine worms and to the slender piercing organs of some insect mouths. The latter are modified mandibles and maxillae.

STYLOLITE. A columnal-like structure, at right angles, or highly inclined to the bedding planes of certain limestones, believed to be produced by differential vertical movements induced under great pressure. The term is derived from the Greek words meaning column and a stone.

STYRENE-MALEIC ANHYDRIDE. A thermoplastic copolymer made by the copolymerization of styrene and maleic anhydride. Two types of polymers are available—impact-modified SMA terpolymer alloys (*Cadon*®) and SMA copolymers, with and without rubber impact modifiers (*Dylark*®). These products are distinguished by higher heat resistance than the parent styrenic and ABS families. The MA functionality also provides improved adhesion to glass fiber reinforcement systems. Recent developments include terpolymer alloy systems with high-speed impact performance and low-temperature ductile fail characteristics required by automotive instrument panel usage.

Copolymers show chemical resistance generally similar to that of polystyrene and terpolymers similar to that of ABS (acrylonitrile-butadiene-styrene). Neither type is recommended for use in strongly alkaline environments. All impact versions have good natural color and products are available in a wide range of colors. Copolymer crystal grades have good clarity and gloss.

Glass-reinforced SMA polymers are used as electrical connectors, consoles, top pads, and as supports for urethane-padded instrument panels. There are several additional automotive uses. SMA are also found in coffee makers, steam curlers, power tools, audio cassette components, business machines, vacuum cleaners, solar heat collectors, electrical housing, and fan blades, among others.

SUBACUTE SCLEROSING PANENCEPHALITIS. This is a form of subacute encephalitis affecting children and young adults; it is due to persistent infection with the measles virus. SSPE and inclusion-body encephalitis are now recognized as the same disease. It most frequently follows by about six years measles contracted at an earlier age and is twice as common in boys as girls. The same vaccination programs which have controlled measles in the United States have also reduced the incidence of SSPE. In the United Kingdom, some twenty cases are notified annually.

It is not known whether persistence of the virus in the brain is the result of an abnormal incomplete form of the virus or of immunodeficiency on the part of the host.

Onset of the syndrome is usually insidious with loss of energy and interest. After some time increasing clumsiness draws attention to the organic nature of the disease and at this stage involuntary movements or myoclonus usually appear. Visual signs may be prominent with chloroidoretinal scarring present in about 30% of cases. Further progression is marked by intellectual deterioration, rigidity, and spacticity, and increasing helplessness with death in a year or two. The disease may be arrested in a state of complete incapacity or rarely some improvement may occur, but not full recovery. No treatment is known to be effective.

R. C. V.

SUBARCTIC PACIFIC WATER. An oceanic water mass extending from the Aleutians on the north to the North Pacific Central Water masses on the south, and far out into the Western Pacific. It is a major surface water mass of low temperature (2–4°C (35.6–39.2°F) at 50° north latitude) and low salinity, 32.2–34.1%, the last figure existing at its lower depths (400–500 meters) (1320–1650 feet). Its eastern part, which extends to the coast of America is higher in temperature and salinity, due to effects of warm air from the land.

SUBGRAPH. A graph containing a subset of the edges of the original graph.

The *complement of a subgraph* is the set of all elements of the graph not in the subgraph.

A *connected graph* is essentially one that is in "one piece." An unconnected graph must therefore be decomposable into several *connected* "pieces." This notion can be made precise by using the notion of a maximal connected subgraph.

Let two vertices β_1 and β_2 of a graph G be defined to be equivalent if there exists a path between them and denote this relation symbolically by $\beta_1 \doteq \beta_2$. Clearly, if

(a) $\beta_1 \doteq \beta_2$ and $\beta_2 \doteq \beta_3$, then $\beta_1 \doteq \beta_3$

(b) and, for the same vertex, $\beta_1 \doteq \beta_1$

(c) If $\beta_1 \doteq \beta_2$, then $\beta_2 \doteq \beta_1$

Thus the relation is an equivalence relation in the strict mathematical sense because it is (a) transitive, (b) reflexive and (c) symmetric. The collection of all vertices of G is partitioned into disjoint equivalence classes such that two vertices belong to the same class if and only if they are equivalent. If G is finite, the number of these classes is also finite and can be enumerated $S_1, S_2, \ldots, S_P$. A little thought reveals that the set of all edges whose two endpoints are vertices in $S_r (r = 1, 2, \ldots, P)$ is a *connected subgraph* G_r of G. Moreover $G_r (r = 1, 2, \ldots, P)$ is a *maximal connected subgraph* of G in the sense that the addition of any more vertices to G_r renders it unconnected. Evidently, $P = 1$ if and only if G is connected.

A *proper subgraph* is a subgraph which does not contain all the edges of the graph.

Disjoint subgraphs are subgraphs of a graph which have no vertices in common.

See also **Graph (Mathematics).**

SUBLIMATION. The direct transition, under suitable conditions, between the vapor and the solid state of a substance. If solid iodine is placed in a tube and slightly warmed, it vaporizes and the vapor reforms into crystals on the cooler parts of the tube. Many crystalline substances, both metallic and non-metallic, may be similarly sublimated in a vacuum; fairly large crystals of selenium have been thus prepared. The most familiar sublimates are frost and snow. As in the case of other changes of state, sublimation is accompanied by the absorption or evolution of heat, the quantity of which per unit mass is called the heat of sublimation of the substance. At pressures near the triple point the heat of sublimation is approximately equal to the sum of the heats of fusion and vaporization. In physical and chemical literature, it is customary to regard as sublimation only the transition from solid to vapor, not from vapor to solid; but meteorologists do not make this distinction.

Sublimation plays a major role in the freeze-drying of foods. See also **Freeze-Drying.**

SUBLIMATION (Heat of). The quantity of heat required at constant temperature (and pressure) to evaporate unit mass of a solid. In sublimation, the change is directly from solid to vapor, without appearance of the liquid phase.

SUBROUTINE (Computer System). 1. The set of instructions necessary to direct a computer to carry out a well-defined mathematical or logical operation. 2. A subunit of a routine. A subroutine is often written in relative or symbolic coding even when the routine to which it belongs is not. 3. A portion of a routine that causes a computer to carry out a well-defined mathematical or logical operation. 4. A routine which is arranged so that control may be transferred to it from a master routine and so that, at the conclusion of the subroutine, control reverts to the master routine. Such a subroutine is usually a closed subroutine. 5. A single routine may simultaneously be both a subroutine with respect to another routine and a master routine with respect to a third. Usually control is transferred to a single subroutine from more than one place in the master routine and the reason for using the subroutine is to avoid having to repeat the same sequence of instructions in different places in the master routine.

Closed Subroutine. A subroutine not stored in the main path of the routine. Such a subroutine is entered by a jump operation and provision is made to return control to the main routine at the end of the operation. The instructions related to the entry and re-entry function constitute a linkage.

Open Subroutine. A subroutine of which a replica must be inserted at each place in the computer program at which the subroutine is used. Although requiring more storage, this approach avoids linkage and housekeeping overhead.

T. J. H.

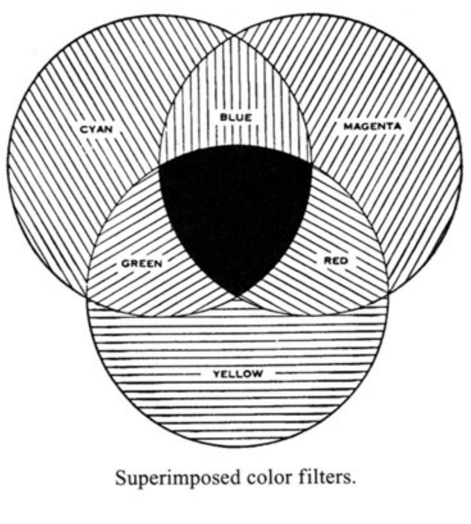
SUBSIDENCE. Subsiding air is sinking air and is associated with lateral divergence. Subsidence in the atmosphere is a stabilizing influence; it also decreases relative humidity within the sinking air as it warms the air. Atmospheric pressure usually rises under the influence of subsidence which is normally associated with anticyclones. Clear or partially clouded skies are the usual weather in a region of subsidence.

SUBTRACTION. An operation that is the inverse of addition. The $-$ sign is used to denote the operation of subtraction. In the equation $M - S = R$, M is the *minuend*, S the *subtrahend*, and R is the *remainder*. To perform the operation, calculate $M + (-S) = R$, i.e., change the sign of the subtrahend and add.

SUBTRACTIVE COLOR PROCESS. A method of photographic color synthesis using two or more superimposed colorants which selectively absorb their complementary colors from white light.

Most modern processes of color photography make use of a subtractive synthesis to yield prints or transparencies. In a three-color process, the colorants cyan, magenta, and yellow are used to control the amounts of red, green and blue in a beam of white light. (See accompanying figure.) This beam of white light may be either that of a projector with its color transparency, or the light reflected from a white support, such as paper, on which the color reproduction is printed. In the first case, the light passes through the colorants once, while in the print viewed by reflection the light must traverse the colorants twice.

The colorants are positive or negative images. A cyan positive image (a cyan colorant controls red light), for example, may be prepared



Superimposed color filters.

from the negative that recorded the red present in the subject. The magenta and yellow colorant images are likewise made from green and blue record negatives. These colorant images are superimposed in register to yield the final reproduction. The three colorants may be in separate removable layers or they may be physically inseparable as in the modern integral tripacks. The contrast of the colorant images must be approximately double for a picture to be viewed by transmitted light as compared to one to be viewed by reflection. The accuracy of color reproduction by a subtractive synthesis as compared to an additive is chiefly dependent on how satisfactorily the three colorants cyan, magenta, and yellow fulfill their role, as red, green and blue absorbers respectively. Color correction is often adopted to improve the accuracy of reproduction when using the colorants generally available.

See also **Photography and Imagery.**

SUCCESSION (Plant). The gradual replacement of one plant association by another. Succession is caused by slow changes in the environmental factors which influence the establishment, development and survival of plants. Many of these changes, such as shading, increase in the humus content or porosity of the soil, decrease in soil temperature, etc., are brought about by plants themselves.

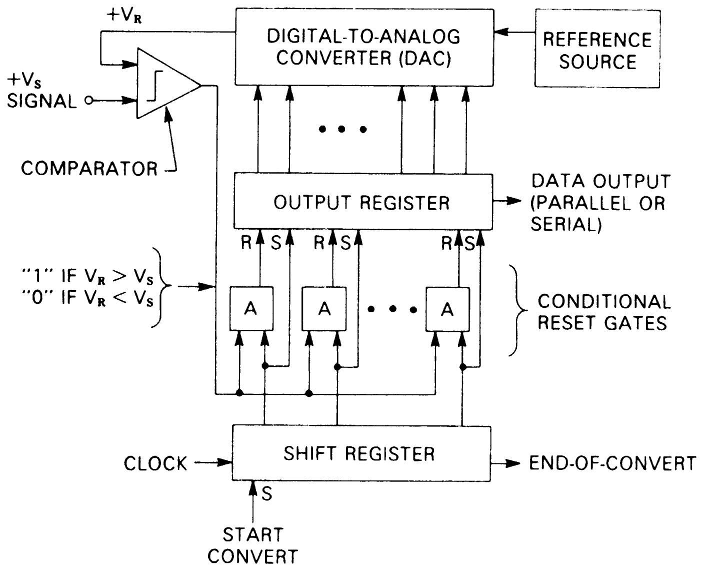
A good example of a plant succession occurs in the filling of a shallow pond. At first the deeper parts of such a pond are occupied only by submersed aquatic species of plants. As the pond becomes shallower, as a result of the accumulation of plant remains and the deposition of the silt caught in such remains, floating-leaf species such as water lilies invade the area and largely shade out the underwater species. With further decrease in the depth of the water the floating leaf aquatics are replaced by cattails, reeds, bulrushes and other typical marsh plants. At a still later stage in succession the former pond will be occupied by a wet meadow composed largely of sedges, rushes, and spike rushes. Still later swamp shrubs such as willows, alders, and buttombush invade the area only to be replaced in turn by trees of species which can tolerate a poorly drained substratum. Still further successional stages may ultimately result in the development of a climax association on the area.

Another example of plant succession which can be observed in many parts of the country is the natural reforestation of abandoned farm land. The pioneer invaders of such an area are mostly herbaceous plants, the very first to appear being mostly annuals. Soon shrubs invade the area to be followed in turn by certain species of trees. The first tree species to occupy the area, however, seldom represent a stable plant association. In large parts of the eastern United States, for example, pines of one species or another are usually the first kind of tree to occupy abandoned fields, but they are in time replaced by other species, most commonly by oaks or hickories or both. In many regions still further successional stages will be passed through before a stable climax association is attained.

Other examples of kinds of areas on which plant succession is initiated are sand dunes, burned over forest lands, cliffs or raw talus slopes, and the bottoms of lakes which have been exposed by drainage. Important consideration should be given to the principles of plant succession in forest and range management and in other land utilization problems.

SUCCESSIVE-APPROXIMATION A/D CONVERTER. This type of analog-to-digital converter makes a direct comparison between an unknown input signal and a reference signal. The converter, as shown schematically in the accompanying diagram, includes a digital-to-analog (D/A) Converter. This provides an output voltage V_R, a precise fraction of the reference source voltage. The setting of analog switches, which are controlled by the digital outputs of the output register, determine the attenuation factor of the converter. Digital-output information of the A/D converter also is stored in the output register. Two input AND logic gates are used as conditional-reset gates. The shift register also functions as a step counter.

The A/D converter shown in the diagram is a binary *n*-bit converter. A "start convert" pulse causes all registers to clear and sets a 1 into the first position of the shift register. In turn, this action also turns on the first switch in the A/D converter and sets the first stage of the output



Successive-approximation analog-to-digital converter.

register to 1. The output of the D/A converter is equal to one-half the full-scale D/A converter output voltage inasmuch as the D/A converter is binary. The D/A output voltage is compared directly with signal voltage V_S by the comparator. The output of the latter is a binary 1 if V_R is greater than V_S and a binary 0 if V_R is less than V_S.

The first bit in the output register may be reset to 0 or may remain as a 1, depending upon the comparator output. Upon appearance of the next clock pulse, the 1 in the shift register moves to the second bit at which instant the comparison process is repeated—with the exception that the output of the D/A converter is either 0.25 or 0.75 of the full-scale D/A converted output. This is dependent upon whether the first bit was reset to 0 or remained as a 1, respectively.

Codes other than binary can be used in successive-approximation A/D converters. What is required is the production of an attentuation that is proportional to the weighting value of the code digit. A binary-coded decimal converter can be constructed by using a binary-coded decimal digital-to-analog converter. The D/A converter will provide attentuation ratios of 0.8, 0.4, 0.2, 0.1, 0.08 and so on. Of course, appropriate coding of the output register is necessary so that the information can be stored in coded form.

Successive-approximation A/D converters can be used at conversion speeds up to approximately 100,000 samples/second at resolutions up to 16 bits (not including sign). At lower resolution, speeds over 250,000 samples/second are practical. Factors to be considered in the design and application of these A/D converters include: (1) stability and regulation of the reference source, (2) overload and recovery characteristics of the comparator, (3) analog-switch characteristics, and (4) speed and response of the ladder network.

Thomas J. Harrison, International Business Machines Corporation, Boca Raton Florida.

SUCCULENTS. Succulent plants are those which are fleshy, that is, have their stems or leaves greatly enlarged to serve as water-storage organs. Succulents occur wild in regions in which there is a very limited supply of available water, as in the desert lands of western America, central Asia, and many parts of Africa. Among stem succulents the outstanding representatives are the members of the Cactus family, with the milkweed and spurge families also ranking high. Leaf succulents include many plants such as *Sedums*, *Crassulas*, *Bryophyllums*, and *Gasterias*. Many of the leaf succulents are plants of extremely curious habit. In some, the leaves are so swollen with stored water that they form a nearly spherical mass growing almost buried in the ground; in others the masses of leaves form compact rosettes at the tips of the branches; while in others the swollen leaves appear like small green beads along a slender stem.

SUCKERS (*Osteichthyes*). Of the group *Cypriniformes*, family *Catostomidae*, suckers live near the bottom of streams, feeding on vegetation and small animals. The mouth is usually provided with fleshy lips and in some species opens at a downward angle. They are not highly valued as food fishes, but the common sucker (*Catostomus commersoni*) often abundant in small streams and lakes is considered of excellent taste although bony.

Buffalo fishes are in this family and have the appearance of a goldfish or carp. The largemouth buffalo (*Ictiobus cyprinellus*) which can attain a length up to 3 feet (0.9 meters) is common throughout the central portion of the United States. *Carpiodes cyprinus* (quillback carpsucker) will often measure up to 26 inches (66 centimeters) in length and weigh up to about 12 pounds (5.4 kilograms). This fish is considered a commercial item in Lake Erie and is also found in the eastern and central regions of the United States.

Suckers are essentially an American fish. Other species include: *Catostomus latipinnis* (flannelmouth sucker) found in the Colorado River; *Minytrema melanops* (spotted sucker) found in the eastern states; *Xyaruchen texanus* (humpback sucker) found in the Colorado River Basin and attains a length up to 2 feet (0.6 meter); and *Myxocyprinus asiaticus*, similar to the *Xyaruchen*, which is one of the few suckers found outside of America, in Asia.

Suckers are sometimes difficult to differentiate from minnows. A method is described in a book by Hubbs and Lagler, "Fishes of the Great Lakes Region," Cranbrook Institute of Science, 1958.

SUDDEN INFANT DEATH SYNDROME. Known as SIDS (sometimes crib or cot death), this syndrome may be defined as the precipitous (sudden), unexpected death of an apparently healthy infant from whom an autopsy fails to identify the cause of death. In other words, these infants die of mysterious causes, which most experts as of the late 1980s consider a puzzle that, to date, has defied statistical research. There may be preventive measures, but physicians question—on the basis of what?

Of each one thousand live births in modern industrialized nations, 990 infants can be expected to live to their first birthday (and normally beyond for many years). Of the ten babies who die before they are one year old, six do not survive the first month of life—succumbing to gross errors in growth and development, usually the result of a very low birth weight, prematurity, or some congenital defect, i.e., causes of death that are explainable. Two or three of the remaining four infants will die in hospital some time during the first year, usually as the result of a serious treatment-resisting infection or because of a late-detected birth abnormality. The remaining two or three infants will die *unexpectedly*, often in their own cribs, usually discovered too late to take to hospital, but even then, death will usually occur and for which a thorough necropsy examination will not yield a plausible cause of death. Even though the number two or three out of a sample of 1000 infants is small statistically, nevertheless in a large nation, there will be several thousand cases of SIDS. (Estimated to be 10,000 cases/year in the United States and the United Kingdom, collectively.)

Golding, Limerick, and Macfarlane in a 1986 book on SIDS (see references) describe a typical SIDS situation: "Three-month-old David's mother is a nurse in her early 30s. He is her third child; happily married to a physician, she does not smoke. Through normal birth she has just mothered a normal, contented baby. At two months, he was screened at the health clinic; no abnormality at all. At three months, not long before his first immunization, she put him in his crib for his afternoon sleep. Warm but not encumbered, he went right to sleep. She heard no sound, but in 30 minutes she went up to check on him. He was very pale; when she picked him up, he was floppy and around his mouth was a blood-tinged froth. She cleared the tiny mouth, phoned for an ambulance and tried resuscitation. Within 10 minutes David was in hospital, still warm but without breath or heartbeat. The emergency staff gave up after an hour's effort. The postmortem was entirely normal except for some vomit in the infant's upper mouth and airways. Certificate of Death—Sudden Infant Death (SID)."

Studies on infant mortality in efforts to relate SIDS statistically with one or more causative factors have been conducted in Britain, Canada, the United States, Australia, New Zealand and in some European countries as well as in Japan and Israel. Statistics thus far have shown that, in contrast with other causes of infant fatality (notable decreases in most countries in recent years), the incidence of SIDS has held steady. (Only exceptions reported are in two large Swedish cities and these have not been explained satisfactorily.) Over the past 15 years, there have been at least five international conferences on the subject.

Although no convincing cause of SIDS is now under present consideration, many hypotheses have been offered over the years, only to be disproven by statistics. In a recent study in the United Kingdom in a medically well-tended region with about 14,000 births per year, some very general criteria have been offered, but these do not contain much help in terms of SIDS prevention.

- Age of mother—SIDS is relatively higher for young mothers, notably those in the lower economic sector of society and who smoke or use drugs. (Yet for one-half of cases, the mothers were not under economic stress and one-third did not smoke.)
- Season of year—SIDS occurs more frequently during winter months.
- Male babies are at higher risk of SIDS than females.
- Babies of exceptionally low birth weight are more susceptible to SIDS. (But, 80% of babies in the sample weighed a normal amount at birth.)
- Siblings of dead infants do not appear to be at much greater risk.
- Infanticide is seldom masked as SIDS.

Some of the past proposed hypotheses that have essentially been ruled out by most authorities today include: (1) sudden arrest of breathing—failure in bioelectronic development; (2) use of the birth control pill; (3) vaccines and medicants normally administered to infants; (4) selenium and other trace element deficiencies in the diet; (5) overheating, perhaps augmented by fever, causing death by heat stroke (febrile convulsions are common in young children, but occur at ages far advanced of the peak of SIDS); (6) vague muscular effects. A point that has added some confusion to the diagnosis of SIDS (as defined here) is respiratory syncytial virus (common virus infection leading to bronchitis and pneumonia)—because this condition matches that of the usual period for SIDS and also has the same ratio of more male than female infants. Similar viruses may show similar patterns. Some authorities believe that anaphylactic shock to the sensitized infant in reaction to respiratory virus infection may precipitate SIDS, but they cautiously suggest that this could account for only some SIDS cases as currently understood. The strongest statistical connection for the virus hypothesis is that SIDS occurs more frequently in winter than in summer.

The real puzzler of SIDS is the usual lack of evidence at post mortem examination.

Updated Appraisals of SIDS

In an 1991 study by J. S. Kemp and B. T. Thach (Washington University of Medicine, St. Louis), data on 25 cases of SIDS were analyzed carefully. Efforts were made to distinguish the difference on postmortem examination of (a) accidental suffucation and (b) SIDS. The researchers point out that three basic assumptions generally can be used to separate the two causes of death: (1) Healthy infants will not suffocate on ordinary bedding. (2) Infants will suffocate only if their heads are firmly restrained when the airway is occluded. (3) Infants lying with their faces straight down and their noses and mouths pressed into the bedding can turn their heads to obtain access to fresh air. In the 25% or more of infants with SIDS who are found with their faces straight down, the posture is thus considered coincidental and unrelated to the cause of death.

The Washington University study produced one very important finding—namely, that the deaths of several of the previously diagnosed SIDS cases actually were deaths from suffocation that occurred in a manner not previously reported in infants. The deaths all occurred on a particular type of cushion filled with polystyrene beads and marketed for infants.

To support their study, the researchers experimented by measuring the mechanics (simulated by machine) involved in breathing with head down on a pliable material. Thus it was found that an infant lying face down on certain materials may, in fact, rebreathe expired gases. In the tests, gas concentrations were measured. Carbon dioxide gas was used in the simulations. Thus, simulated infants with faces pressed down on certain materials resulted not in a sudden cessation of breathing, but in breathing high concentrations of carbon dioxide. Not only is air flow through such materials impeded, but the carbon dioxide content rose.

In mid-1991, the researchers concluded: "Accelerated suffocation by rebreathing was the most likely cause of death in most of the 25 infant cases studied. Consequently, there is a need to reassess the cause of death in the 28 to 52 percent of the victims of SIDS who are found with their faces straight down. Safety regulations setting standards for softness, maleability, and the potential for rebreathing are needed for infant bedding."

Obviously, the foregoing findings cannot explain all cases of SIDS, and the findings have not gone unchallenged. See Balding reference listed.

Another avenue of research on SIDS includes that of a medium-chain acyl coenzyme A dehydrogenase deficiency. See Dino reference listed.

The unusual breathing habits of seals also is being investigated for possible leads to causes of SIDS in humans. It has been known for decades that the seal has the ability to hold its breath for long periods of time. Seals halt breathing for several minutes when diving and also when they sleep. A form of sleep apnea in seals is positive in that this slows down their metabolism and conserves energy in the form of blubber. It has been found, however, that breath-holding in baby seals will cause random fluctuations in heart rate. Part of the baby seal's learning process is that of finding how to stabilize its heartbeat during apnea. Researcher M. Castellini (University of California, Santa Cruz) is studying sleep apnea in baby seals in an effort to learn how they overcome the early heart-rate problem; also, this knowledge may provide insights on human SIDS. Another group at the University of Pennsylvania is studying a similar phenomenon that occurs in bulldogs.

Additional Reading

Balding, L. E.: "SIDS and Suffocation," *N. Eng. J. Med.*, 1806 (December 19, 1991).

Ding, J-H, Roe, C. R., and A. K. Iafolla: "Medium-Chain Acyl-Coenzyme A Dehydrogenase Deficiency and Sudden Infant Death," *N. Eng. J. Med.*, 61 (July 4, 1991).

Flam, F.: "The SIDS-Seal Connection," *Science*, 1613 (June 21, 1991).

Gilbert, E. F., and K. Kenison: "Fetal Hemoglobin Levels in SIDS," *N. Eng. J. Med.*, 1281 (November 1, 1990).

Golding, J., Limerick, S., and A. Macfarlane: "Sudden Infant Death: Patterns, Puzzles, and Problems," University of Washington Press, Seattle, Washington, 1986.

Kemp, J. S., and B. T. Thach: "Sudden Death in Infants Sleeping on Polystyrene-Filled Cushions," *N. Eng. J. Med.*, 1858 (June 27, 1991).

Lagercrantz, H., and T. A. Slotkin: "The 'Stress' of Being Born," *Sci. Amer.*, 100–107 (April 1986).

Long, W., et al. "A Controlled Trial of Synthetic Surfactant in Infants Weighing 1250 g or more with Respiratory Distress Syndrome," *N. Eng. J. Med.*, 1696 (December 12, 1991).

McCormick, M. C.: "The Contribution of Low Birth Weight to Infant Mortality and Childhood Morbidity," *N. Engl. J. Med.*, **312**(2), 82–90 (January 10, 1985).

Perlman, J. M., and F. Moya: "Synthetic Surfactants in Infants with Respiratory Distress Syndrome," *N. Eng. J. Med.*, 1703 (June 18, 1992).

Ziai, M., Ed.: "Pediatrics," 3rd Ed., Little, Brown & Co., Boston, Massachusetts, 1984.

SUFFICIENT STATISTIC. A sufficient statistic is one which summarizes all the information in the sample concerning the relevant parameter. Thus if t_0 is a sufficient estimator of a parameter θ, and t_1 is an alternative estimator calculated from the same sample, a combination of t_0 and t_1 (such as their mean) is no better as an estimate of θ than t_0 alone. A sufficient statistic is necessarily efficient.

SUGARCANE. A tall tropical perennial grass (*Saccharum officinarium*) and the source of over half of the world's supply of sucrose (table sugar, saccharose), the major sweetener for foods and bevarages. The remaining sucrose comes from the sugar beet (*Beta vulgaris*) and the grain sorghum. Honey and maple trees furnish a small amount of sugar.

Sugarcane grows from 5 to 15 feet (1.5 to 4.5 meters) in height, with stalks that range from 1 to 2 inches (2.5 to 5 centimeters) in diameter. The sugarcane stalk is made up of a series of nodes and internodal sections, ranging from 1 to 10 inches (1.5 to 25 centimeters) in length. At each node there is a vegetative bud and root primordia that will "germinate" or sprout to produce a new shoot and nodal roots when placed in the proper environment. The saccharine content of the nodes is very low, whereas the internodes contain most of the sugar. As the plant commences to grow, leaves are formed on alternate sides of the nodes, reaching a length of some 3 feet (0.9 meter). At the base, the leaves are about 2 inches (5 centimeters) wide and taper gradually throughout their length to a sharp point at the end. The plant begins to display nodes and internodes when only a few weeks old. Early in the life of the plant, these parts are green, but after a while some of these parts may change color, depending upon variety. Although some may continue green, others will take on a striped pattern or a yellow-green or purple coloration. If disease-free, maturity of the plant is first indicated by yellowing and dropping of the lower leaves. Although the bottom part of the plant may be quite barren, the upper stalk will be green and vigorous. In some regions, frost may halt maturity, but if permitted to mature this will take the form of development of flowers and seeds. From 12 to 15 months are required to achieve full maturity by some varieties and in some regions, notably in the more tropical areas.

Sugarcane is not an agricultural crop that is likely to diminish over the years ahead, even though the role of sucrose in food products has been the subject of penetrating analysis and criticism during the past few years—critiques that go far beyond the more traditional negative roles of sucrose in terms of dental caries. The emergence of corn (maize) and other sugar sources also has posed a serious threat. Sugar is frequently the first substance to be attacked when obesity is mentioned, and dieting (even though usually not of a long-term nature or a permanent accomplishment) is almost a status symbol in some countries. Sugar has been identified as an ugly "empty calorie." Nevertheless, despite many of the very well justified critiques of sugar excesses in various diets, population increases in some countries, coupled with increased consumption of snacks and convenience foods in other countries, has enabled sugar to maintain a relatively high volume of production.

But, even if improvements in nutrition ultimately pose a relatively serious threat to the continued high production of sugarcane and sugar beets, it is now believed by many authorities that sugarcane production, in particular, will increase severalfold during the next decade or two—not as a food product, but rather as a raw material for the chemical industry, replacing to some extent the present declining raw materials for the petrochemical industry.

Efforts of the Brazilian government to produce alcohol from sugarcane and blend that with hydrocarbons for automotive fuel are exemplary of the important role that sugarcane can play as an industrial chemical raw material.

SUINES *(Mammalia, Artiodactyla)*. Pigs, boars, hogs, and peccaries *(Suines)* comprise one of the more primitive groups of the order *Artiodactyla* (even-toed hoofed animals). The group is not large in terms of identifiable species in the wild, as contrasted, for example, with the antelopines and even the bovines, but because they are so highly valued in terms of the domestic breeds, they do comprise an extremely large population group of mammals.

Pigs are native to the warmer parts of Europe and Asia, the Oriental region, and Africa. None are attributed to originating in North or South America, although the closely related Peccaries have inhabited these continents for many, many centuries. Pigs are not ruminants, i.e., they do not chew their cuds. They all dig with their muzzles and have a preference in nature for vegetable matter. They are characterized by having large litters. There are certain misconceptions concerning pigs that should be clarified. In general, these animals are not at all stupid as sometimes portrayed, but are very intelligent and rate only second to the Great Apes among the mammals, except humans, for their mental abilities. By nature, they are not dirty. When in the wild and left to their own habits, these animals are exceptionally orderly and clean. Thus, likening a person to a pig really should not be considered uncomplimentary. It is true that most species of *Suines* enjoy mud wallows where they go to cool off, to remove external parasites from their bodies, and to cleanse themselves. Mud, as some beauticians acclaim, is an excellent cleaner, containing helpful antibiotics. The pig in captivity that wallows in mud that is littered with excrement, rotting garbage, etc., does not do this out of choice. In the wild, these animals never excrete in their mud wallows.

In connection with the domesticated pigs, certain terminology often requires clarification. The terms *pig* and *swine* are generally considered synonymous in most parts of the world, but in the United States, swine usually refers to pigs that are under three months of age. In England, Canada, and Australia, pigs are swine of any age or weight. Informally, untamed, wild pigs are sometimes referred to as *wild swine*. Adult animals are often referred to as hogs, particularly the marketable and commercial animals. Hog also refers to a castrated boar. A *gilt* is a young female pig, often referring to an animal with its first litter. A *sow* is an adult female pig. A *boar* is a well-developed male used for breeding service. A *boar pig* is a male animal under breeding age (usually less than six months old). A *stag* is a male animal that has been castrated in maturity—after the tusks, shields, enlarged sheath, crest, and other characteristics have developed. A *barrow* is a male animal that has been castrated before sexual characteristics have developed. A *shote* is an immature animal of either sex.

There are over 400 breeds of domestic pigs. Some of the more important are listed here. *Berkshire* (Fig. 1), originated in England and Japan. It is of medium size, black, with white on face, legs, and tail tip. *Chester White* (Fig. 2) originated in Chester County, Pennsylvania. The animal has a pink skin and is highly regarded for its lard. *Duroc* originated in New York State, is of red coloration and considered an excellent pig for large-scale production. *Essex* is black with white saddle on shoulder, legs, nose, and tail. *Gloucestershire* originated in England, is lop-eared, white with black spots, and is highly considered for bacon and cross breeding. *Hampshire* is derived from the Saddleback and appears much like it; it is highly regarded for bacon. *Hereford* is red with white head, a good producer. *Iberian* is a red-colored animal with short ears; it resembles pigs of medieval times. It is of no commercial importance outside Spain and Portugal. *Landrace* is a white animal with lop ears, popular in Denmark, Germany, Norway, and Sweden. This breed was exported from Sweden to England in 1953. It is a highly regarded breed. The bacon from these animals is often called Wiltshire bacon. *Cornwall* is an all-black animal with lop ears; it originated in England and is mainly raised for bacon and cross-breeding. *Mangalitza* is a white pig, is much like the Chester White, but smaller and dishfaced. *Pietrain* originated in Belgium, is lop-eared and off-white color. This breed has attracted much attention since 1960 and essentially has been confined to Belgium, France, England, Denmark, and the Netherlands. *Poland China* (Fig. 3) originated in Ohio (Warren and Butler counties). It is a very large animal, black with white face, feet, and tail tip, and drooping ears. *Spotted Poland China* originated in Indiana. It is much like the Poland China except with white spots over body. It is an excellent lard producer. *Tanworth* originated in England and is found in Canada, Australia, and New Zealand. It is golden red with long snout and erect ears. Its major use is for bacon and cross breeding. *Welch* is white with lop ears; excellent for bacon; found in Africa, France, Poland, and the former U.S.S.R. *Wessex Saddleback* is black with white saddle. It originated in England; is extensively produced in Australia. Major uses are for bacon and crossing with white boars. It was introduced into the United States several years ago as a special breed. *West French White* is lop-eared, white; a major pork producer. *Yorkshire* is one of the most widely distributed of all breeds and a good producer. It is a white animal with erect ears. Other important breeds include the Cheshire, Duroc-Jersey, Large Black, and Razorback. Breeders look for: (1) prolific sows that farrow and raise large litters; (2) animals that grow rapidly and that show economical gain during feeding periods; and (3) animals that are resistant to infection and parasites and that have carcasses that withstand handling, transporting, and processing well. In particular, boars are sought that are prolific, and have a high degree of masculinity, good disposition, and an ability to sire strong, healthy pigs. In selecting sows, the length of the body and femininity are important, particularly a well developed udder and two rows of teats with a minimum of six teats in a row.

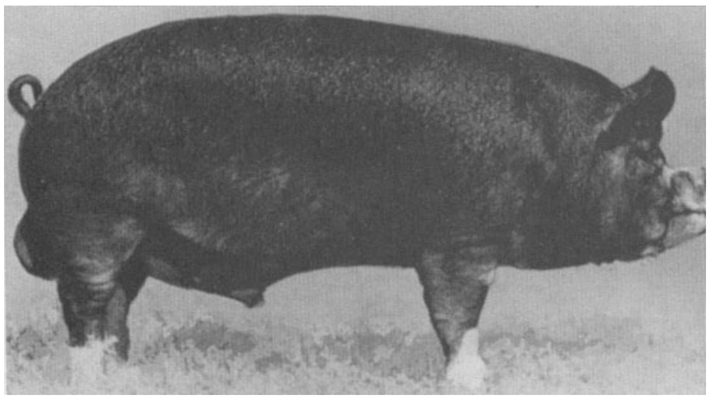


Fig. 1. Berkshire boar. *(USDA.)*

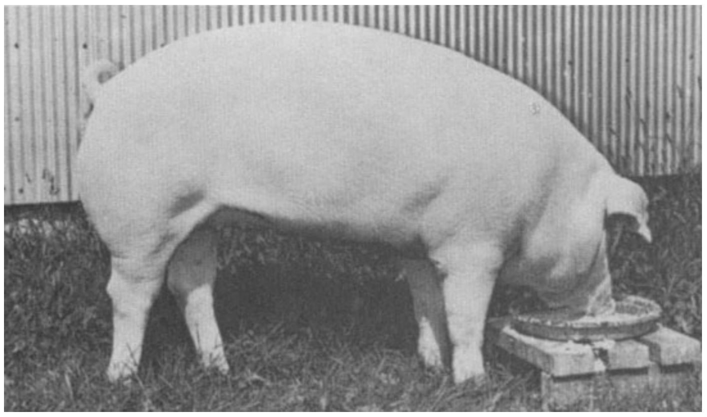


Fig. 2. Chester White gilt. *(USDA.)*

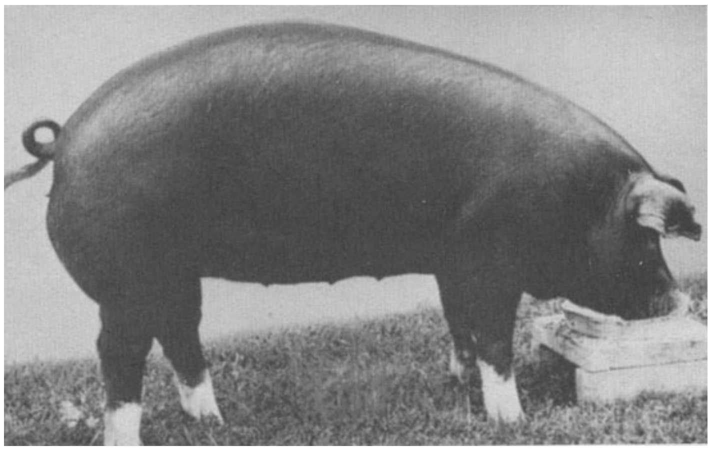


Fig. 3. Poland china gilt. *(USDA.)*

Young boars may breed at about 8 months of age. They usually will be mated to 20 to 30 sows during their first breeding season. Mature boars will be mated with from 40 to 60 sows per season. The sow's period of estrus (heat) is about 3 days, occurring every 3 weeks. Mating usually occurs on the second or third day. Gestation period is 114 days. A gilt will have her first litter at about one year of age.

Pigs also can be classified by the main type of product for which they are best suited. Lard producers are large of frame. Animals with longer frames are more suitable for bacon. Because of increased demands for lean meat, an animal with a lean carcass is desired. Usually whitepigs require more intensive care than the colored varieties. Most of the white breeds originated in Canada and the United States. They usually are excellent lard producers.

Of the Eurasian Pigs found in the wild state, most species are found in the Oriental Region. However, the Wild Boar, of which there are many races, occurs in western Europe, north Africa, across India and to southeast and central Asia. However, the animal is not found north of the Caspian Sea, in the U.S.S.R., or the great Asian mountain chains. They are characterized by long, pointed muzzles. Coloration is red-brown-chocolate-black. The hair is stiff and bristly. The amount of hair varies with climate. Some authorities consider the bite of the wild boar the worst of any mammal with the exception of the Killer Whale. The bite is more of a ripping than a slicing nature and is considerably worse than that inflicted by the Great Cats. The smaller Pigmy Pig is found in the Himalayas. The mature boars stand only about 1 foot (0.3 meter) high at the shoulders. On the other hand, the Giant Bornean Pig will attain a length of some 6 feet (1.8 meters), with very large heads. They are characterized by an upturned moustache.

The African Bush-Pigs and Forest-Hogs are found widely throughout that continent. The bush-pigs are known for their rooting abilities. These animals like to turn over fallen logs for the snakes, rats, snails, and fungi found thereunder. It is interesting to note that the Giant Forest-Hog was not discovered by naturalists until the present century. It is a giant, low-slung animal with big head and widespread ears. Warts are located below the eyes. The body is covered sparsely with long, stiff bristles which, as the animal matures, could almost be described as spines.

The Wart-Hog is well known for its ugly appearance, characterized by a large flattened head out of which grow most unattractive warts. They are known for their erratic behavior. The Babirusa of the East Indies is a nocturnal, forest-loving beast. The males have pronounced teeth. They are known for their excellent flesh.

It is believed that the peccaries migrated to North America when there was a land bridge with Asia. The collared peccary or muskhog ranges from the southwestern United States to southern South America, and the white-lipped peccary is found only from Belize to Paraguay. The latter species is gregarious, living in large bands. Its vicious nature makes it dangerous to encounter, although a single animal is too small to trouble a human being. The collared peccary lives singly or in small groups and is inoffensive. The collared peccary attains a length of about 3 feet (1 meter) and a height of close to 16 inches (40 centimeters).

The early Chinese in America are accredited with bringing the breed of domesticated true pig to the New World. This animal readily takes to the wilds and breeds with wild species. This gave rise to the feral or so-called razorback hogs of the American south.

More details on the suines can be found in the "Foods and Food Production Encyclopedia," (D. M. Considine, editor), Van Nostrand Reinhold, New York, 1982.

SULFONE POLYMERS. Polysulfone is a transparent, heat-resistant, ultrastable and high-performance engineering thermoplastic. It is amorphous and has low flammability and smoke emission. Electrical properties are good; the material remains essentially unchanged up to near its glass transition temperature, 190°C (374°F). The molecular structure of polysulfone features the diaryl sulfone group, a group that tends to attract electrons from the phenyl rings. Oxygen atoms para to the sulfone group enhance resonance and produce oxidation resistance. High resonance also strengthens the bonds spatially, fixing the grouping into a planar configuration. Thus, the polymer has good thermal stability and rigidity at high temperatures. Ether linkages provide chain flexibility, thus imparting good impact strength.

The resistance to acids, alkalies, and salt solutions is high and also good in terms of detergents, oils, and alcohols even at elevated temperatures under moderate stress. Polysulfones, however, are attacked by polar organic solvents, such as ketones, chlorinated hydrocarbons, and aromatic hydrocarbons. The material can be used continuously in steam up to temperatures of 93°C (300°F). Maximum stress in water at about 82°C (180°F) is 2000 psi (steady loads) and 2500 psi (intermittent loads). In long-term performance at 150°C (300°F), polysulfone increases about 10% in strength and modulus values, retaining 90% of its dielectric strength and 70% of its impact strength.

Polysulfone is widely used in medical instrumentation and trays for holding instruments during sterilization. Food processing applications, such as piping, scraper blades, steam tables, microwave oven cookware, and beverage dispensing tanks, are numerous. Electrical/electronic applications include connectors, automotive fuses and switch housings, soil bobbins and cores, television components, capacitor film, and structural circuit boards. In chemical processing equipment, uses include corrosion-resistant piping, tower packing, pump parts, filter modules, and membranes. Polysulfone is available in both molding and extrusion grades. A special medical grade is available. Also available are polysulfone compounds with glass fiber or beads, as well as fillers, such as *Teflon*®.

SULFONAMIDE DRUGS. In 1935, Domagk, a German researcher, was the first to observe the clinical value of *prontosil*, a red compound derived from azo dyes. Paraaminobenzenesulfonamide was shown to be the effective portion of the prontosil molecule. This substance was given the name *sulfanilamide*. This was the first of a group of related drugs to receive wide clinical trial. It was found to be effective in the treatment of hemolytic streptococcal and staphylococcal infections. Within a short span of years, related drugs were synthesized and given clinical trials. These included *sulfapyridine*, *sulfathiazole*, *sulfaguanidine*, *sulfadiazine*, and *sulfamerazine*. These drugs acted by inhibiting the growth of bacteria rather than by killing organisms.

Even though numerous adverse side effects were observed over a period of time, the sulfonamides played an important role in medicine prior to the advent of the antibiotics. In recent years, the importance of the so-called *sulfa drugs* has diminished considerably, but for certain situations they are still considered important antimicrobials. Presently the sulfonamides are mainly used to treat uncomplicated urinary tract infections, including prostatitis, due to *E. coli*. They are also used to treat a number of noncardial infections. At one time the sulfa drugs were widely used in the treatment of meningococcal meningitis and bacillary dysentery. Unfortunately, the bacilli responsible for these diseases developed, over the years, a resistance to the drugs, severely reducing their efficacy.

Within the last few years, some new sulfa drugs have been introduced, including trimethoprim-sulfamethoxazole. This drug has broadened the scope in treatment of urinary tract infections derived from species in addition to *E. coli*, namely, *Klebsiella*, *Enterobacter*, and *Porteus* species. This drug also is used for the treatment of acute otitis media in children, particularly those instances where strains of *H. influenzae* and *streptococcus pneumoniae* may be suspected. The drug is also used to treat systemic infections that may arise from chloramphenicol- and ampicillin-resistant *Salmonella*; as well as infections attributed to *Pneumocystis carinii*.

Also, the nature of sulfonamide compounds (relatively short duration of action, capability of entering into synergism with other drugs, poor absorbability, and topical effectiveness, not to mention relatively low cost) is taken advantage of in what is sometimes called short-acting sulfonamides. Short-acting sulfonamides include sulfisoxazole, sulfadiazine, and trisulfapyrimidines. An intermediate-acting sulfonamide in current use is sulfamethoxazole. This drug does tend to cause renal damage arising from sulfonamide crystalluria.

Sulfacetamide eyedrops continue to be used for treatment of superficial ocular infections. Sometimes silver-sulfadiazine cream is applied to burn surfaces to minimize or prevent bacterial growth, as well as preventing invasive infection.

The adverse effects of sulfonamides include hypersensitivity reactions, as manifested by rashes, photodermatitis (allergic reaction to light), so-called drug fever, nausea, and vomiting. These reactions occur with some frequency when sulfonamides are administered. Less frequently encountered is crystalluria, previously mentioned, but with the risk lessened in the case of sulfisoxazole. Sulfa drugs also occasionally cause hemolytic anemia, agranulocytosis, and kernicterus (in infants) when the drugs are given to nursing mothers. In rare instances, sulfa drugs may precipitate hepatitis, aplastic anemia, renal tubular necrosis, and certain blood disorders.

SULFUR. Chemical element, symbol S, at. no. 16, at. wt. 32.064, periodic table group 16, mp 112.8°C (rhombic), 119.0°C (monoclinic), 120.0°C (amorphous), bp 444.7°C (all forms), sp gr 2.07 (rhombic), 1.96 (monoclinic), 2.046 (amorphous). Atomic weight varies slightly because of naturally occurring isotopes 32, 33, 34, and 36, the total possible variation amounting to ± 0.003.

The stable isotopes of sulfur are ^{32}S, ^{33}S, ^{34}S, and ^{36}S. There are three known radioactive isotopes, ^{31}S, ^{35}S, and ^{37}S, with ^{35}S having the longest half-life (87.1 days). See also **Radioactivity.** Electronic configuration $1s^2 2s^2 2p^6 3s^2 3p^4$. Ionic radius S^{2-} 1.855Å, S^{6+} 0.29Å (Pauling). Covalent radius 1.07Å. In terms of abundance, sulfur ranks fourteenth among the elements occurring in the earth's crust, with an estimated 520 grams per metric ton. In seawater, the element ranks fifth, with an estimated 894 grams per metric ton.

First ionization potential 10.357 eV; second, 23.3 eV; third, 34.9 eV; fourth, 47.08 eV; fifth, 63.0 eV; sixth 87.67 eV. Oxidation potentials $H_2S(aq) \rightarrow S + 2H^+ + 2e^-$, -0.141 V; $H_2SO_3 + H_2O \rightarrow SO_4^{2-} + 4H^+ + 2e^-$, -0.20 V; $S + 3H_2O \rightarrow H_2SO_3 + 4H^+ + 4e^-$, -0.45 V; $SO_3^{2-} + 2OH^- \rightarrow SO_4^{2-} + H_2O + 2e^-$, 0.90 V; $S^{2-} \rightarrow S + 2e^-$, 0.508 V; $HS^- + OH^- \rightarrow S + H_2O + 2e^-$, 0.478 V. Other important physical properties of sulfur are given under **Chemical Elements.**

Sulfur has a large number of allotropes. The ordinary form, α-sulfur, is rhombic having a crystal unit cell composed of sixteen S_8 molecules. At 95.5°C it undergoes transition to β-sulfur, which is monoclinic and also has a molecular weight (in solution in carbon disulfide) corresponding to S_8. Four other monoclinic forms have been identified microscopically: γ-sulfur, prepared by heating α-sulfur to 150°C, cooling to 90°C, and inducing crystallization by friction, ρ-sulfur, S_6, prepared by extracting an acidulated sodium thiosulfate solution with toluene, as well as υ-sulfur, and δ-sulfur. There is also a tetrahedral form, θ-sulfur, crystallized from a carbon disulfide solution of rhombic sulfur treated with balsam. The first liquid form to appear is λ-sulfur, a pale yellow liquid, obtained on heating sulfur to 120°C. Above 160°C, this form changes to a viscous, dark-brown liquid consisting mainly of μ-sulfur. A third liquid allotrope, π-sulfur is considered to exist in molten sulfur, in equilibrium with the other two forms, having its greatest concentration at about 180°C. Sulfur vapor has been shown to contain S_8, S_6, S_4, and S_2 molecules. Several other allotropes of sulfur have been produced, including two paramagnetic forms, purple and green in color, by low-temperature processing.

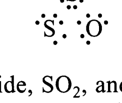
Sulfur occurs as free sulfur in many volcanic districts, and may have been formed in part by sublimation, by decomposition of hydrogen sulfide, or metallic sulfides, or by organic agencies. It is often associated with limestones and gypsum. Sulfur is found in Spain, Iceland, Japan, Mexico, and Italy. It occurs especially in Sicily, which was the producer for the world until about the beginning of the twentieth century, when Herman Frasch, by inventing the superheated-water method of mining sulfur, made available the great Louisiana and Texas deposits. This method of mining is at the same time a method of purifying sulfur, because in the process of heating, accompanying materials remain unmelted at the temperature at which sulfur melts and is drawn off. In the Louisiana and Texas deposits the sulfur is associated with gypsum, occurring in the caprock overlying the salt plugs that have pierced the strata underlying the Gulf coastal plain. In the United States, sulfur is also found in California, Colorado, Nevada, and Wyoming. Sulfur also occurs as (1) sulfides, e.g. cobaltite, iron disulfide, pyrite, FeS_2, lead sulfide, galenite, PbS, copper iron sulfide, copper pyrite, $CuFeS_2$, zinc sulfide, zinc blende, ZnS, mercury sulfide, cinnabar, HgS; and (2) as sulfates, e.g., calcium sulfate, gypsum, $CaSO_4{\cdot}2H_2O$, barium sulfate, barite, $BaSO_4$. Several of these minerals are described under separate alphabetical entries.

Sulfur Production and Use: The manufacture of H_2SO_4 accounts for nearly 90% of all sulfur consumed. Of this, about 50% of the H_2SO_4 goes into fertilizer production, nearly 20% into chemical manufacture, 5% into pigments, about 3% each for iron and steel production and the manufacture of rayon and synthetic films, and about 2% for various petroleum processes. The balance of over 15% of H_2SO_4 is consumed by a large number of other industries, this all giving credence to the use of H_2SO_4 production figures as an overall economic index. The 10% of the sulfur not going into H_2SO_4 is converted into numerous chemicals that are consumed by a variety of industries, the largest among these being pulp and paper production and the manufacture of carbon disulfide.

Sulfur Compounds: In addition to the compounds described in the following paragraphs, see also **Hydrogen Sulfide, Mercaptans, Sodium Thiosulfate, Sulfuric Acid, Sulfurous Acid, Thiocyanic Acid, Thioethers, Thiophene** and **Thiourea.**

Sulfur-Oxygen Compounds: Due to its $3s^2 3p^4$ electron configuration sulfur, like oxygen, forms many divalent compounds with two covalent bonds and two lone electron pairs, but *d*-hydridization is quite common, to form compounds with oxidation of 4+ and 6+.

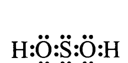
A number of suboxides of sulfur have been reported, but in general their composition has not been clearly established. Polysulfur oxides of formula $S_{8-16}O_2$ are formed by reaction of hydrogen sulfide and sulfur dioxide. Also, when sulfur is burned with oxygen in very limited supply disulfur monoxide, S_2O is formed. This has the structure



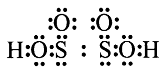
A mixture of sulfur dioxide, SO_2, and sulfur vapor, at low pressure and with an electric discharge, forms sulfur monoxide, SO. Its presence is shown from its absorption spectrum, but upon separation it disproportionates at once to sulfur and SO_2. Sulfur sesquioxide, S_2O_3, is formed by reaction of powdered sulfur with anhydrous SO_3; S_2O_3 also disproportionates (at 20°C in nitrogen) to sulfur and SO_2. Sulfur dioxide, SO_2, is formed by the combustion in air or oxygen of sulfur and sulfur compounds generally, except those in which sulfur is in a higher state of oxidation. Sulfur dioxide has an O—S—O bond angle of 119.5°. The sigma bonds utilize essentially sulfur *p* orbitals, with *dp* hybridization for the pi bonds. Its oxidation to sulfur trioxide, SO_3, by atmospheric oxygen attains a significant rate only at higher temperatures, but can be materially increased by catalysts. Sulfur trioxide is also evolved from oleum on heating. It exists in the vapor state chiefly as the planar monomer, in which the oxygen atoms are spaced symmetrically (120° angles) about the sulfur atom, and it has S—O bond lengths of 1.43Å. Liquid SO_3 is partly trimerized, and exists in three physical forms.

Sulfur tetroxide is formed by reaction of pure oxygen and sulfur dioxide under the silent electric discharge. It is not obtained pure, but in a variable SO_3/SO_4 ratio, and as a polymerized white solid. Another peroxide, $(SO_2OOSO_2O)_x$, which is written as S_2O_7, is known.

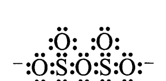
Of the 16 oxyacids of sulfur that are recognized, only four have been isolated. The more important oxyacids of sulfur are: (1) Thiosulfurous acid, $H_2S_2O_2$, structure not established, existing only in compounds, an oxidizing agent for Fe^{2+}, H_2S and HI; (2) Sulfoxylic acid, H_2SO_2, existing only in salts and other compounds, e.g., $ZnSO_2$, SCl_2, $S(OR)_2$, structure probably



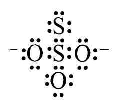
(3) Dithionous acid (or hydrosulfurous acid), $H_2S_2O_4$, existing only in compounds, widely used reducing agent, chiefly as the sodium salt, for organic substances, also reduces Sb^{3+}, Ag^+, Pb^{2+}, Cu^{2+} to the elements, structure



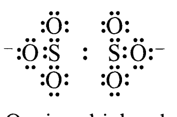
(4) Sulfurous acid, H_2SO_3, produced by hydration of SO_2, not isolated but existing in many salts, the sulfites and acid sulfites, and many organic compounds, including the dialkyl or diaryl sulfites and the alkyl or aryl sulfonic acid esters, which suggests two possible structures, $(HO)_2SO$ and $H(HO)SO_2$, although the acid dissociation constants (first, 1.25×10^{-2}, and second, 5.6×10^{-8}) suggest the structure with only one unhydrogenated oxygen atom. Sulfurous acid and sulfites are fairly strong reducing agents, but the HSO_3^- ion may act as an oxidizing agent, as for formates and related compounds. Other compounds of SO_2 are the metabisulfites or pyrosulfites, containing the ion



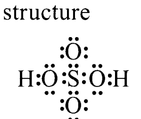
which enters into equilibrium with water to form acid sulfite. (5) Thiosulfuric acid, $H_2S_2O_3$, existing only in compounds, the anion having the structure



and widely used as a coordinating ion for forming complexes with metals; it also is an oxidizing agent, and is used in iodometric titrations. (6) Dithionic acid, $H_2S_2O_6$, existing only in compounds but stable in dilute solution at room temperature, and differing in its stability to hydrolysis and oxidation from the polythionates,



(7) Polythionic acids, $H_2S_nO_6$, in which n has values of 3, 4, 5, 6 and others, some of which have been reported to have values indefinitely high (20–80), structure not established, though there is evidence that they consist of two sulfonic acid groups connected by a linear chain of sulfur atoms. An interesting property of the polythionates that are very rich in sulfur ($n > 20$) is their slight tendency to decompose to give free S. (8) Sulfuric acid, H_2SO_4, structure



strong acid, formed by hydration of sulfur trioxide, completely dissociated (first ionization) in aqueous solutions up to 40%; above that concentration dissociation decreases and hydrate formation occurs. Both normal and acid sulfates are formed by metallic elements, though the products of their direct reaction with the acid vary with temperature. (9) Sulfuric acid dissolves SO_3, the product of a 1:1 ratio being pyrosulfuric or disulfuric acid. $H_2S_2O_7$, which forms the pyrosulfates, also obtainable by heating acid sulfates, structure HO(O)(O)SOS(O)(O)OH. Two series of alkali metal pyrosulfates are known: those formed from SO_3 and the metal sulfates and those formed from H_2SO_4 and the metal sulfates, which have the pyrosulfuric acid structure. (10) Peroxymonosulfuric acid is produced by addition of SO_3 to concentrated H_2O_2, its salts are fairly stable, and it has the structure HOS(O)(O)OOH. (11) Peroxydisulfuric acid is produced by reaction of concentrated H_2O_2 on H_2SO_4 or by electrolysis of acid sulfate solutions; its salts are fairly stable and it has the structure HOS(O)(O)OOS(O)(O)OH.

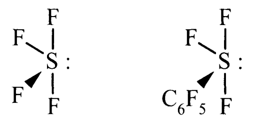
Hydrogen Sulfide: H_2S is a weak acid ($pK_{A1} = 7.00$), ($pK_{A2} = 12.92$) stronger than water but weaker than H_2Se, as expected from its position in the periodic system; its reducing strength exhibits the same relation. Its long use in analytical chemistry is due to the differential solubility of many sulfides with variation of the pH of an aqueous solution. Hydrogen persulfide, H_2S_2, structure HSSH, with an S—S bond distance of 2.05Å, formed from an alkali metal polysulfide solution and HCl at low temperatures, is the first of a group of hydrogen polysulfides of the general formula H_2S_x.

Sulfur Halides: Many are known. Those that have been identified and whose properties have been determined include the fluorine compounds, S_2F_2, SF_4, SF_6, S_2F_{10}, the chlorine compounds, S_2Cl_2, SCl_2, SCl_4 and the bromine compound, S_2Br_2. Sulfur chlorides of general formula S_nCl_2 are known up to $n \approx 20$. A similar series of cyanides, $S_n(CN)_2$, is known. Derivatives of SCl_4, e.g., SCl_3CN, have been prepared and the list of derivatives of SF_6 is rapidly growing, including S_2F_{10}, $(SF_5)_2O$, $(SF_5)_2O_2$, SF_5Cl, SF_5Cl_3, $SF_4(CF_3)_2$, $(SF_5)_2CF_2$, SF_5OF, SF_5OSO_2F, etc. Derivatives of SF_4 include $C_6H_5SF_3$, $(SF_3)_2CF_2$, etc. All of them except the higher fluorides hydrolyze readily, are essentially covalent in character, and the simple compounds can be prepared directly from the elements, the activity of the halogen determining the product obtained, fluorine yielding SF_6 and S_2F_{10} and the other fluorides being prepared from those, chlorine and bromine yielding the monohalides from which the others are obtained by continued halogenation.

Sulfur Oxyhalides: Four general compositions of oxyhalides of sulfur have been known for many years. In one of these, sulfur has a 4+ oxidation state, the thionyl halides, SOX_2, and in three of which it has a 6+ oxidation state, thionyl tetrafluoride, the sulfuryl and pyrosulfuryl halides, SOF_4, SO_3X_2 and $S_2O_5X_2$, respectively. As is the case for the simple halides, no iodine compounds are known, but polyhalogen ones, such as SOFCl and SO_2FCl exist.

Isolable Oxysulfuranes: Sulfuranes, as described by Musher (1969), are compounds of sulfur(IV) in which four ligands are attached to sulfur and have in common with rare-gas compounds such as XeF_2 an electronic structure involving a formal expansion of the valence shell of the central atom from 8 to 10 electrons. Martin and Perozzi (1976) pointed out that the incorporation of oxygen ligands makes possible a wide range of new structural types that illustrate structure-reactivity relationships in a particularly illuminating way.

For many years, it was postulated that most types of sulfuranes were intermediate (not isolable) compounds. However, the isolable halosulfuranes have been well established for many years. The first known of these, SCl_4, was prepared by Michaelis and Schifferdecker in 1873. In 1911, it was found that SF_4, while highly reactive, was thermally stable. However, the compound was not fully described until 1929. Development of SF_4 led to the creation of a family of stable fluorosulfuranes and their derivatives. It was found that the fluorines in these compounds can be replaced by aryl or perfluoroalkyl groups (Tyczkowski, 1953). Kimura and Bauer (1963) described the geometry of SF_4 as a distorted trigonal bipyramid with two fluorines and a lone pair of electrons occupying equatorial positions, with the other two fluorines in apical positions. The postulated structures of SF_4 (a), and of a derivative (b) are shown below:

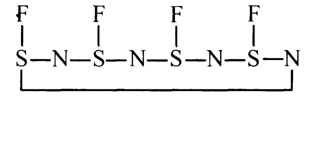


(a) (b)

In the early 1970s, Sheppard, by reacting SF_4 with pentafluorophenyllithium, prepared an isolated sulfurane with four carbon-centered ligands, namely, *tetrakis*-(pentafluorophenyl)sulfurane, $(C_6F_5)_4S$. Martin and Perozzi (1976) prepared the first isolable diaryldialkyloxysulfurane. If it is protected against moisture, the researchers found the compound to be stable over an indefinite period at room temperature. The research in this interesting area continues, some of the details of which are well described by Martin–Perozzi (1976). Summarizing the situation, the researchers observe that the development of synthetic methods for oxysulfuranes has made a wide range of isolable compounds of hypervalent sulfur available for study. Structure-reactivity correlations are now becoming evident as a result of such study. The fact that oxygen is dicoordinate makes it possible to sythesize cyclic oxysulfuranes and to use the pronounced changes of reactivity which accompany cyclization to design new, potentially useful sulfurane reagents stable enough to allow isolation.

Sulfur-Nitrogen Compounds: Many of the sulfur-nitrogen compounds are sulfuric acid derivatives. Three of these compounds correspond to replacement of the hydrogen atoms of ammonia with one, two and three—SO_3H radicals, the monosubstituted compound being aminesulfonic (sulfamic) acid, and being readily separated, the others known only in their salts, the aminedisulfonates (imidodisulfonates) and aminetrisulfonates (nitrilotrisulfonates). Other amines, such as hydroxylamine and hydrazine have similarly related compounds. (See **Hydrazine; Hydroxylamine.**) Diamino derivatives of the sulfoxy acids are also known, such as sulfamide, $H_2NSO_2NH_2$. Imidosulfinamide, HN $(SONH_2)_2$, has been prepared by reaction of $SOCl_2$ and ammonia (also directly from SO_2 and ammonia), and a trimer of sulfimide, $(O_2SNH)_3$, by ammoniation of SO_2Cl_2. It is cyclic in structure, composed of alternate >NH and >SO_2 groups. Nitrosulfonates, containing the ion SO_3NO^- and dinitrososulfonates, containing $SO_3N_2O_2^{2-}$, are also known.

The most important sulfur-nitrogen compound is tetrasulfur tetranitride, S_4N_4, prepared in many ways, including the direct reaction of ammonia and sulfur. All data on its structure are in accord with a puckered eight-member ring, or a cage with N—S connections. The question as to whether there are also transannular N—N or S—S bonds has not been clearly settled. On hydrogenation it adds 4 H atoms, on fluorination it forms $S_4N_4F_4$, structure



and SN_2F_2 the latter reacting with SNF to form SNF_3, structure F_2SNF. Other thiazyl compounds, prepared from S_4N_4 and the halogens or sulfur halides, include $(ClSN)_3$, S_4N_3Cl, S_4N_3Br, S_4N_3I. These last are salts, i.e., $[N_4S_3]X$, and salts of other anions can also be prepared. Other sulfur-nitrogen compounds known are SN_2, S_4N_2, S_5N_2, and S_2N_2, the last being formed by heating S_4N_4.

Thiocyanogen, $(SCN)_2$, is formed by treatment of a metal thiocyanate with bromine in an organic solvent. It reacts with organic compounds in a manner completely analogous to the free halogens, lying between bromine and iodine in oxidizing power. The alkali metal and alkaline earth metal thiocyanates are prepared by fusing the cyanides with sulfur, and the other metal thiocyanates, as well as the organic ones, are usually prepared from the alkali metal thiocyanates. (See entries under Thio-.)

Many selenium analogs of thio compounds can be made, including $SeSO_3$, SO_3Se^{2-}, SSe^{2-}, etc.

In addition to carbon disulfide (odorless when pure), carbon subsulfide, S=C=C=C=S, an evil-smelling red oil and carbon monosulfide, (CS_x), are known as well as COS, CSSe and CSTe. Because of its similarity to oxygen, and the reactivity of its acids, sulfur enters widely into organic compounds.

Biological Aspects of Sulfur: See **Sulfur (In Biological Systems).**

Additional Reading

Dalrymple, D. A., and T. W. Trofe: "An Overview of Liquid Redox Sulfur Recovery," *Chem. Eng. Progress*, 43 (March 1989).

Meyers, R. A.: "Handbook of Chemicals Production Processes," McGraw-Hill, New York, 1986.

Mollare, P. D.: "From Calcasieu to Caminada: A Brief History of the Louisiana Sulfur Industry," *Chem. Eng. Progress*, 73 (March 1989).

Sax, N. R., and R. J. Lewis, Sr.: "Dangerous Properties of Industrial Materials," 8th Edition, Van Nostrand Reinhold, New York, 1992.

Staff: "Handbook of Chemistry and Physics," CRC Press, Boca Raton, Florida, 1992–1993.

Trofe, T. W., Dalrymple, D. A., and F. A. Scheffel: "Stetford Process Status and R&D Needs," Topical Report GRI-87/0021, Gas Research Institute, Chicago, Illinois, 1987.

SULFURIC ACID. Infrequently termed "oil of vitriol," sulfuric acid, H_2SO_4, is a colorless, oily liquid, dense, highly reactive, and miscible with water in all proportions. Much heat is evolved when concentrated sulfuric acid is mixed with water and, as a safety precaution to prevent spluttering, the acid is poured into the water rather than vice versa. Sulfuric acid will dissolve most metals. The concentrated acid oxidizes, dehydrates, or sulfonates most organic compounds, sometimes causing charring. There are numerous commercial and industrial uses for H_2SO_4 and these include the manufacture of fertilizers, chemicals, inorganic pigments, petroleum refining, etching, as a catalyst in alkylation processes, in electroplating baths, for pickling and other operations in iron and steel production, in rayon and film manufacture, in the making of explosives, and in nonferrous metallurgy, to mention only some of its numerous uses. Because of its wide use industrially, some economists over the years have included sulfuric acid consumption among their economic indicators.

Most countries with significant industrial activity and particularly in chemicals production will have significant capacities for making sulfuric acid. In some countries, H_2SO_4 is the leading chemical in terms of tonnage production. Depending upon suppliers, H_2SO_4 is commercially available in a number of strengths, ranging from 77.7% H_2SO_4 (60° Baumé, sp. gr. 1.71) through 93.2% H_2SO_4 (66° Baumé), 98% H_2SO_4, 99% H_2SO_4, and 100% H_2SO_4 (sp. gr. 1.84).

Fundamentally, there are two kinds of sulfuric acid plants: (1) those that use the dry gas (sulfur burning) process; and (2) those that use the wet gas process. In the first type, the raw materials are elemental sulfur and water. In the second type, the sulfur dioxide feed may come from a variety of sources, including metallurgical smelters (copper, zinc, lead, etc.), pyrite roasters, waste acid decomposition furnaces, and hydrogen sulfide burners. In these plants, the SO_2 gas stream enters the acid plant containing a large amount of water vapor. The gas is usually hot (260–430°C) and dusty, and also may contain a number of impurities, such as fluorides, that could harm the catalyst in the contact section of the plant. These incoming gases thus require cooling and purification in the series of scrubbers and electrostatic precipitators, followed by drying prior to entering the contact section of the plant.

In either type of plant, sulfur dioxide is converted to sulfur trioxide in the contact portion of the plant. The reaction $SO_2 + \frac{1}{2}O_2 \rightarrow SO_3$ is effected by passing the SO_2 over a catalyst, usually vanadium pentoxide (V_2O_5). The catalyst in the converter vessel is usually in the form of small pellets and typically arranged in four layers. Provision is made for removal of the heat of reaction after each layer or stage. The catalyst may be used for a number of years with only a very moderate decrease in activity.

From this fundamental point, the sulfuric acid plant designer has a number of alternatives and options to consider. Two factors are of major import in sulfuric acid plant design today, namely, recovery and conservation of energy, and minimizing environmental impact. For example, in the relatively simple plants of a few years age, the SO_2 need contact the catalyst but once and the absorption of the resulting SO_3 in water (a solution of sulfuric acid) could be handled in a single absorption tower. Recycling could be kept to a minimum. In the modern sulfuric acid plant, double contact (DC) of the gases with catalyst and double absorption (DA) of the gases is commonly practiced. Designs are available in numerous configurations, each offering various advantages in terms of energy conservation, pollution minimization, initial and operating costs. A typical sulfur-burning DC/DA sulfuric acid plant is shown in the accompanying diagram.

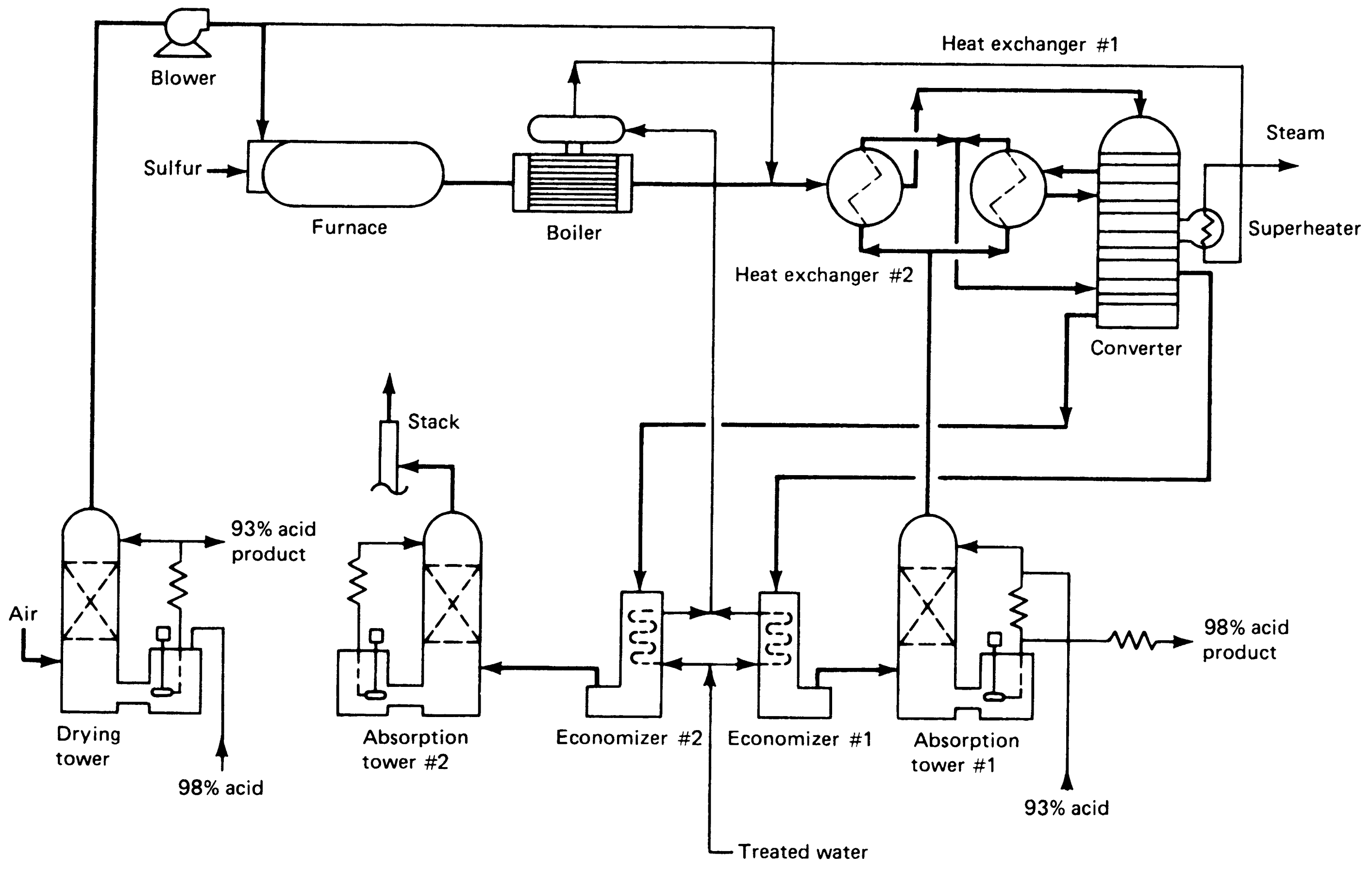
Of the approximately 40 million tons (36 million metric tons) of sulfuric acid manufactured in the United States per year, about 90% is used in the production of fertilizers and other inorganic chemicals. Much of the remaining 10% of H_2SO_4 is used by the petroleum, petrochemical, and organic chemicals industries. Much of this latter acid is involved in recycling kinds of processes. As pollution regulations in various countries become more restrictive, spent acid may become a much more attractive raw material than has been the case in the past.

As pointed out by Sander and Daradimos (1978), a regeneration of sulfuric acid of high quality can only be attained by thermal decomposition back to sulfur dioxide at high temperatures, where all organic impurities are completely burned—followed by reprocessing the SO_2 gases by the contact process to concentrated acid or oleum.

Reactivity of Sulfuric Acid. Dilute sulfuric acid reacts: (1) with many hydroxides, e.g., sodium hydroxide, to yield two series of sulfates (the acid is dibasic), e.g., sodium sulfate or sodium hydrogen sulfate, depending upon the ratio of acid to base reacting, (2) with many ordinary oxides, e.g., magnesium oxide, to yield the corresponding sulfate, e.g., magnesium sulfate solution, (3) with some carbonates, e.g., zinc carbonate, to yield the corresponding sulfate, e.g., zinc sulfate solution plus carbon dioxide gas (calcium carbonate is soon coated by a layer of calcium sulfate, which prevents further reaction), (4) with some sulfides, e.g., ferrous sulfide, to yield the corresponding sulfate, e.g., ferrous sulfate plus hydrogen sulfide gas, (5) with many metals, e.g., zinc, if not too pure (but not copper), to yield the corresponding sulfate, e.g., zinc sulfate solution plus hydrogen gas, (6) with solutions of some salts to yield the corresponding sulfate, e.g., barium chloride, changed to barium sulfate precipitate, calcium citrate, malate, tartrate to calcium sulfate precipitate and the free organic acid in solution.

Higher strengths of sulfuric acid react similarly in kind to the cases of (1), (2), (3), (6) above, but not, in general, as in cases (4) and (5) above. Copper and concentrated sulfuric acid yield copper sulfate and sulfur dioxide gas. Iron reacts similarly, yielding ferric sulfate in the place of copper sulfate.

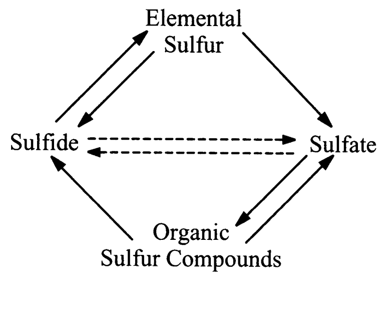
A number of other reactions of sulfuric acid are characteristic of its higher strengths. Concentrated sulfuric acid is thus (7) an oxidizing agent, and a further example is the oxidation of sulfur to sulfur dioxide (the reacting sulfuric acid is reduced to sulfur dioxide), (8) a sulfonating agent, e.g., naphthalene sulfonated to naphthalene-sulfonic acids (mono-, alpha or beta, di- several), (9) an esterification agent, e.g., methyl alcohol esterified to dimethyl sulfate $(CH_3O)_2SO_2$, melting point −32°C, boiling point 189°C, or methyl hydrogen sulfate $CH_3O \cdot SO_2OH$, ethyl alcohol esterified to diethyl sulfate $(C_2H_5O)_2SO_2$, melting point −26°C, boiling point 208°C, or ethyl hydrogen sulfate $C_2H_5O \cdot SO_2OH$, (10) a dehydration agent, e.g., formic acid into carbon monoxide, sugar blackened with separation of carbon, (11) an addition agent, e.g., ethylene into ethyl hydrogen sulfate, (12) a non-volatile acid upon heating, e.g., with sodium chlorite or nitrate, hydrogen chloride or nitric acid, respectively, is volatilized and sodium sulfate or sodium hydrogen sulfate remains as a residue.



Representative sulfuric acid plant of the sulfur-burning, double contact (DC), double absorption (DA) type.

SULFUR (In Biological Systems). Sulfur, in some form, is required by all living organisms. It is utilized in various oxidation states, including sulfide, elemental sulfur, sulfite, sulfate, and thiosulfate by lower forms and in organic combinations by all. The more important sulfur-containing organic compounds include the amino acids (cysteine, cystine, and methionine, which are components of proteins); the vitamins thiamine and biotin; the cofactors lipoic acid and coenzyme A; certain complex lipids of nerve tissues, the sulfatides; components of mucopolysaccharides, the sulfated polysaccharides; various low-molecular-weight compounds, such as glutathione and the hormones vasopressin and oxytocin; and many therapeutic agents, such as the sulfonamides and penicillins, as well as oral hypoglycemic agents sometimes used in treatment of diabetes mellitus. Sulfhydryl groups of the cysteine residues in enzyme proteins and related compounds, such as hemoglobin, play a key role in many biocatalytic processes; sulfhydryl-disulfide interchange reactions involving the cysteine residues of proteins are critical events in the immune processes, in transport across cell membranes, and in blood clotting. The S—S bridges between these residues are important in the maintenance of the tertiary structure of most proteins.

The electronic structure of sulfur is such that a variety of oxidation states are readily obtainable. It can be said that a sulfur cycle exists in nature, as noted by



The oxidation and reduction of elemental sulfur and sulfide occur in different species of bacteria, e.g., the oxidation of sulfides via elemental sulfur to sulfate takes place in *Chromatia*, the alternative oxidation to sulfate in *Thiobacillus*. The reduction of sulfate to sulfide occurs in *Desulfovibrio*. The biosynthesis of organic sulfur compounds from sulfate takes place mainly in plants and bacteria, and the oxidation of these compounds to sulfate is characteristic of animal species and of heterotrophic bacteria.

The amino acids cysteine and cystine are interconverted by oxidation-reduction reactions, as shown by

$$\begin{array}{llcl}
S\text{———} & S & & SH\\
| & | & & |\\
CH_2 & CH_2 & & CH_2\\
| & | & & |\\
CNNH_2 & CNNH_2 + 2H & \rightleftharpoons 2 & CNNH_2\\
| & | & & |\\
COOH & COOH & & COOH\\
\multicolumn{2}{c}{\text{(Cysteine)}} & & \text{(Cystine)}
\end{array}$$

Cystine was first isolated from a urinary calculus by Wollaston in 1805. It was shown to be a component of protein by Morner in 1899 and independently by Embden in 1900. Proof of its structure was given by Friedman in 1902. See also **Amino Acids; Coenzymes; Proteins;** and **Vitamins.**

In the chain from soils to plants to humans, inorganic sulfur, or more accurately, the sulfate ion (SO_4^{2-}), is taken up by plants and converted within the plant to organic compounds (the sulfur amino acids). These amino acids combine with other amino acids to make up plant protein. When the plant is eaten by a human or by livestock animals, the protein is broken down and the amino acids are absorbed from the digestive tract and recombined in the proteins of the animal body. The most important feature of sulfur in the food chain is that plants use inorganic sulfur compounds to make sulfur amino acids, whereas animals and humans use the sulfur amino acids for their own processes and excrete inorganic sulfur compounds resulting from the metabolism of the sulfur amino acids.

Ruminants, such as cattle, sheep, and goats, can use inorganic sulfur in their diets because the microorganisms in the rumen convert the inorganic sulfur into sulfur amino acids and these are then absorbed farther along in the digestive tract.

Soils very low in available sulfur are common in a number of regions of the world. In the United States, low-sulfur soils are frequently found in the Pacific Northwest and in some parts of the Great Lakes states. For many years, sulfur in the form of calcium sulfate was an accessory part of most commercial phosphate fertilizers, and this probably helped to prevent development of widespread sulfur deficiency in crops grown where these fertilizers were used. Volatile sulfur compounds from smoke, particularly before tight pollution controls, were an important source of sulfur for plants growing near industrial centers. In some cases, excessive sulfur in the air can cause injury to the plants. The trend toward high-analysis fertilizers without sulfur and air pollution abatement diminishes some of the inadvertent sources of sulfur for plants and crops and creates a need for more deliberate use of sulfur-containing fertilizers.

The extent to which any plant will convert inorganic sulfur taken up from the soil into amino acids and incorporate these into protein is controlled by the genetics of the plant. Increasing the available sulfur in soils to levels in excess of those needed for optimum plant growth will not increase the concentration of sulfur amino acids in plant tissues. To meet the requirements for sulfur amino acids in human diets, the use of food plant species with the inherited ability to build proteins with high levels of sulfur amino acids is required in addition to that supplied by way of the soil.

Since animals tend to concentrate in their own proteins the sulfur amino acids contained in the plants they eat, such animal products (meat, eggs, and cheese) are valuable sources of the essential sulfur amino acids in human diets. In regions where the diet is composed almost entirely of foods of plant origin, deficiencies of sulfur amino acids may be critical in human nutrition. Frequently, persons in such areas (also voluntary vegetarians) are also likely to suffer from a number of other dietary insufficiencies unless supplemental sources are used.

Diets of corn (maize) and soybean meal are usually fortified with sulfur amino acids for pigs and chickens. Sometimes fishmeal, a good source of sulfur amino acids, is added to the diets, or sulfur amino acids synthesized by organic chemical processes may be used.

Since ruminants can utilize a wide variety of sulfur compounds, any practice to increase the sulfur in plants may help to meet the requirements of these animals. Sheep appear to have a higher requirement for sulfur than most other animals, perhaps because wool contains a fairly high level of sulfur. Adding sulfur fertilizers to soils used to produce forage for sheep may improve growth and wool production, even though no increased yield of the forage crop per se may be noted.

Sulfate and Organic Sulfates. Inorganic sulfate ion (SO_4^{2-}) occurs widely in nature. Thus, it is not surprising that this ion can be used in a number of ways in biological systems. These uses can be divided primarily into two categories: (1) Formation of sulfate esters and the reduction of sulfate to a form that will serve as a precursor of the amino acids cysteine and methionine; and (2) certain specialized bacteria use sulfate to oxidize carbon compounds and thus reduce sulfate to sulfide, while other specialized bacterial species derive energy from the oxidation of inorganic sulfur compounds to sulfate.

Among the variety of sulfate esters formed by living cells are the sulfate esters of phenolic and steroid compounds excreted by animals, sulfate polysaccharides, and simple esters, such as choline sulfate. The key intermediate in the formation of all of these compounds has been shown to be 3′-phosphoadenosine-5′-phosphosulfate (PAPS). This nucleotide also serves as an intermediate in sulfate reduction.

In organisms that utilize sulfate as a source of sulfur for synthesis of cysteine and methionine, the first step in the reduction process is the formation of PAPS. This is not surprising since the direct reduction of sulfate ion itself is an extremely difficult chemical process. It is known that the reduction of esters and anhydrides occurs much more readily than the reduction of corresponding anions. Following activation, the sulfuryl group of PAPS is reduced to sulfite ion (SO_3^{2-}) by reduced triphosphopyridine nucleotide (TPNH) and a complex enzyme system. Following the reduction of PAPS to sulfite, additional reduction steps readily produce hydrogen sulfide, which appears to be a direct precursor of the amino acid cysteine.

Sulfur Compounds in Onion and Garlic. Dating back to antiquity, there have claims made for the curative and preventative physiological powers of onion and garlic. Dr. Eric Block (State University of New York at Albany), a specialist in the organic chemistry of sulfur, and colleagues have investigated the chemistry of onion and garlic over a period of years, some of the results of which were reported in the Block (1985) reference listed. As pointed out by Block, the cutting of an onion or a garlic bulb releases a number of low-molecular-weight organic molecules that incorporate sulfur atoms in bonding forms rarely encountered in nature. These molecules are highly reactive and they change spontaneously into other organic sulfur compounds, which in turn participate in further transformations. Researchers have cataloged a number of biological effects of the extracts from these bulbs, including antibacterial and antifungal properties. Other extracts act as antithrombotic agents (inhibit blood platelets). As early as 1721, a drink consisting of wine and macerated garlic (*vinaigre des quatre voleurs*) was used as an antibiotic in France and is still available today! Pasteur (1858) reported on the antibacterial properties of garlic. Albert Schweitzer is reported to have used garlic in the treatment of amoebic dysentery in Africa. As reported by Block, laboratory investigations have shown that garlic juice diluted in one part in 125,000 inhibits the growth of bacteria of the genera *Staphylococcus, Sterptococcus, Vibrio* (including *V. cholerae*) and *Bacillus* (including *B. typhosus, B. dysenteriae*, and *B. Enteritidis*). Lacrimatory factors contained in these bulbs are well known.

Serious research commenced in 1844 by Theodor Wertheim, a German chemist. He attributed some of the properties of garlic, "*mainly to the presence of a sulfur-containing, liquid body, the so-called garlic oil. All that is known about the material is limited to some meager facts about the pure product, which is obtained by steam distillation of bulbs of Allium sativum. Since sulfur bonding has been little investigated so far, a study of this material promises to supply useful results for science.*" Wertheim suggested the name *allyl* for the oil. Today, allyl is used for chemicals in the C_3H_5 series ($CH_2{=}CHCH_2$).

Another German investigator (Semmler, 1892) also produced garlic oil via steam distillation. The oil yielded diallyl disulfide, $CH_2{=}CHCH_2SSCH_2CH{\equiv}CH2$, with minor amounts of diallyl trisulfide and diallyl tetrasulfide present. The oil yielded by similar experimentation with onions was different, containing essentially propionaldehyde, C_2H_5CHO, plus a number of sulfur compounds, of which dipropyl disulfide, $C_6H_{12}S_2$, was one.

Using less harsh methods, Cavallito (1944) produced an oil, $C_6H_{10}S_2O$, by extracting the garlic with ethyl alcohol (room temperature). This oil was found to be more powerful than penicillin or sulfaguanidine against *B. typhosus*. The exact formula of Cavallito's oil was found to be allyl-2-propenethiosulfinate, $Ch_2{=}CHCH_2S(O)SCH_2CH{=}CH_2$. Cavallito gave this substance the common name, *allicin*. Precursor molecules for allicin have been identified and it has been established that allicin is not developed in garlic until it is initiated by an enzyme, termed *allinase*.

As pointed out by Block, allinin is the "first natural substance found to display optical isomerism due to mirror-image forms at sulfur as well as at carbon." The research by Block and others is well delineated in the Block (1985) reference. Although much remains to be learned, there is now a long line of hard scientific evidence that garlic and onion have beneficial physiological properties. Most scientists to date suggest that the properties of these bulbs are best exploited by consuming the fresh products, rather than from extracts, particularly when the latter are derived from harsh methodologies, such as steam distillation.

Preservatives. Sulfur compounds, such as sulfur dioxide and sodium bisulfite, are used commercially to preserve the color of various food products, such as orange juice, dehydrated fruits and vegetables, such as apricots, carrots, peaches, pears, potatoes, and many others. Concentrated sulfur dioxide is used in wine-making to destroy certain bacteria. The color preservation of canned green beans and peas is enhanced by dipping the produce in a sulfite solution prior to canning. In 1986, some of these compounds and uses were put under closer regulation in the United States.

The sulfatases are a widely distributed group of enzymes that hydrolyze simple sulfate esters to inorganic sulfate.

Sulfur-Based Pesticides. Sulfur (elemental) has been used as an effective acaricide, fungicide, and insecticide. For ease of use, a number of special formulations are available, ranging from sulfur dusts (up to 95% sulfur); a wettable powder (30 to 90%); and pastelike solutions in which the sulfur is ground to a fine colloidal form. Such formulations may contain up to 50% sulfur. Target plant diseases of sulfur when used as a fungicide include: apple scab, brown rot, downy and powdery mildew, and peach scab. Against insects, sulfur is effective for mite, scale, and thrip. Most formulations are not injurious to honeybees.

Specific sulfur control chemicals include: (1) Calcium polysulfide (lime-sulfur), dating back to the 1850s and available as a solution (up to 31% sulfur) or as a dry powder (up to 70% sulfur). The compound is effective against anthracnose, apple scab, brown rot, powdery mildew, and peach leaf curl—and against mite and scale insects. (2) Sodium polysulfide and sodium thisulfate mixtures—used for spraying and dipping fruit, adding color to the product, and prolonging the period during which the fruit can be picked. (3) Sodium thiosulfate pentahydrate, which prevents discoloration of some green vegetables (use is regulated).

Role of Sulfur in Tidal Wetlands. As pointed out by Luther, et al. (1985 reference listed), the biogeochemical role of sulfur in tidal wetlands presently is subject to considerable research. Sulfur is an important redox element under natural aquatic conditions and is responsible for several important biogeochemical processes, including (1) sulfate reduction, (2) pyrite formation, (3) metal cycling, (4) salt-marsh ecosystem energetics, and (5) atmospheric sulfur emissions. These processes depend upon the formation of one or more sulfur intermeidates, which may have any oxidation state between $+6$ and -2. The intermediate oxidation states may be organic or inorganic. In their study, Luther and colleagues analyzed sulfur species in pore waters of the Great Marsh, Delaware. Anticipated findings reported were bisulfide increases with depth due to sulfate reduction and subsurface sulfate excesses and pH minima, the result of a seasonal redox cycle. Not expected was the pervasive presence of thiols, such as glutathione, particularly during periods of biological production. It appears that salt marshes may be unique among marine systems in producing high concentrations of thiols. Polysulfides, thiosulfate, and tetrathionate also showed seasonal subsurface maxima. The findings suggest a dynamic seasonal cycling of sulfur in salt marshes involving abiological and biological reactions and dissolved and solid sulfur species. The researchers suggest that the chemosynthetic turnover of pyrite to organic sulfur is the likely pathway for this sulfur cycling. It follows that the material, chemical, and energy cycles in wetlands appear to be optimally synergistic.

SULFUROUS ACID. H_2SO_3, formula weight 82.08, colorless liquid, prepared by dissolving SO_2 in H_2O. Reagent grade H_2SO_3 contains approximately 6% SO_2 in solution. As a bleaching agent, sulfurous acid is used for whitening wool, silk, feathers, sponge, straw, wood, and other natural products. In some areas, its use is permitted for bleaching and preserving dried fruits. The salts of sulfurous acid are sulfites.

Sulfurous acid is a strong reducing agent, being oxidized to H_2SO_4 (1) on standing in contact with air, (2) by chlorine, bromine, iodine, yielding HCl, HBr, or HI, respectively, (3) by HNO_3 or nitrous acid yielding nitric oxide, and (4) by permanganate. Sulfurous acid is itself reduced by zinc and dilute H_2SO_4 to H_2S. Sulfurous acid also may be formed by the reaction of a sulfite or bisulfite solution and an acid.

Sodium sulfite Na_3SO_3 and sodium hydrogen sulfite $NaHSO_3$ are formed by the reaction of sulfurous acid and NaOH or sodium carbonate in the proper proportions and concentrations. Sodium sulfite, when dry and upon heating, yields sodium sulfate and sodium sulfide. Sodium pyrosulfite (sodium metabisulfite) $Na_2S_2O_5$ is a common sulfite. Crystalline sulfites are obtained by warming the corresponding bisulfite solutions. Calcium hydrogen sulfite Ca $(HSO_3)_2$ is used in conjunction with excess sulfurous acid in converting wood to paper pulp. Sodium sulfite and silver nitrate solutions react to yield silver sulfite, a white precipitate, which upon boiling decomposes forming silver sulfide, a brown precipitate.

An esterification agent, sulfurous acid forms dimethyl sulfite $(CH_3O)_2SO$, bp 126°C and diethyl sulfite $(C_2H_5O)_2SO$, bp 161°C. Sulfites give a white precipitate with barium chloride, soluble in HCl with evolution of SO_2. Sulfites decolorize iodine in acid solution.

SUM. The answer when two or more quantities are combined by addition. In group theory, the term direct sum has a special meaning (see **Representation of Groups**).

The usual symbol for the operation of summartion is the Greek capital letter *sigma* (Σ). Thus,

$$\sum_{n=0}^{\infty} a_n x^n = f(x)$$

means to add the terms $a_n x^n$, assigning every integral value from zero to infinity to the letter n. When two or more sums are to be taken with respect to several subscripts or indices only one summation sign may be used and the summation limits are often omitted, if they are understood from the accompanying text. Even the summation sign can be omitted and this is common, especially in tensor analysis. Thus, an expression like

$$A^m = \frac{\partial x^{-m}}{\partial x^i} A^i$$

by convention means that summation is to be made over the repeated index, i. Still further condensation is customary in matrix algebra, where $\mathbf{Ab} = \mathbf{C}$ means that the matrix $\mathbf{C}$ has elements C_{ij} calculated by the equation

$$C_{ij} = \sum A_{ik} B_{kj}$$

with A_{ik} and B_{kj} as elements of **A** and **B.** The summation over k is to be extended from unity to n, where n is the number of columns in **A** and the number of rows in **B.**

SUMNER LINE. A line of position obtained from the observation of altitude of some celestial object. This method for obtaining a line of position was discovered by Captain Thomas H. Sumner in 1837, and circumstances leading to the discovery are described in "The American Practical Navigator," Bowditch.

The method employed by Sumner for obtaining a celestial line of position was standard procedure for American ship masters until the early part of the twentieth century. For this reason, it is worthy of consideration here, in spite of the fact that it has been almost completely superseded by the methods described in the article on celestial navigation. Captain Sumner used the old-fashioned method for determining latitude and longitude at sea from an observation of the altitude of a celestial object. These methods require the solution of the astronomical triangle. To solve this triangle, at least three parts must be known. At sea, the altitude of the object, as obtained from sextant observations, will give one part, and the declination of the object will give another. To obtain the third part, either the latitude or the longitude of the observer must be known. Since latitude and longitude are the coordinates that the navigator is seeking, it would seem at first glance that a solution of the problem is impossible. However, if the object observed is approximately due east or west, a change in latitude of several miles will produce but slight effect on the computed longitude. On the other hand, if the object is nearly due north or south, a change in longitude will produce but slight effect in the computed latitude. An approximate value of latitude and longitude can be obtained by methods of dead reckoning (DR). If the observed object is within 45° of east or west, the DR latitude is used and the longitude is computed. When the object is within 45° of the meridian, the DR longitude is used in the computation of the latitude. If the DR position is known to within 20 miles (32 kilometers), the computed latitude or longitude will be accurate to within $\frac{1}{4}$ of a mile, unless the observed object is close to the zenith.

In Captain Sumner's case, his ship had experienced gales and fog for a number of days, and the DR position was very uncertain. When the clouds broke away, he obtained an altitude of the sun in the forenoon. Since his DR position was uncertain, he assumed several values for latitude separated by about 20 miles (32 kilometers), and computed the corresponding longitudes. On plotting these positions, he found that they lay along a straight line on a mercator chart. He made the assumption, which has since been established as sound, that his ship must be on the line.

After the publication of this discovery, the use of Sumner lines became the standard procedure for most navigators. If two objects, differing in bearing by at least 45°, are available for observation, two Sumner lines can be obtained and a fix determined. If, as is frequently the case during daylight hours, only one object is available, this one object is observed twice, with an interval of time between the two observations sufficient to produce a change in bearing of at least 45°. The fix of these two lines is obtained by moving one line to the time of the other by the method of running fix.

A Sumner line is, in reality, a small circle on the earth, with a point on the earth directly under the observed object as center. Unless the object is within 10° of the zenith, the curvature is so slight as to be negligible in drawing the line on a mercator chart.

See also **Course;** and **Navigation.**

SUMPTNER PRINCIPLE. When a source of light is placed at any point inside a sphere with perfectly diffusing walls, every part of the interior is equally illuminated.

SUN (The). The sun is a self-luminous body; it is our nearest star, 93,000,000 miles (~150 million kilometers) from the Earth, from which we receive light and heat to generate and maintain all life processes on the Earth. The sun is 109 times the earth's diameter and 300,000 times its mass, but when we compare it to other stars for size, mass, and brightness we find that it falls in the middle of their range. It is a typical dwarf star of spectral class G_2. Its energy output in the form of light and heat, generated by nuclear processes, appears to be remarkably constant and to have been so over hundreds of millions of years, and is expected to be so for many billions of years into the future. Being the closest star, we can examine its surface in great detail. As such it serves as a testing ground for astrophysical theories. The sun's atmosphere consists of the visible surface or photosphere and two higher-temperature envelopes, the chromosphere and corona. Telescopic views show the surface of the sun to be in great turmoil, covered by a granular pattern of convection cells and punctuated with sunspots, great magnetic storms in the sun's atmosphere. Magnetism plays the dominant role in the sun's variability; in addition to sunspots it is fundamental to flares, with their x-ray and radio bursts, it controls the motions and behavior of prominences, governs the solar wind streaming past the earth, and indirectly produces a wide variety of solar-terrestrial events: auroras, magnetic storms, radio fadeouts and possible changes in the climate. Physical data for the sun are given in the accompanying table.

Modern solar astronomy can be roughly divided into three major subdivisions:

1. *Velocity fields*: Solar rotation, granulation and supergranulation, solar oscillations.
2. *Magnetic fields*: General field of the sun, sunspots, flares, prominences.
3. *Composition*: Interior structure, chemical composition, energy output.

These will be discussed in the following sections of this entry.

The Photosphere

The visible surface of the sun is called the photosphere—literally, light sphere. If the surface of the sun is examined carefully through a telescope of 4 inches (~10.2 centimeters) aperture or larger, at the times of best viewing the surface is seen to be mottled, a structure referred to as *granulation*. Figure 1 shows granules in a photograph taken on very high contrast film. Also clearly visible, even in a small telescope, and sometimes with the naked eye, are *sunspots*, dark only in contrast to their brilliant surroundings.

Three to four million granules, each about the size of an American state (200–1000 miles; 322–1609 kilometers across), cover the surface of the sun. Similar to the convection in a boiling pot of cereal, granulation is one of the ways energy is brought to the surface from the deeper, hotter layers. Time lapse photography reveals the 5–10 minute life history of a granule: each starts out as a small bright area that grows in several minutes to about 1000 miles (1610 kilometers) diameter, divides into several smaller units, which may coalesce with other granules or they fade and then are replaced by a new granule. A spectrum of the photosphere shows Doppler shifts of the spectrum lines indicating that each granule rises in the atmosphere with an average velocity of $\frac{1}{3}$ mile (~0.5 kilometer) per second. As the granule cools and dissipates the gas returns to the sun in the darker intergranular spaces. Groups of granules appear to be in a constant up and down oscillation with a 5-minute period.

The photospheric layer is about 500 miles (805 kilometers) thick—hotter (10,000°K), denser, and more opaque at the bottom, cooler (4200°K), much less dense, and quite transparent at the top. The energy radiated from the photosphere can be characterized by a continuous spectrum with a temperature of 5740°K—the effective temperature of the sun. In appearance the surface is brilliant white, brighter at the center of the disk where we see to the deeper hotter layers and darker at the limb of the sun where we see only the outermost cooler layers of emitting gas. The variation in brightness, center to limb, is known as the solar *limb darkening* (visible also in spectrum lines, i.e., Hα of Fig. 2).

DIMENSIONS AND PHYSICAL DATA FOR THE SUN

Diameter	864,000 miles (1,390,180 kilometers); 110 × diameter of the earth
Volume	1.412×10^{33} cm^3 (1.3 million times the volume of the Earth)
Mass	1.989×10^{33} gm (330,000 times mass of the Earth)
Mean density	1.41 gm cm^{-3} ($\frac{1}{4}$ the earth's mean density)
Central density	148 gm cm^{-3}
Force of gravity of surface	Nearly 28 times that at earth's surface (A mass weighing 100 pounds on Earth would weigh nearly 1.4 tons on the sun)
Mean distance from Earth	92,955,807 miles (149,597,870 kilometers = 1.0 astronomical unit)
Time for light to travel from sun to Earth	499 seconds (slightly over 8 minutes)
Velocity in space	About 12.4 miles per second (20 kilometers per second) toward a direction in space not far from the star Vega
Solar constant (rate at which solar radiation is received outside the Earth's atmosphere on a surface normal to the incident radiation and at the Earth's mean distance from the sun)	1.94 calories per square centimeter per minute
Candle power of sun	2.4×10^{27} candles
Average illumination of sun at zenith	100,000 meter-candles

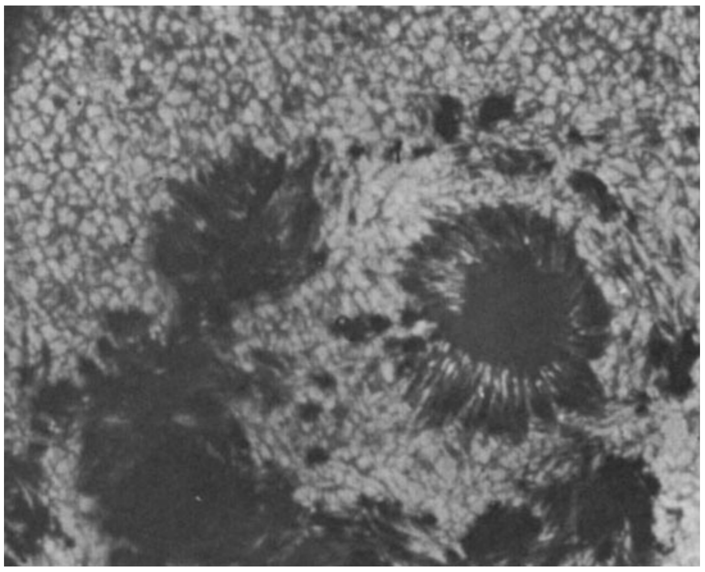


Fig. 1. Portion of sun's disk photographed from a balloon at an altitude of about 24 kilometers (80,000 feet). (*Princeton University.*)

The amount of limb darkening varies with wavelength. In the red and infrared there is very little variation between center and limb, whereas in the violet the intensity falls rapidly away from disk center to about 10% of the central value. From the observed limb darkening we can derive from an integral equation the temperature variation with depth in the sun's atmosphere.

The Chromosphere

A layer of the sun's atmosphere about 6200 miles (10,000 kilometers) thick just above the photosphere was named the chromosphere by eclipse observers. At eclipse time, when the moon has covered the bright photosphere, it is revealed for a few seconds before and after totality, as a thin crescent colored bright pink (from the hydrogen Hα spectrum line). Examined through a telescope the top of the layer is resolved into a great multitude of small, short-lived (5–15 minutes), geyserlike jets called *spicules*, each about 310 miles (500 kilometers) diameter and 3100–6200 miles (5000–10,000 kilometers) tall, projecting outward at all angles to the sun's surface.

Spicules are not randomly distributed over the sun's surface. High-resolution spectroheliograms show that they are arranged in a "network" pattern that covers the entire sun. Piled up at the network boundaries by slow, outward directed motion, each element of the net is a giant convection cell, typically 18,600 miles (30,000 kilometers) across, in which material flows up and outward from the center carrying with it the chromospheric magnetic fields concentrated in the spicules, to the cell boundary, and then descends. The motion is slow, horizontally about $\frac{1}{3}$ mile ($\frac{1}{2}$ kilometer) per second. The lifetime of a cell is about a day.

The disk of the sun when viewed in the light of Hα through a spectroheliscope or monochromatic filter presents a totally different view from its white light photospheric image. Prominences, bright at the limb show very black on the disk (Fig. 2). Bright areas appearing around sunspots, given the French name for beach, are called *plages*. In the central regions of the disk we see a composite of very small, overlapping, wormlike prominence structures referred to as *chromospheric mottling*. Near sunspots these structures are drawn out into long threads and filaments which often exhibit a spiral structure tracing out magnetic fields emerging from the spot.

The gases of the chromosphere are transparent. When viewed at the extreme limb, and not projected onto the bright underlying photosphere, the chromospheric spectrum consists of emission lines which in the last crescent phase of an eclipse momentarily flash into view, i.e., the *flash spectrum*. See Fig. 3. From the temperature minimum (4200°K) at the top of the photosphere the temperature rises to a plateau of 6000°K for the layers 620–1240 miles (1000–2000 kilometers) high then rises very rapidly to 30,000°K at 1550 miles (2500 kilometers) and then to coronal values. The lower pressure and higher temperature of the chromosphere results in greater ionization and excitation than in the photosphere, thus the spectrum consists of many lines from singly ionized elements.

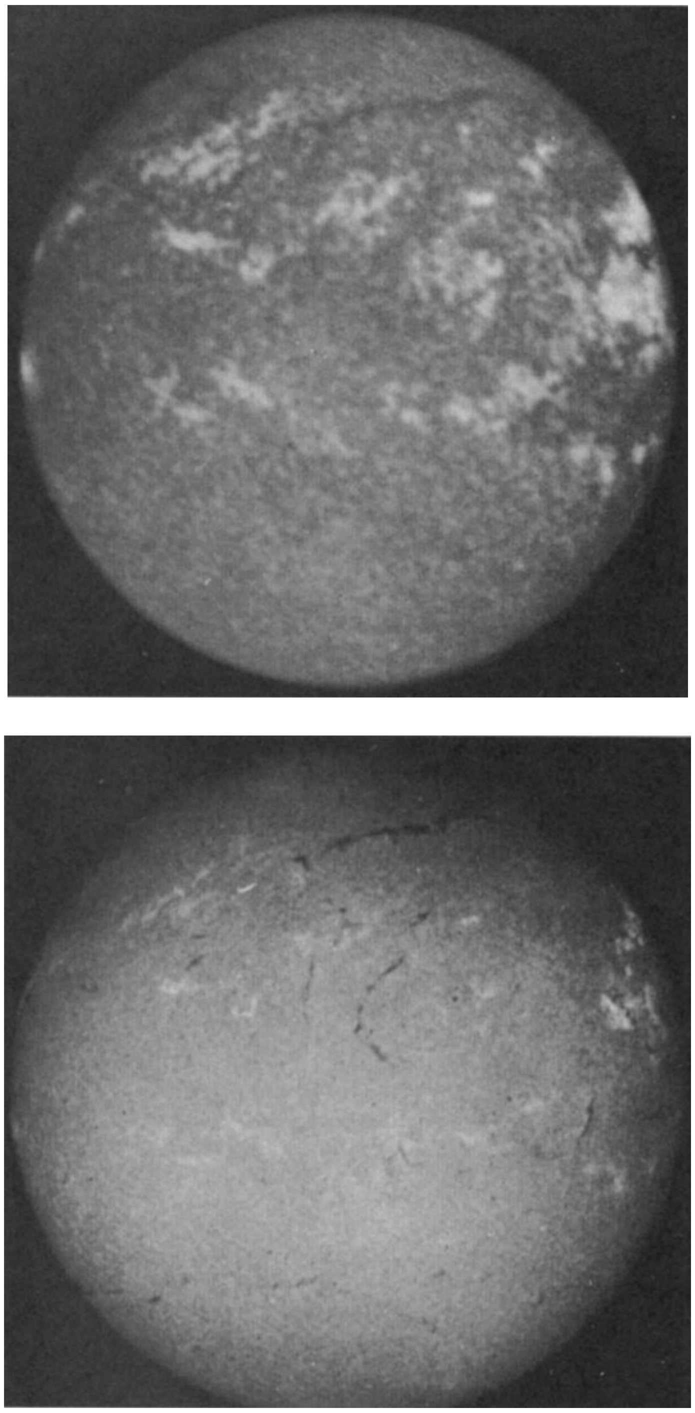


Fig. 2. Photographs of sun taken in the K line of Ca II. (*Top*) Calcium plages are indicated. (*Bottom*) Simultaneous photograph in the Balmer H alpha line. (*U.S. Navy.*)

Corona

At the time of a total solar eclipse a pearly white radiance, called the corona, can be seen around the sun. See Fig. 4. The corona is composed of three parts, labeled L, K and F. The L-corona refers to a low level portion, the light from which is emitted by highly ionized atoms. The light of the K-corona, one hundred times brighter than L, is sunlight scattered by fast moving free electrons. The F component is produced by sunlight scattered by interplanetary dust particles; it is an extension of zodical light inward to the sun. Only the L and K components represent the true coronal atmosphere of the sun. Though the inner corona

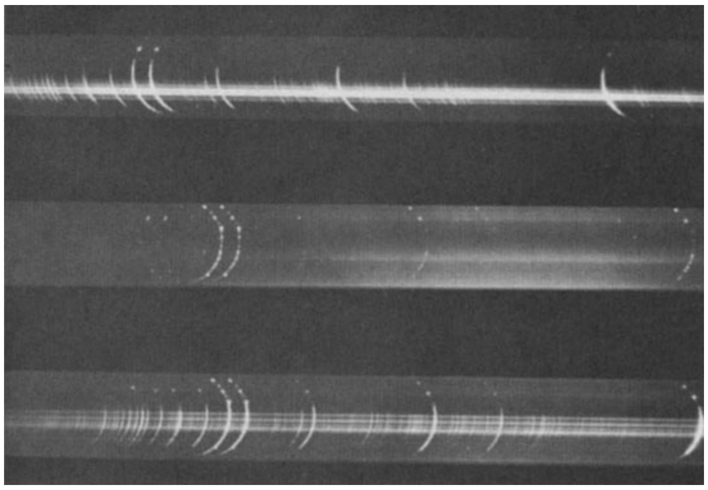


Fig. 3. Solar flash spectrum. (*Top*) Just before second contact; (*Center*) Bailey's—beads light passing through valleys in the moon's limb form two rings that are the violet lines of ionized calcium; (*Bottom*) Just after third contact. (*Mount Wilson and Palomar Observatories.*)

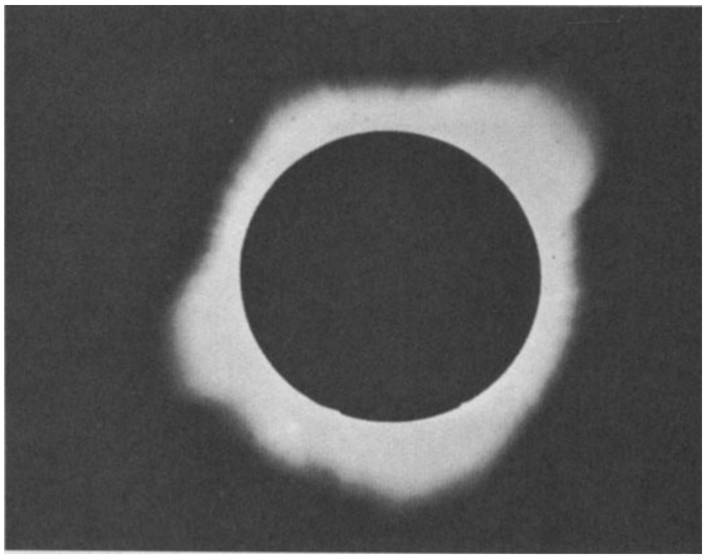


Fig. 4. The corona shown at time of total eclipse of the sun.

can be studied with a very specialized telescope called a Coronagraph or from space, its beauty can only be seen and appreciated at the time of a total eclipse, when one can discern outward streaming rays, arches, filaments, and brilliant red prominences, apparently projected on the corona. At sunspot maximum short streamers extending outward in all directions give the corona a globular shape; at sunspot minimum there may be only a few long equatorial streamers together with a polar plume or crown composed of fine rays curving outward from the north and south poles like iron filings at the poles of a magnet.

Though normally visible only at the limb of the sun, when the bright surface is occulted, the corona can also be seen projected on the disk, if photographed from space in x-ray wavelengths. The Japanese YOHKOH satellite, launched at the peak of the 1991 sunspot cycle, aimed in x-rays at the million-degree corona and yielded spectacular motion pictures of intense, active regions in the sun's corona, visible in these wavelengths both on the sun's disk and off the sun's limb. The hope is to solve the puzzle of the flare mechanism. Photographs from Skylab revealed a complex structure of bright points surrounded by rays and arch structures. Between them large dark irregular areas appear, *coronal holes*. These are low-density regions from which magnetic lines of force lead out into interplanetary space allowing the plasma to escape which, moving with a velocity of about 400 km/sec near the earth, is called the *solar wind*. This represents a steady mass flow from the sun of 10^{-14} solar masses/year. Coronal transients, outwardly expanding loops or clouds, contribute another 5%. Parenthetically this is a very small fraction of the loss shown by many early-type hot stars and cool giants and supergiants.

The spectrum of the corona, first observed in 1869, exhibited many unidentified spectrum lines. Not duplicated in the laboratory, the composition of the corona remained a mystery for many years, leading the suggestion of a light hypothetical gas, "coronium," as their source. W. Grotrian and B. Edlen (1939–1942) found the clue and showed that the spectrum was that of common elements—iron, calcium, nickel, etc.—which had lost 10 to 15 electrons, thus indicating a temperature of several million degrees for the corona. As yet no satisfactory theory accounts for the temperature rise from about 4200°K at the top of the photosphere to 2,000,000°K in the corona.

Spectrum Analysis

The latter half of the nineteenth century saw not only great advances in the observations of the solar spectrum, but also the beginnings of the theory by which those observations could be understood. The first breakthrough occurred when Gustav Kirchoff (1859) noted that incandescent solids or liquids gave off continuous spectra of all colors, while hot gases emitted light in bright lines, their color or wavelength being characteristic of the chemical composition of the gas. If the continuous spectrum of an incandescent solid was passed through a cooler gas, dark lines would appear in the spectrum at exactly the same positions as the bright lines displayed by the incandescent gas when viewed alone. This fact formed the basis for the analysis of the solar spectrum. Kirchoff applied his results to derive a model of the sun as a hot liquid sphere covered by an atmosphere of gases which produced the dark lines in the spectrum. We now know that a dense gas may also produce a continuous spectrum and that the sun is totally gaseous, but for the 1860s Kirchoff's ideas were quite reasonable.

Since each gaseous element produces a characteristic pattern of spectral lines it was possible to analyze the chemical composition of the sun, a task first undertaken by Kirchoff together with a Heidelberg Professor of Chemistry, R. Bunsen. The solar spectrum was set beside the spectrum produced by a laboratory sample (usually heated to an incandesent vapor in a spark discharge or an arc source) and visually noting the coincidences of the lines; in the later years of the 19th century a photographic comparison was made with far greater precision. Kirchoff and Bunsen identified solar spectrum lines from the following elements: sodium, calcium, barium, strontium, magnesium, copper, iron, chromium, nickel, cobalt, zinc, and gold. Later workers such as Angstrom extended this list to include hydrogen, which we now know to be the major constituent of the sun. To date most (about 65) of the naturally occurring nonradioactive elements have been found to be present in the solar atmosphere. At the time of the total solar eclipse of 1869, Lockyer and Frankland observed in the flash spectrum a bright yellow line which had never before been seen in laboratory spectra. The element producing this line was not known on the earth and was named helium (after *helios*—the Greek word for sun). It was not until 1895 that helium was first detected on the earth. Despite its rarity on earth helium is the second most abundant gas in the sun (after hydrogen).

Cosmology deals with the chemical formation and evolution of the universe. The chemical composition of the sun, earth, and meteorites serve as standards of comparison; hence a great deal of effort has been expended in their determination. Solar abundances can only be determined for the sun's atmosphere, but with few exceptions the sun is considered to have the same composition throughout.

Solar (and stellar) atomic and molecular abundances can be determined from spectrum analysis, i.e., by knowing the atomic parameters from laboratory determinations and by comparing the observed intensity of Fraunhofer lines with theory. As the abundance increases a Fraunhofer line becomes darker in the center until black, i.e., saturated, then slowly broadens, finally developing resonance wings—the so-called *curve of growth*. On the theoretical side the strength of a line is determined by: (1) the abundance of the element, (2) the number of atoms of the element in a state able to produce the line—determined by the pressure and temperature of the gas—and (3) the atomic transition probability between the atomic levels connected with the line. A plot of the theoretical intensity vs. observed intensity yields the atomic abundance of the element. Hydrogen and helium make up 99.9% by volume of the sun. For every 1,000,000 atoms of hydrogen there are 63,000 of He, 690 of 0, 32 of Fe, 3 of Al, 2 of Na, etc.

Sunspots

The discovery and first scientific study of sunspots was carried out telescopically by Galileo Galilei in 1610. Many of his contemporaries to whom he showed them refused to believe their eyes for they considered the sun to be immaculate and hence they had to be objects in front of the telescope. However, Galileo noted their daily westward movement, their motion in parallel lines, and their slower apparent motion near the limb, from which he concluded that the sun was a sphere rotating in about 27 days and that the spots were on its surface.

When examined through a telescope a large sunspot is seen to consist of two distinct parts: a dark central *umbra* surrounded by a less dark *penumbra*. On birth, one sees several small round pores each about 2000 miles (3220 kilometers) in diameter separated by tens of thousands of miles. If their development continues—most small sunspots last less than a day—the leader and follower grow rapidly in size and separate, resulting in two large spots having umbra and penumbra together with a number of smaller spots forming a complex group. Old age sets in after a week or two; one of the large spots disintegrates, as it becomes covered by light bridges, leaving a single spot which over several days diminishes in size until gone. On occasion exceptionally large, complex groups develop, exhibiting great variety of shape, often covered by brilliant "bridges" of light, giving the impression of great turmoil. Such spots, which may extend over 100,000 miles (160,000 kilometers) of the sun's surface, are easily visible to the naked eye when the sun's light is reduced by fog or by absorption of the earth's atmosphere at sunset. Russian records of naked eye sunspots describe them as looking like nail heads. Chinese annals, some before Christ, record many objects (sunspots) on the sun and describe them as like a flight of birds or as having the appearance of pigeon eggs.

Sunspots are dark only in contrast to the photosphere. Brighter than an electric arc, examined spectroscopically one finds atomic and molecular lines from which a temperature of about 3800°K is obtained. George Ellery Hale (1914) found that some of the atomic lines were divided and polarized in the sunspot spectrum (the Zeeman effect), demonstrating the strong magnetic field associated with sunspots. Further work showed that pairs of spots were of opposite polarity, that the leader spot in the Northern Hemisphere had a magnetic polarity opposite that of the leader in the Southern Hemisphere, and that the polarity in each hemisphere reversed every 11 years.

Telescopic records for 400 years and scattered naked eye observations for 2000 years show that the number of sunspots varies with an average period of 11.2 years. Wolf, of Zurich, introduced the term *sunspot number*, $R = f + 10g$, as a measure of solar activity, where f is the total number of individual sunspots and g is the number of spot regions, either groups or spots. This number is quite arbitrary, but in practice serves very well to define solar activity. See Fig. 5. The last maximum occurred in 1979, at which time as many as 100 spots were visible on the sun's surface. At minimum, months may pass without a single record of a sunspot. The cause of this periodicity is not known. It has been suggested that Jupiter, which revolves around the sun with a period of 11.8 years and has a mass $\frac{1}{1000}$ that of the sun, could introduce tidal action and sunspots. Other suggestions, such as swarms of meteors impacting the surface, have been made—all without success.

We should here remark that the sun rotates not like the earth, as a rigid body, but rather nonuniformly. Being gaseous throughout, it rotates once in 25 days in the equatorial regions, whereas the polar regions require 35 days. This nonuniform rotation results in deep-seated stirring of the interior. The sun is a magnet like the earth. According to a recent theory due to Babcock, its nonuniform rotation causes the magnetic lines of force to wind up like string on a top. As the magnetic lines deep in the interior become drawn out parallel to the equator and get tighter and closer together, magnetic buoyancy and turbulent motion carries a kink in the lines to the surface. The visible manifestation at the surface then is a pair of spots of opposite magnetic polarity with a field strength of 1000 to 3000 gauss, a field equal to that in a modern power-plant generator but distinguished from it by the fact that it extends over an area larger than the size of the earth. The time required for the windup of the magnetic lines of force is 3–4 years. The time required for them to appear at the surface is 6–7 years; hence, the 11-year cyclic period. Each solar cycle begins with appearance of spots in two zones north and south about 30° from the equator. During the cycle the zone of spots drifts equatorward, ending at ±5°.

Strangely, the cycle of sunspot activity died out for about 75 years in the late 17th century. Indirect evidence of solar activity before the invention of the telescope in 1610, from isotopic carbon 14 found in tree rings, suggests that even longer periods of quiescence have occurred. Eddy has noted that the periods of low solar activity coincides with climatic lows in European temperatures and thus a connection between solar activity and climate.

Prominences

Prominences first seen at solar total eclipses stand as cloudlike forms above the sun's surface. See Fig. 6. They are classified by appearance and formation into the following types: quiescent, coronal, eruptive, sunspot, and tornado. A typical quiescent prominence is a large, thin, caterpillerlike form with several of its feet attached to the sun's surface; some may extend over 125,000 miles (201,125 kilometers) in length; 30,000 miles (48,270 kilometers) in height by 5,000 miles (8,045 kilometers) thick. They generally develop in the sunspot zone, then drift north or south to the polar regions; drawn out by differential solar rotation they may last for several years as a polar crown. Coronal forms appear to condense from the hot coronal gas and stand free of the sun's surface. Active prominences are generally associated with underlying or nearby sunspots.

Though formerly viewed at an eclipse or spectroscopically, today prominences are more easily seen and studied through a narrow pass-

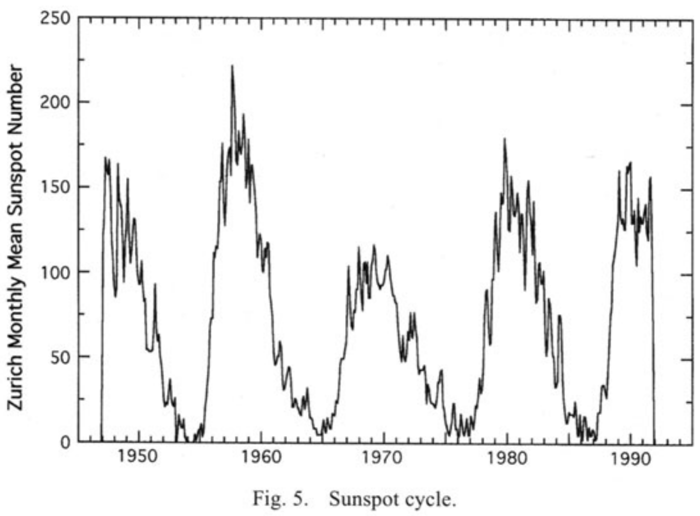


Fig. 5. Sunspot cycle.

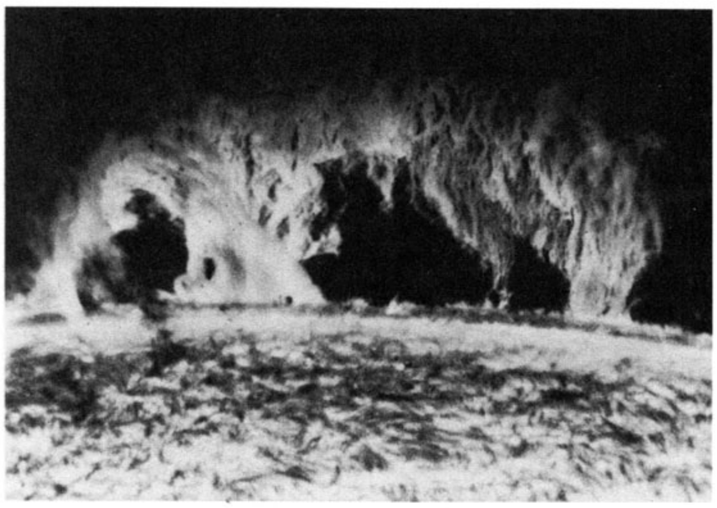


Fig. 6. Hedge-row prominence. (*Big Bear Solar Observatory.*)

band interference filter (5 angstroms or less) or through a Lyot-Ohman polarizing filter (0.2–1.0 Angstrom passband) centered on Hα—the red line of hydrogen, $\lambda = 6563$ angstroms. Time-lapse motion pictures of active prominences show the strong control exerted by magnetic fields on their support and behavior. Often appearing as great inverted funnels, condensing from the corona, there is a general downward flow of prominence material along curved arcs into neighboring sunspots. Over a sunspot region, particularly after a flare, multiple arches may appear, condensing at several apices, the material flowing downward along curved lines of magnetic force with velocities of 12–24 miles (20–40 kilometers) per second. Greatly enhanced activity is often triggered by a nearby solar flare that destroys the equilibrium of the magnetic field structure; the whole or a portion of a prominence may erupt from the sun with great acceleration reaching velocities of several hundreds of miles/second.

Spectroscopic studies of prominences show that they may be divided into cool, quiescent (8000–10,000°K) forms and hot, active arch-prominences with temperatures up to 100,000°K. The same elements that appear in the sun's atmosphere exist in the prominence, their appearance modified only by its temperature and pressure.

Flares

A flare is a short-lived brightening of a small area of the sun comprising an evolving bipolar magnetic structure. Optically, except for the very rare and outstanding flares visible in white light, most flares are seen through the spectrohelioscope or filter as intense bursts in the Hα light of hydrogen; they are one of the most spectacular sights connected with the sun. Flares are chromospheric phenomena. They are by no means uncommon; at the peak of the sunspot cycle there are about 20–30 flares per year of class 3, having a duration 2–4 hours, and a thousand of class 1 lasting on the average 25 minutes. Preflare activity is often noted in quiescent prominences which takes the form of enhanced internal motions, a slow rise into the corona, and when the flare starts sudden acceleration and complete disappearance, sometimes to reappear with hours or days in nearly the same shape and position. The flare phenomenon often includes the ejection of material: surges from small flares take place along lines of magnetic force, generally at an angle to the sun's surface, the material returning to the sun along the same path. Great flares are accompanied by sprays of explosively ejected material with velocities as high as 930 miles (1500 kilometers)/second.

The flare phenomenon is exceedingly complex. We will describe some of the events in a great flare. Typically, individual brightenings merge along the sides of a disappearing quiescent prominence to form two ribbons (see Fig. 7), or starting as a bright single ribbon in a young active region it divides into two which initially separate with speeds up to 62 miles (100 kilometers)/second, but soon stop. The flare starts with a very rapid increase of temperature to 5–10 million degrees Kelvin high in the corona, accompanied by a soft x-ray burst, followed by a slow exponential decrease. Next the flare energy moves down into the chromosphere, heating a thin layer to 10,000°K and becoming visible as ribbons in Hα. In the flash phase, within the first 10 minutes, electrons accelerated to 10–100 keV produce hard x-rays, impulsive radio, and far-ultraviolet bursts. Streams of electrons moving outward at 100,000 km/sec produce intense radio frequency bursts and can be tracked at low radio frequencies to the vicinity of the earth and recorded by spacecraft. In some large flares a blast wave originates, accelerating electrons and protons to hundreds of millions of electron volts which in turn when bombarding the photosphere produce nuclear reactions and associated gamma rays.

Among the many terrestrial effects of a flare is the short-wave radio fade-out (SWF). Hard x-rays emitted at the onset of large flare illuminate the sunlight side of the earth, causing an SWF for all transmission paths over this hemispherical cap. The x-rays ionize the lower layers of the earth's atmosphere (60 km), freeing electrons which, excited to oscillation by radio waves, quickly lose their energy of oscillation by collisions with abundant air molecules. The loss of signal by absorption of the radio waves may last for several hours. Very large flares have been known to produce marked changes in the earth's geomagnetic field, which in turn induce electric currents in long-distance transmission lines strong enough to trip circuit breakers and to stop telegraph communication.

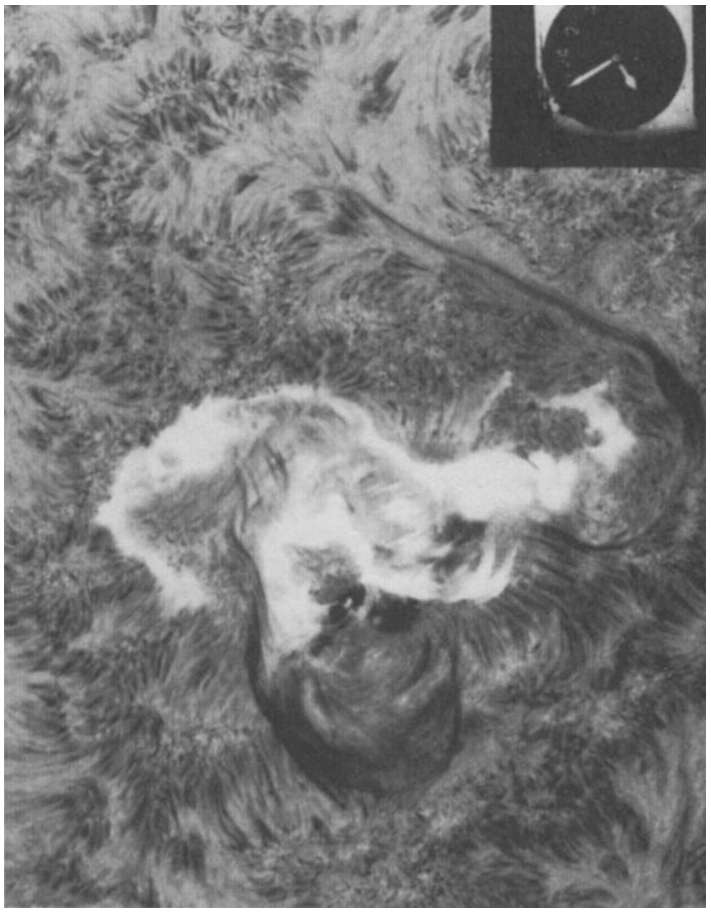


Fig. 7. An intense two-ribbon flare with associated magnetic arches. (*Big Bear Solar Observatory.*)

On March 10, 1989, a great flare occurred in a large sunspot region on the sun. Protons and electrons arriving at the earth two days later produced brilliant auroras over Canada and the United States. Transformers tied to the Hydro-Quebec power grid and to the United States saturated, overheated, and were destroyed from induced, nearly DC, earth-surface potentials of 1–10V per kilometer of transmission line. The cost of outages and loss of transformers exceeded tens of millions of dollars from this one event.

The Solar Constant

The solar constant is defined as the quantity of solar energy received at normal incidence outside the earth's atmosphere at the earth's mean distance from the sun. It is expressed in calories per square centimeter per minute or kilowatts per square meter. The latest measurements give 1.968 cal $\mathrm{cm}^{-2}\ \mathrm{min}^{-1}$ or 1.373 kW m^{-2}.

Variations in the solar output in x-ray, far infrared, and radio wavelengths are well known, but very little energy comes from these wavelengths (less than 1.0%), hence the larger question of any appreciable variation in the total luminosity of the sun over long periods of time goes begging. The evidence from paleontology suggests that it has been nearly constant (less than 10–15% variation) over several billion years. Today a variation of 0.5% or less is of interest to climatologists. For decades, C. G. Abbot with instruments of his own devising made a long series of measurements from mountains located in arid lands, particularly the Atacama desert of Northern Chile. He observed variations of up to 1.5% which he attempted to correlate with rainfall. Recent measures from satellites show variations up to $\pm$ 0.04%.

Solar Energy Generation—The Neutrino Puzzle

Mass can be converted into energy. Substituting the energy output of the sun (3.86×10^{33} ergs/sec) into Einstein's famous formula $E = mc^2$, we find that m, the mass loss, amounts to 4,700,000 tons per second. Large as this is, it is only 10^{-11} of the solar mass per year. In the sun's core, between center and one-tenth of the solar radius, the most important fusion process is the proton-proton reaction in which hydrogen nuclei, called protons, are converted into a stable helium nucleus. The reaction is:

$$2 \times \begin{cases} H^1 + H^1 \rightarrow H^2 + e^+ + \nu \\ H^2 + H^1 \rightarrow He^3 + \gamma \\ He^3 + He^3 \rightarrow He^4 + H^1 + H^1 \end{cases}$$

In this chain reaction, two protons fuse giving deuterium, a positron, and a neutrino. The neutrino escapes the sun at the velocity of light, carrying a small amount of energy with it. The positron collides with an electron and both are annihilated, with the release of gamma rays. The fusion of the deuterium nucleus with another abundant proton results in the light isotope of helium, with two protons and one neutron, and the emission of more gamma rays. Finally, two helium-3 atoms combine to make a stable helium nucleus plus two nuclei of hydrogen. With a loss of six hydrogens and two electrons we gain one helium and two hydrogens, additionally two neutrinos and 5 gamma rays are released in the process. From the known weights of He^4 and H^1 one calculates that 4.3×10^{-5} ergs is released for every He^4 nucleus formed.

Ninety percent of the sun's energy is produced through this reaction. Another reaction chain of the proton-proton cycle gives beryllium and boron with the emission of gamma rays, a positron, and a high energy (14 MeV) neutrino. It is this neutrino that can be detected by its absorption in a chlorine 37 nucleus, giving a radioactive argon 37 atom. The experiment performed by Davis uses 100,000 gallons (3785 hectoliters) of perchloroethylene enclosed in a sealed tank situated in a deep gold mine in South Dakota. One atom per day of radioactive argon is observed. This rate falls short by a factor of 3 from the rate calculated from the standard model of the sun, that is, its temperature, pressure, energy output, and chemical composition. The discrepancy can be equated either to a lack of knowledge of the sun's interior, or to a lack of knowledge of the neutrino's properties. This experiment has stimulated a great deal of theoretical work in solar and neutrino physics.

Low-energy neutrinos from the proton-proton fusion reaction, which contributes most of the neutrinos, can be detected through a nuclear transmutation of gallium, by neutrino capture, giving radioactive germanium. At the earth's distance from the sun, there are 1000 billion billion of these neutrinos passing through our bodies every second. A neutrino, moving with the velocity of light and having passed from the sun's core to its surface (and through the earth if the sun is on the other side and we are in darkness), spends very little time in the neighborhood of a gallium atom. The chance of capture is very, very small. Thirty tons of gallium in solution sits in a laboratory under a high peak in the Italian Apennines, and in another experiment 57 tons of gallium metal is deep under a mountain in the Caucasus. Thirty tons of gallium should yield 1.2 atoms of germanium per day. From 300 days of collection, the number of captures was 63% of that predicted. Further tests are under way.

Solar Seismology

Like the earth, the sun also oscillates with a variety of shapes and frequencies. Though such oscillation is extremely difficult to detect from 93 million miles (~150 million kilometers) away, from precise measurements of its shape and from Doppler measurements of the up and down motion of its surface we can now probe the sun's interior. The 5-minute oscillation, discovered in the 1960s, of large groups of solar granules covering areas 2480–9300 miles (4000–15,000 kilometers) across, represents a standing pattern of sound waves trapped within the convection zone. Globally the entire convection zone (124,000 miles; 200,000 kilometers deep) is oscillating, with several nodes around the sun's circumference and with a number of nodes between the surface and the bottom of the zone. See Fig. 8. The observations are interpreted in terms of spherical harmonics of low-angular (l) and high-radial ($n \sim 15$) overtones with an amplitude of 10–100 cm/sec. l-values of 0, 1, 2, 3, 4 have been observed in a 5-day doppler-shift measurement recorded in Antarctica at the South Pole during its summer, where the sun may be seen continuously for long periods of time. Because of the global coherence in space and time it has been possible to derive the depth of the convection zone and the solar rotation rate down to the bottom of the zone, where it is found that the sun rotates 5% faster than at the surface. In addition to the well established 5-minute oscillation, work

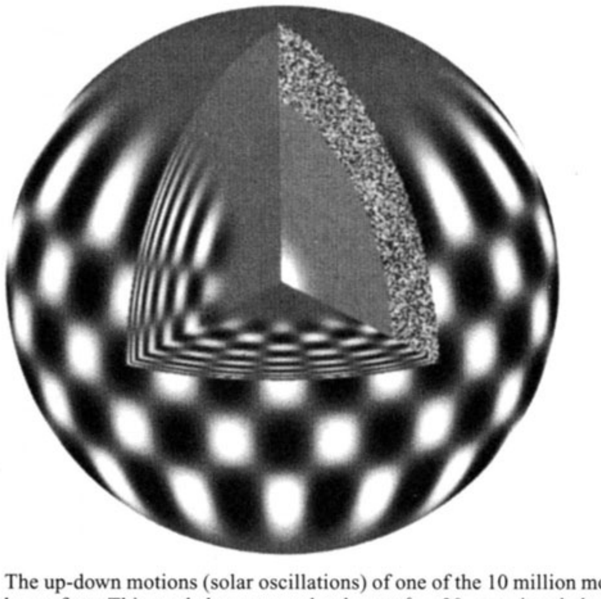


Fig. 8. The up-down motions (solar oscillations) of one of the 10 million modes of the solar surface. This mode has an angular degree $\ell = 20$, an azimuthal order $m = 16$, a radial order $n = 15$, and a frequency corresponding to a period of 5 minutes. (*Global Oscillation Network Group, National Solar Observatories.*)

in Crimea, the United Kingdom, and the United States has yielded a 160-minute period of oscillation.

The Sun's Motion in Space

Our sun, together with the stars in its immediate neighborhood, slowly orbits around the galactic center, located 33,000 light years away in the direction of the Milky Way constellation of Sagittarius, moving with a velocity of 155 miles (250 kilometers) per second. The "cosmic year" or period of rotation is 2400 million years. Additionally the sun has its own motion within the local group of naked eye stars. A study of radial velocities and proper motions places the direction in which the sun is moving near the summer constellation of Lyra. From the radial velocities its speed toward the solar apex is found to be approximately 12.1 miles (19.5 kilometers) per second.

Solar Telescopes and Instrumentation

Because of the enormous light and heat of the sun, and because the full surface of the star can be examined, unusual telescopes and instruments have evolved. The size of the sun's image in a camera or telescope is $\frac{1}{108}$th of the focal length. Thus, in the common 35-millimeter camera of 50 millimeters focal length, the sun's image on the film is only $\frac{1}{2}$ millimeter $\frac{1}{50}$th inch). The McMath-Pierce Solar Telescope Facility at Kitt Peak, Arizona, with a 300-foot (91.4-meter) focal length, yields an image 3 feet (0.9 meter) in diameter. Like most solar telescopes, the McMath-Pierce instrument is not pointed to the sun, but is fixed in direction. Sunlight goes 500 feet (152 meters) down the south polar axis at an angle of 32°. The heliostat mounting is driven to follow the sun during the course of the day. The beam, focused by a 60-inch (152-centimeter) concave mirror, is returned slightly below the incoming light to a 48-inch (122-centimeter) flat that directs the image vertically downward to spectrographs and other instrumentation located in the observing room. Great care has been taken to eliminate thermal disturbances which distort the image. Thus, the 80-inch (203-centimeter) mirror is located 100 feet (30.5 meters) above the local terrain to eliminate ground effects. Additionally, the whole telescope, two-thirds of it underground, is cooled to ambient air temperature. The problem of internal seeing in solar telescopes is so great that the bold step of evacuating the whole telescope has recently been taken. The newer 30-inch (76.2-centimeter), 180-foot (54.9-meter) focal length tower at Sacramento Peak in New Mexico, and the 24-inch (61-centimeter) Cassegrain instrument located at San Fernando, California are good examples of vacuum telescopes.

The McMath-Pierce solar telescope is illustrated in the diagram of Fig. 9. An overall external view of the telescope is shown in Fig. 10. An image is shown in Fig. 11.

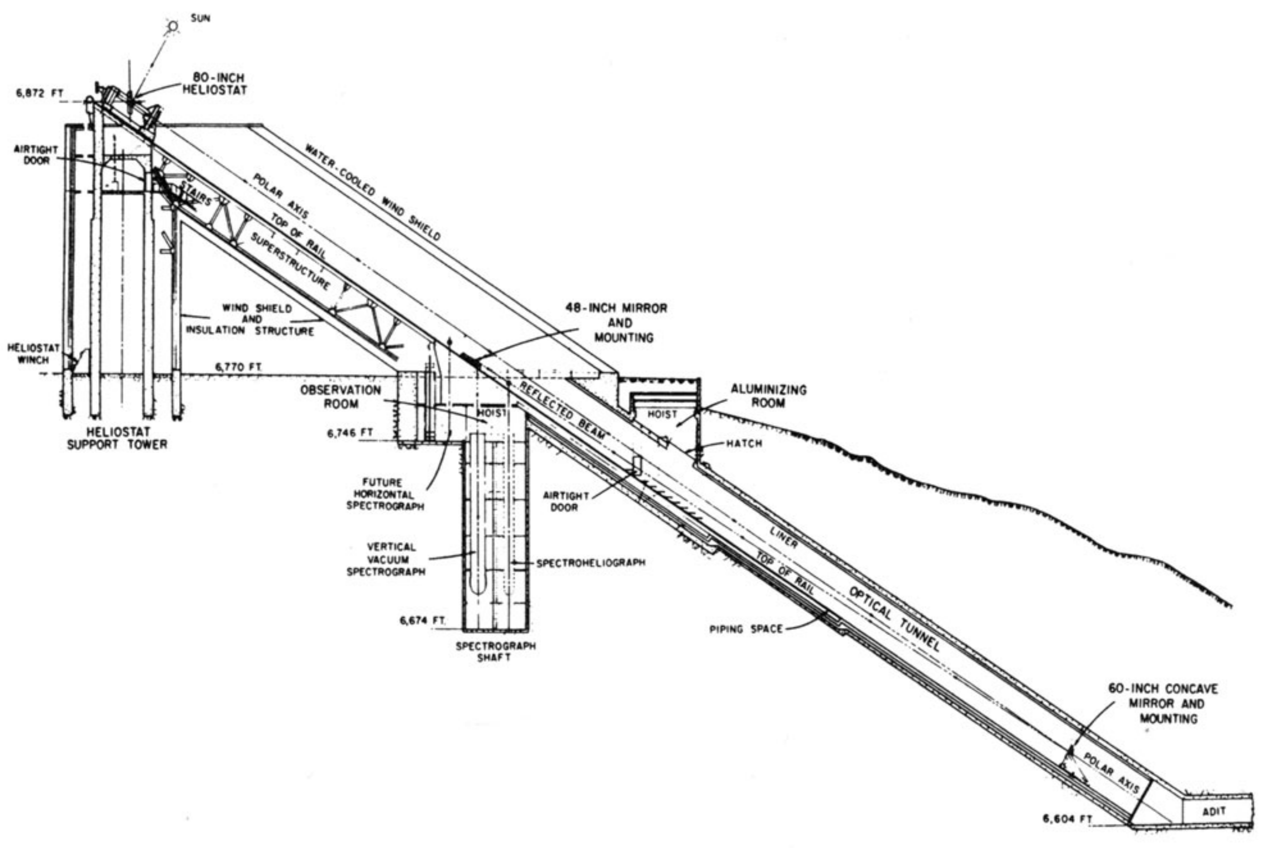


Fig. 9. Sectional view of the McMath-Pierce solar telescope located at Kitt Peak, Arizona. (*Kitt Peak National Observatory.*)

Coronagraph. A telescope for observing the corona surrounding the sun, at times other than at a solar eclipse, is called a *coronagraph*. An artificial eclipse of the sun is produced by placing an opaque circular disk slightly larger than the image of the sun at the focus of the telescope. Light from the corona passing the edge of the occulting disk is reimaged by a lens onto a photographic plate or other detector. Suitable diaphragms capture the defracted light from the objective. The brightness of the corona is one millionth that of the sun, hence, the success of the instrument depends upon the great care in selecting the glass for the singlet objective, free of bubbles and stria, and in polishing the surfaces free of all scratches. Furthermore, great pains must be taken to avoid dust on the objective. To minimize the atmospheric scattering, coronagraphs are mounted at high elevation sites (10,000 feet; 3050 meters).

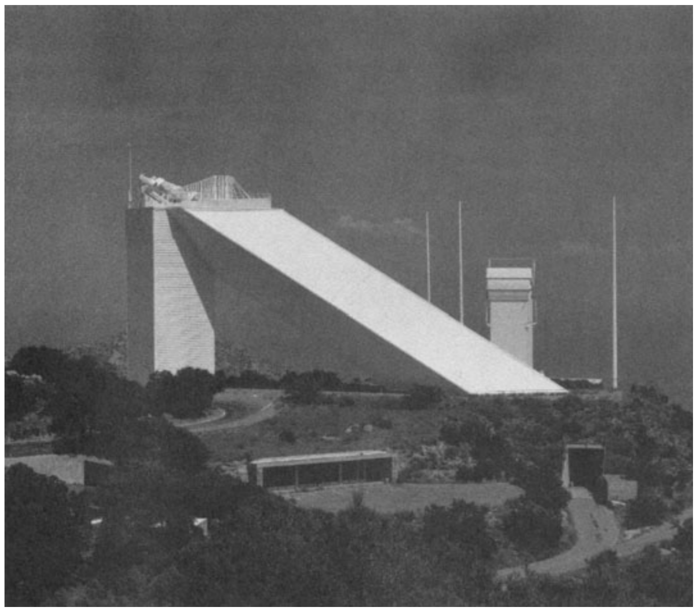


Fig. 10. Heliostat support tower and part of water-cooled wind shield of the McMath-Pierce solar telescope. (*Kitt Peak National Observatory.*)

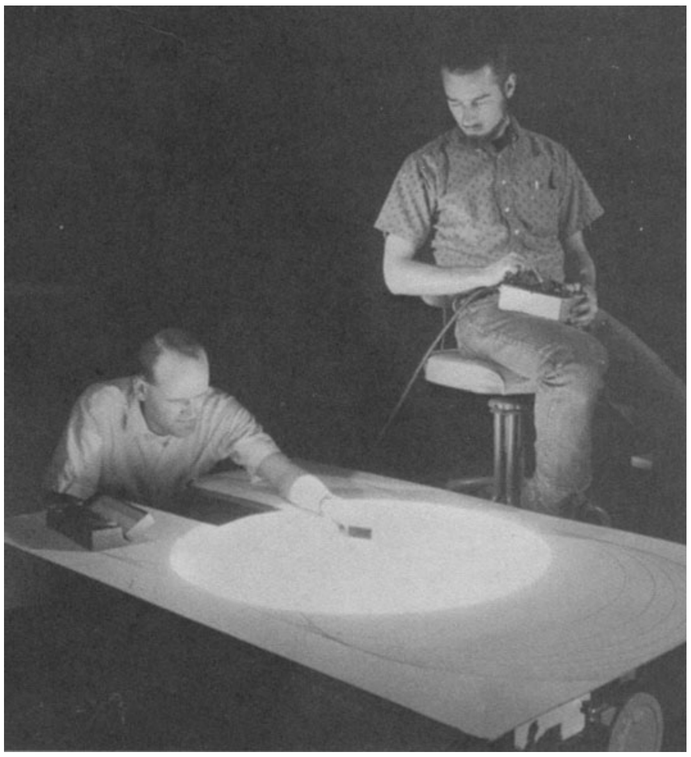


Fig. 11. Inspection of solar image. (*Kitt Peak National Observatory.*)

The appearance of strong emission lines in the spectrum of the corona permits the instrument to be used monochromatically with narrow band polarizing filters or with a spectrograph.

Spectroheliograph. Essentially the instrument consists of a high-dispersion spectrograph with a second slit placed directly in front of the photographic plate so that the radiation from only one spectral line is received on the plate. If the instrument is so placed that the first slit is in the principal focus of a telescope directed toward the sun, a narrow strip of the sun's image will be admitted by the first slit and an image of that narrow section of the sun will be formed on the photographic plate in the particular radiation for which the second slit is adjusted. The instrument is so constructed that the first slit may be moved across the image and at the same time the second slit will move across the photographic plate at the same rate. Hence, it is possible to obtain a photograph of the sun in the almost monochromatic radiation of any particular element, say calcium.

Birefringent Filters. B. Lyot (1933) investigated the transmission of a series of crystalline plates in polarized light. Independently in 1937, Y. Ohman built such a filter capable of isolating the Hα line of hydrogen and was able to photograph prominences through it. The filter consists of a series of quartz plates in which each plate has a thickness of one-half the preceding plate. Polaroid is sandwiched between each plate. By suitable design, a narrow passband (0.25–5 Å) can be obtained anywhere within the spectrum. With their use, the whole sun can be examined in the light of one element, usually hydrogen. These filters have been widely used particularly in cinematography of solar phenomena.

Solar Experiments in Space

The selective absorption of ozone, carbon dioxide, and water vapor in the earth's atmosphere effectively blocks the ultraviolet solar spectrum short of 2900 angstroms and in bands beyond 7000 angstroms in the infrared. The earliest (1946) rockets carrying small spectrographs above 70 kilometers (43 miles) extended the solar spectrum from the ultraviolet ozone cut off at 2900 angstroms to 2100 angstroms. These unstabilized instruments were soon superseded by high resolution spectrographs on stabilized platforms capable of guiding to a fraction of a second of arc. The *Skylab* manned mission of 1973 mounted a complex of instruments designed to obtain high resolution direct photographs of the sun, prominences and flares in x-ray and far UV wavelengths. The disk of the sun was observed to be covered with thousands of x-ray bright points bringing magnetic fields to the surface. *Skylab*'s white light coronagraph observed for the first time huge blast waves—coronal transients—moving outward through the corona at speeds of 600 km (375 mi) per second, set in motion by an underlying flare.

The solar maximum mission of 1980 carried seven instruments into space: a gamma ray spectrometer, several x-ray spectrometers, a coronagraph, and an active cavity radiometer for measuring the solar irradiance. This instrument is still (1987) providing data. Future programs look to high resolution imaging of the solar granulation.

The Six Basic Problems of Solar Physics

Giovanelli has listed five problems of solar physics to which we add a sixth, i.e., the neutrino puzzle: (1) the cause of the sunspot cycle, (2) the structure of the convection zone comprising the outer 200,000 km (125,000 mi) of the sun's radius, (3) differential rotation with latitude and depth in the sun's atmosphere, (4) the cause of flares, and (5) the heating mechanism of the chromosphere and corona.

Additional Reading

Allen, C. W.: "Astrophysical Quantities," 3rd edition, Athlone, Univ. of London (1973).
Bahcall, J. N.: "Neutrino Astrophysics," Cambridge Univ. Press, New York (1989).
Bahcall, J. N.: "The Solar Neutrino Problem," *Sci. American*, **262**, May, 54 (1990).
Billings, D. E.: "A Guide to the Solar Corona," Academic, New York (1966).
Bray, R., and R. Loughead: "Sunspots," Longmans, London (1964).
Christensen-Dalsgaard, J.D. Gough, and J. Toomre: "Seismology of the Sun," *Science*, **229**, 923 (1985).
Cox, A. N., W. C. Livingston, and M. S. Matthews, Editors: "Solar Interior and Atmosphere," Univ. of Arizona Press, Tucson, Arizona (1991).
Eddy, J. A.: "The New Sun, the Solar Results from Skylab," National Aeronautics and Space Administration, Washington, D.C. (1979).
Giovanelli, R. G.: "Secrets of the Sun," Cambridge Univ. Press, New York (1984).
Harvey, J. W., J. R. Kennedy, and J. W. Leibacher: "GONG: To See Inside Our Sun," *Sky and Telescope*, **470** (1987).
Kappenman, J. G., and V. D. Albertson: "Bracing for the Geomagnetic Storms," IEEE, Spectrum, 27, March, p. 27 (1990).
Meadows, A. J.. "Early Solar Physics," Pergamon, Oxford (1970).
Noyes, R. W.: "The Sun, Our Star," Harvard Univ. Press, Cambridge, Massachusetts (1982).
Pasachoff, J. M., and M. L. Kutner: "University Astronomy," Saunders, Philadelphia (1978).
Pepin, R. A., J. A. Eddy, and R. B. Merrill: "The Ancient Sun," *Geochimica et Cosmochimica Acta*, Supp. 13, Pergamon Press, New York (1980).
Pierce, A. K.: "The McMath Solar Telescope of the Kitt Peak National Observatory," *Applied Optics*, **3**, 12, 1337–1346 (1964).
Smith, F. G.: "Radio Astronomy," Penguin Books, Baltimore (1960).
Sonett, C. P., M. S. Giampapa, and M. S. Matthews, Editors: "The Sun in Time," Univ. of Arizona Press, Tucson, Arizona (1991).
Svestka, Z.: "Solar Flares," Reidel, Dordrecht, Holland, 1976.
White, O. R.: "The Solar Output and Its Variation," Univ. Colorado, Boulder, Colorado (1977).
Zirin, H.: "Astrophysics of the Sun," Cambridge Univ. Press, Cambridge (1988).
Zirker, J. B.: "Total Eclipses of the Sun," Van Nostrand Reinhold, New York (1984).

SUN COMPASS. A device utilizing the direction of the sun for direction or orientation purposes. The sun compass operates on much the same principle as that of the sun dial. In the sun dial, the gnomon for casting the shadow is set accurately parallel to the earth's axis of rotation, and the direction of the shadow indicates local apparent time. In the sun compass, the dial is set for local apparent time, and the direction of the shadow is used in connection with a compass card. The instrument is quite complicated, for it must be set for terrestrial latitude, longitude, and local apparent time. It has been of great service in connection with flights in the polar regions of the earth, where the weakness and uncertainty of the horizontal component of the earth's magnetic field render the use of the magnetic compass most unreliable.

SUNDIAL. It is logical to suppose that from the earliest times mankind has used the apparently moving sun as a means for reckoning time. As the sun appears to move across the heavens during the day, the position and length of the shadow cast by an opaque rod will continually change. The positions of lengths of this shadow may be used for the purpose of subdividing the period between sunrise and sunset. Any device that utilizes the shadow cast by the sun for the purpose of subdividing the day into equal parts is known as a sun dial.

It is difficult to say just when the first sun dial was constructed. The earliest written record is found in Isaiah 38:8, which was written approximately 700 years before the Christian era. The earliest instrument still existing is one built in Egypt, for which the exact date of construction is unknown. Sun dials came into general use during the thirteenth century, and the different types were then rapidly developed. By the time that mechanical clocks and watches made their appearance in the fifteenth century, a multitude of types of sun dials had been constructed, and many volumes had been written regarding the theory of the various devices.

There are two fundamental types of sun dials: fixed and portable. The most common fixed type marks the divisions of the day by means of the shadow thrown by the sun. The dial itself may be set at any desired angle, but generally, the plate is horizontal, and the style, which casts the shadow, is so placed as to be parallel to the axis of rotation of the earth. The portable dial makes use of the fact that the length of the shadow of the sun varies throughout the day, being the shortest at noon and the longest at sunrise and sunset. The great difficulty with this type is that because of the change in declination of the sun with season, it is necessary to have different scales of time for different periods of the year.

A multitude of ingenious and beautiful types of sun dials have been used in the past for the purpose of keeping time and are in use at present as ornaments or items of curiosity. In adjusting the horizontal type of sun dial, such as may be purchased from a number of dealers in garden supplies or curios, it is important to remember that the style should be parallel to the axis of the earth. That is, it should lie exactly in the true north-south plane, and the north end should be so elevated that the angle which the style makes with the horizontal plate is equal to the latitude of the observer. When properly adjusted, the sun dial will read local apparent time. This time will differ from ordinary clocktime by the longitude difference between the position of the dial and the standard time meridian, and also by the equation of time.

See **Equation of Time.**

SUNFISHES (*Osteichthyes*). Of the suborder *Percoidea*, family *Centrarchidae*, various species of sunfishes are among the favorite food and sporting fishes, particularly in North American fresh waters. The largemouth black bass, crappie, and bluegill are members of this family. Although initially American fishes, a number of species have been introduced elsewhere, notably Europe. The largest of the species is *Micropterus salmoides* (largemouth black bass) which measures up to about 30 inches (76 centimeters) and weighs up to 25 pounds (11 kilograms). The *Micropterus dolomieui* (smallmouth) is somewhat smaller, ranging up to 27 inches (69 centimeters) and weighing up to 12 pounds (5.4 kilograms). The *Lepomis macrochirus* (common bluegill) is frequently used as a small forage fish when stocking the larger species. Capable of attaining a length of about 15 inches (38 centimeters) and weight of nearly 5 pounds (2.3 kilograms), the average adult bluegill usually does not exceed about 4 inches (10 centimeters). Crappies are popular with sportsmen and have been distributed widely throughout North America. The *Pomoxis annularis* (white crappie) prefers muddy and turbid waters. In contrast, the *Pomoxis nigromaculatus* (black crappie) normally is found in clear waters. Records indicate that crappies up to 21 inches (53 centimeters) in length and 5 pounds (2.3 kilograms) have been caught. *Archoplites interruptus* (Sacramento perch), attaining a length of about 12 inches (30 centimeters), is found in the San Joaquin and Sacramento basins of California.

SUPERCHARGER. The performance of an internal combustion engine is indicated, among other things, by the brake horsepower output. A review of the factors affecting power indicates that atmospheric conditions have a significant effect. A naturally aspirated (unsupercharged) engine is able to draw into the cylinders on suction strokes only from 70–85% of the fuel charge which it is theoretically capable of inducing. Consequently, the mean effective pressures are not as large as they might be, and power output per cubic inch of piston displacement does not reach its maximum possible value. Compression of the incoming air, or air-fuel mixture, somewhat above ambient pressure is a natural way of increasing output at sea level or of regaining it at altitudes. A compressor used for this purpose is designated a *supercharger*.

In recent years in connection with automotive engines, there has been considerable emphasis on the use of a turbocharger boost device to increase downsized engine power while increasing fuel economy and, in essence, retaining packaging advantages.

In 1987, Uthoff and Yakimow (Eaton Corp.) in a paper for the Society of Automotive Engineers, observed that recently there has been a high level of activity involving the use of the mechanically driven supercharger as a boost device. The supercharger can provide improved engine torque response at low engine speeds as compared to the turbocharger. The latest production turbochargers using variable geometry housings and ceramic turbines still take four times as long as a positive displacement, mechanically driven supercharger to produce maximum boost. In addition, the turbocompressor does not reach its maximum efficiency range until high speeds and airflows are achieved later in the vehicle acceleration event. This contrasts with the almost immediate boost response of the supercharger which takes approximately 0.4 second to produce 50 kPa boost. The supercharger is continuously driven at full boost speed for the given rpm, and as soon as the bypass valve can be closed, the intake system is pressurized. There is no need to accelerate a mechanical device to high rotational speeds prior to production of boost pressure.

In testing commercially available vehicles, some equipped with turbochargers and others with superchargers, the researchers found that in attempting to accelerate from a stop in a turbocharged car resulted in a sluggish feel until 2 seconds after pressing the accelerator pedal. Then, the power rapidly climbed making it necessary to readjust the pedal position. Usually, the lack of turbocharger response at low engine speeds resulted in a large throttle opening and rich fuel during the first 3 seconds of acceleration then a gradual throttle closing. Fuel economy suffered and a high level of driver interaction was required to maintain vehicle control.

When driving a supercharged car, much less throttle modulation was needed to set the desired level of vehicle acceleration. The engine torque level increased almost immediately upon application of the throttle and did not suddenly change seconds later as with the turbocharged vehicles. This reduced the amount of driver interaction needed to maintain control of the vehicle and gave the same feel as a much larger, naturally aspirated engine under the same conditions.

SUPERCONDUCTIVITY. A property of a material that is characterized by zero electric resistivity and, ideally, zero permeability. The phenomenon of superconductivity was discovered in 1911 by Heike Kamerlingh Onnes (University of Leiden) as the outcome of a remarkable achievement in those years—the liquefaction of helium for the first time. Helium condenses at atmospheric pressure at 4.2 K. Onnes, using the newly available very low temperature substance, proceeded to investigate the electrical resistance of various metals at low temperatures. Even prior to quantum mechanics, it had been predicted that if absolute zero could be achieved in a metal having a perfectly regular interatomic structure, the electrical resistance of the metal would be zero. Onnes found that the resistance of a mercury wire suddenly dropped to zero at 4.2 K, which indeed is a very low temperature, but still well above absolute zero. Onnes and other investigators at Leiden researched superconductivity in other metals, such as lead, which was found to become superconductive at 7.2 K. It should be mentioned at this point that scientists of that period were biased in their thinking of conductivity and superconductivity in terms of metals. The first stable conducting organic material was not synthesized until 1960 and it was not until 1979 that a superconducting organic material was isolated. The so-called "hottest" superconductor was announced in mid-1993 by a research team at Eidgenossische Technische Hochschule in Zurich. The new material, considered toxic, is made of two distinct compounds that commence to be superconductive at 133 K. Prior record was claimed for $Tl_2Ca_2Ba_2Cu_3O_{10}$ at 127 K.

Rediscovery. The topic and potential of superconductivity essentially was rediscovered in the late 1970s and early 1980s, during which period the research activity safely can be described as zealous. This was fueled by the scores of probable applications of superconductors, but which to date largely remain as promises. Research continues at a good pace, and the topic is much better understood as compared with a decade ago. As pointed out later in this article, much excellent technological fallout has occurred from multimillions of dollars invested in research. The search for the ultimately practical superconductor, although still elusive, has reinforced the multidisciplinary sciences involved, notably physics and chemistry.

Application Targets. Among the ultimately practical applications for superconductivity, especially those of significant future commercial values, are: (1) magnetic shielding, (2) magnetic resonance imaging magnets for medicine and research, (3) electric utility transmission lines and load-leveling storage coils, (4) magnetic separators for materials processing, (5) higher-speed (10×) switching and signal transmission for computers, (6) more compact electronics with finer interconnect lines, (7) extremely compact electric motors and actuators, (8) no-loss portable electrical storage "batteries," and (9) noncontact bearings and magnetically levitated vehicles. Special areas of interest by the military in superconductors include (1) infrared optical detector elements, (2) high-speed millimeter and submillimeter-wave electronics for advanced radar countermeasures, (3) magnetic anomaly detection of submarines, and (4) free-electron laser components.

Considerable research of a contemplative or assumptive nature continues to go forward—that is, studying how the "ideal" superconductor can be applied, once developed. Examples of progress are shown in Figs. 1, 2, and 3.

Fundamental Research Findings

For several years following the previously mentioned work of the Louden scholars, investigators concentrated principally on metallic elements and alloys, including indium, tin, vanadium, molybdenum, nio-

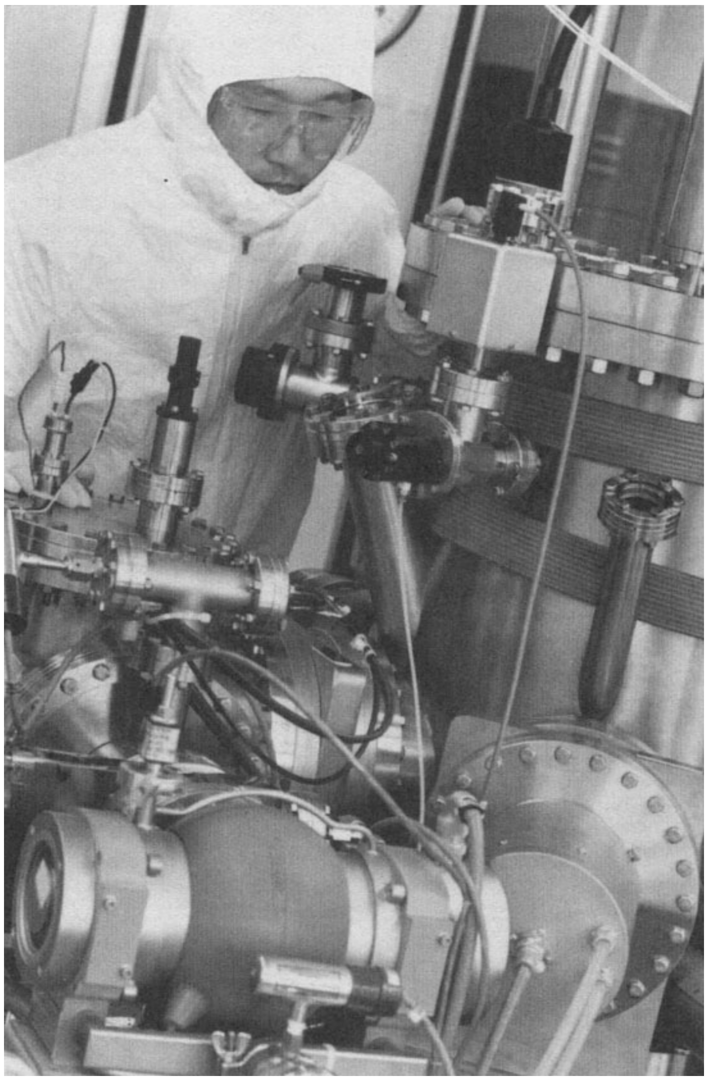


Fig. 1. Special equipment required to fabricate low-temperature superconducting junctions. Josephson junctions are comprised of aluminum oxide sandwiched between layers of niobium. These trilayer devices are considered vital to the very-high-speed signal processing demands of next-generation computers, radar, and communication systems. Shown in illustration is scientist Dr. Joonhee Kang. (*Westinghouse Electric Corporation.*)

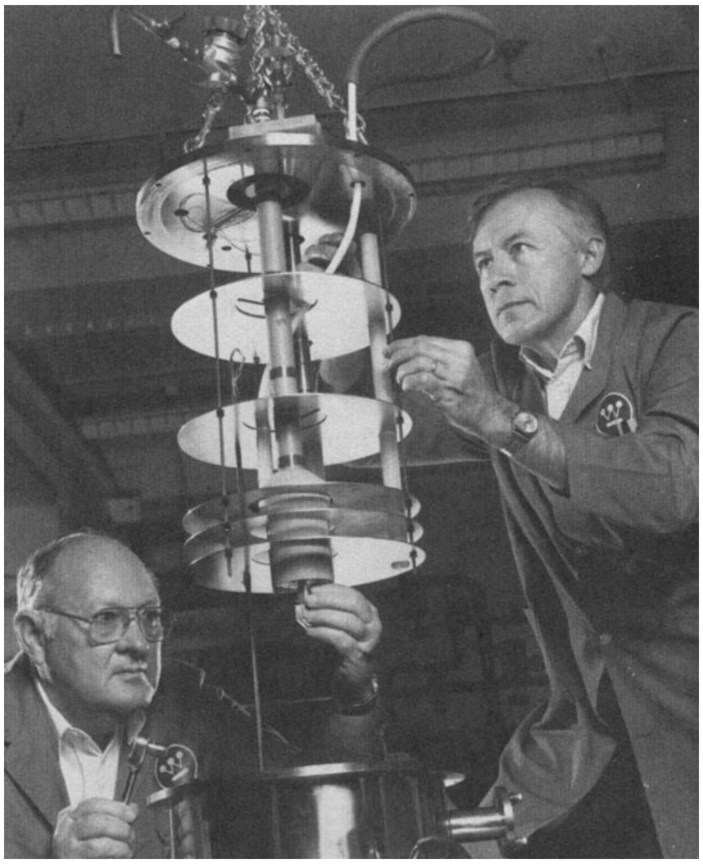


Fig. 2. Electrical lead comprised of a high-temperature superconductor can carry a current of 2000 amperes. A variety of uses include magnetic resonance imaging and superconducting magnetic energy storage. (*Westinghouse Electric Corporation.*)

bium-zirconium, and niobium-tin, among many others. A number of basic discoveries were made. For example, finding that the property of superconductivity could be destroyed by the application of a magnetic field equal to or greater than a critical field H_c. This H_c, for a given superconductor, is a function of the temperature given approximately by

$$H_c = H_0(1 - T^2/T_c^2) \tag{1}$$

where H_0, the critical field at 0K, is in general different for different superconductors and has values from a few gauss to a couple of thousand gauss. For applied magnetic fields less than H_c, the flux is excluded from the bulk of the superconducting sample, penetrating only to a small depth λ into the surface. The value of λ (called the penetration depth) is in the range 10^{-5} to 10^{-6} centimeter. Thus the magnetization curve for a superconductor is

$$\begin{array}{ll} B\text{ (inside)} = 0 & \text{for } H < H_c \\ B\text{ (inside)} = B\text{ (outside)} & \text{for } H > H_c \end{array}$$

This magnetization behavior is reversible and cannot therefore be explained entirely on the basis of the zero resistance. The reversible magnetization behavior is called the Meissner effect.

The existence of the penetration depth λ suggests that a sample having at least one dimension less than λ should have unusual superconducting properties, and such is indeed the case. Thin superconducting films, of thickness d less than λ, have critical fields higher than the bulk critical field, approximately in the ratio of λ to d. This result follows qualitatively from the thermodynamics of the Meissner effect: the metal in the superconducting state has a lower free energy than in the normal state, and the transition to the normal state occurs when the energy needed to keep the flux out becomes equal to this free energy difference. But in the case of a thin film with $d < \lambda$, there is partial penetration of the flux into the film, and thus one must go to a higher applied field before the free energy difference is compensated by the magnetic energy.

It is clear that the existence of the critical field also implies the existence of a critical transport electrical current in a superconducting wire, i.e., that current I_c which produces the critical field H_c at the surface of the wire. For example, in a cylindrical wire of radius r, $I_c = \frac{1}{2}rH_c$. This result is called the Silsbee rule.

All of the above properties distinguish superconductors from "normal" metals. There is another very important distinction, which contains a clue to understanding some of the properties of superconductors. In a normal metal at 0K, the electrons, which obey Fermi statistics, occupy all available states of energy below a certain maximum energy called the Fermi energy ζ. Raising the temperature of the metal causes electrons to be singly excited to states just above the Fermi energy. There is for all practical purposes a continuum of such excited energy states available above the Fermi energy. The situation is quite different in a superconductor; it turns out that in a superconductor, the lowest excited state for an electron is separated by an energy gap ϵ from the ground state. The existence of this gap in the excitation spectrum has been confirmed by a wide range of measurements: electronic heat capacity, thermal conductivity, ultrasonic attenuation, far infrared and microwave absorption, and tunneling. The energy gap is a monotonically decreasing function of temperature, having a value $\sim 3.5kT_c$ at 0K (where k is the Boltzmann constant) and vanishing at T_c.

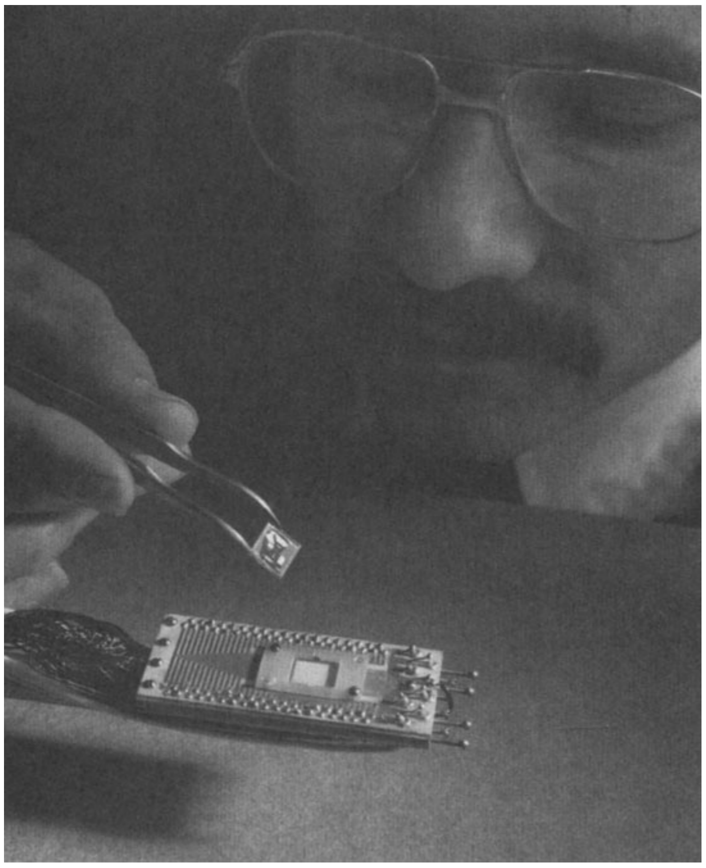


Fig. 3. Scientist Donald L. Miller holds an integrated circuit chip comprising a high-resolution superconducting analog-to-digital converter. The one-square-centimeter chip, known as a counting converter, holds promise as an unprecedented combination of high resolution and low power consumption, as needed in future air traffic control radar and infrared space-tracking applications. The 12-bit circuit (Josephson junction) has a resolution of 1 part in 4000. (*Westinghouse Electric Corporation.*)

The superconducting state has a lower entropy than the normal state, and therefore one concludes that superconducting electrons are in a more ordered state. Without, for the present, inquiring more deeply into the nature of this ordering, one can state that a spatial change in this order produced say by a magnetic field will occur, not discontinuously, but over a finite distance ζ, which is called the *coherence length*. The coherence length represents the range of order in the superconducting state and is typically about 10^{-4} centimeter, though we shall see later that it can in some superconductors take much lower values and lead to some remarkable properties.

Measurements of the transition temperature on different isotopes of the same superconductor showed that T_c is proportional to $M^{-1/2}$, where M is the isotopic mass. This isotope effect suggests that the mechanism underlying superconductivity must involve the properties of the lattice, in addition to those of the electrons. Another indication of this is given by the behavior of allotropic modifications of the same element: white tin is superconducting, while grey tin is not, and the hexagonal and face-centered cubic phases of lanthanum have different transition temperatures. A third, and most striking, indication is that the current vs voltage characteristic of a superconducting tunneling junction shows a structure which is intimately related to the phonon spectrum of the superconductor.

The superconducting properties of alloys present a bewildering variety of phenomena. They show a great deal of magnetic hysteresis, with little indication of a perfect Meissner effect. The Silsbee rule is inapplicable, and the resistive transition occurs at fields generally very much higher than in pure superconductors. For example, a wire of Nb_3Sn can carry a current of 10^5 amperes/cubic centimeter in an applied field of 100 kilogauss, while a similar wire of lead would carry about 10^3 amperes/cubic centimeter in a field of only 100 gauss. When experiments are done using well-annealed (preferably single-crystal) alloys, it is found that the critical currents drop considerably, and the magnetic behavior becomes reversible but still quite unlike that of pure superconductors. The flux is excluded from the interior of the sample up to a well-defined field H_{c1}. When the applied field is raised further, flux begins to penetrate, even though the resistance remains zero, until a second critical field H_{c2} is reached, at which the flux penetration is complete, and normal resistance is abruptly restored.

Superconductivity Theory

The theory of superconductivity has developed along two lines, the phenomenological and the microscopic. The phenomenological treatment was initiated by F. London, who modified the Maxwell electromagnetic equations so as to allow for the Meissner effect. His theory explained the existence and order of magnitude of the penetration depth, and gave a qualitative account of some of the electrodynamic properties. The treatment was extended by V. L. Ginzburg and L. D. Landau, and by A. B. Pippard, who in particular emphasized the concept of the range of coherence. A. A. Abrikosov used these ideas to develop a model for alloy superconductors. He showed that if the electronic structure of the superconductor were such that the coherence length ζ becomes smaller than the penetration depth λ, one would get magnetic behavior similar to that observed in alloys, with two critical fields H_{c1} and H_{c2}. The problem of high critical currents in unannealed (or otherwise metallurgically imperfect) alloys and compounds is more complicated because it involves the interaction between the microscopic metallurgical structure and the superconducting properties. This is an area of great research activity because of the technological implication to be mentioned later.

The microscopic theory of superconductivity was initiated by H. Fröhlich, who first recognized the importance of the interactions of electrons with lattice vibrations and in fact predicted the isotope effect before its experimental observation. The detailed microscopic theory was developed by J. Bardeen, L. N. Cooper and J. R. Schrieffer in 1957, and represents one of the outstanding landmarks in the modern theory of solids. The BCS theory, as it is called, considers a system of electrons interacting with the phonons, which are the quantized vibrations of the lattice. There is a screened coulomb repulsion between pairs of electrons, but in addition there is also an attraction between them via the electron-phonon interaction. If the net effect of these two interactions is attractive, then the lowest energy state of the electron system has a strong correlation between pairs of electrons with equal and opposite momenta and opposite spin and having energies within the range $k\theta$ (where θ is the Debye temperature) about the Fermi energy. This correlation causes a lowering of the energy of each of these Cooper pairs (named after L. N. Cooper who first pointed out their existence on the basis of some general arguments) by an amount ϵ relative to the Fermi energy. The energy ϵ may be regarded as the binding energy of the pair, and is therefore the minimum energy which must be supplied in order to raise an electron to an excited state. We see thus that the experimentally observed energy gap follows from the theory. The magnitude ϵ_0 of the gap at 0K is

$$\epsilon_0 \approx 4k\theta \exp\left(-\frac{1}{NV}\right)$$

where N is the density of electronic states at the Fermi energy and V is the net electron-electron interaction energy. The superconducting transition temperature T_c is given by

$$3.5kT_c \approx \epsilon_0$$

It has been shown that the BCS theory does lead to the phenomenological equations of London, Pippard and Ginzburg and Landau, and one may therefore state that the basic phenomena of superconductivity are now understood from a microscopic point of view, i.e., in terms of the atomic and electronic structure of solids. It is true, however, that we cannot yet, *ab initio*, calculate V for a given metal and therefore predict whether it will be superconducting or not. The difficulty here is our ignorance of the exact wave functions to be used in describing the electrons and phonons in a specific metal, and their interactions. However, we believe that the problem is soluble in principle at least.

The range of coherence follows naturally from the BCS theory, and we see now why it becomes short in alloys. The electron mean free path is much shorter in an alloy than in a pure metal, and electron scattering tends to break up the correlated pairs, so that for very short mean free paths one would expect the coherence length to become comparable to the mean free path. Then the ratio $\kappa \approx \lambda/\zeta$ (called the Ginzburg-Landau order parameter) becomes greater than unity, and the observed magnetic properties of alloy superconductors can be derived. The two kinds of superconductors, namely those with $\kappa < 1/\sqrt{2}$ and those with $\kappa > 1/\sqrt{2}$ (the inequalities follow from the detailed theory) are called respectively type I and type II superconductors.

Challenges to Established Theories. It is interesting to note that some theoreticians struggle with describing how superconductivity occurs at high temperatures in the newer, ceramic superconductors. This is understandable because the classic theory of superconductivity is tied to metals. Most ceramic superconductors discovered to date incorporate distinctive layers of copper and oxygen atoms. One question posed by some researchers, "Is the mechanism of high-temperature superconductivity the same in hole superconductors as it is in electron superconductors?"

Researchers at the Brookhaven National Laboratory, in applying x-ray techniques to a cerium-doped electron superconductor developed at the University of Tokyo, found that the holes of a hole superconductor are linked to oxygen atoms in the copper-oxygen layers, whereas in an electron superconductor the electrons are associated with copper atoms. This is exemplary of how easy it is for former theories to become outdated when new material combinations are tested for their superconductivity.

Superconductivity Research

In 1962, B. Josephson recognized the implications of the complex order parameter for the dynamics of the superconductor, and in particular when one considers a system consisting of two bulk conductors connected by a "weak link." This research led to the development of a series of weak link devices commonly called Josephson junctions. See also **Josephson Tunnel-Junction.** These devices hold much promise for achieving ultra high-speed computers where switching time is of the order of 10^{-11} second.

Good success also has been achieved in the use of certain type II superconductors, such as Nb-Zr and Nb-Ti alloys, and Nb_3Sn, in making electromagnets. In a conventional electromagnet employing normal conductors, the entire electric power applied to the magnet is consumed as Joule heating. For a magnet to produce 100 kilogauss in a reasonable volume, the power requirement can run into megawatts. In striking contrast, a superconducting magnet develops no Joule heat because its resistance is zero. Indeed, if such a magnet has a superconducting shunt placed across it after it is energized, the external power supply can be removed, and the current continues to flow indefinitely through the magnet and shunt, maintaining the field constant. Superconducting magnets have been constructed producing very strong fields in usable volumes. There is a natural upper limit to the critical field possible in such superconductors, given by the paramagnetic energy of the electrons (due to their spin moment) in the normal state becoming equal to the condensation energy of the Cooper pairs in the superconducting state. This leads to a limit of about 360 kilogauss for a superconductor with a T_c of 20K.

As investigators accumulated data upon data, many emphasized the practical as well as theoretical aspects of superconductors. The ultimate superconductor, of course, would be one that operated at room temperature or above. The materials must be manufacturable in a useful form, such as strong ductile wires for high-field magnets, electrical machinery, and power transmission lines, situations which could be even more important in commercial and industrial application than their value to science per se. (Traditionally, superconducting materials have been hard, brittle, and difficult to process.) Although superconductors that would operate at room temperature and above present a long-range target, lesser targets, including practical ways to cool them with liquid nitrogen instead of liquid helium and possibly, even better, operate them within a closed-cycle refrigeration system is the goal in the shortrange. Useful superconductors in large-scale applications must retain their properties not only at high temperatures, but also in the presence of high magnetic fields and while carrying large electrical currents. Praveen Chaudhari (IBM) has observed that new superconductors will enter the marketplace rapidly when intensive materials engineering produces easily cooled, mechanically robust conductor configurations that can handle high current densities (100,000+ A/cm^2) under powerful magnetic fields (10+ T), while maintaining stable superconductivity.

Johannes Georg Bednorz and Karl Alexander Mueller (IBM Zurich Research Laboratory) after several years devoted to a study of oxide compounds (not in terms of superconductivity) proceeded with the working hypothesis that an increase in the density of charge carriers in a material (either as electrons or as positively charged "holes") possibly would lead to a rise in transition temperature. They commenced a search for nickel- and copper-containing oxides. Early in 1986, they found a certain form of barium lanthanum copper oxide that evidenced the onset of superconductivity at temperatures as high as 35 K (12 degrees over the previous record). They encountered skepticism because the facts did not square with accepted theory that limits the phenomenon to well below 35 K. Shortly thereafter, however, researchers at the University of Tokyo, the University of Houston, and AT&T Bell Laboratories confirmed the Bednorz-Mueller findings. On October 14, 1987, the Nobel prize in physics was awarded to these two researchers and a speaker for the Royal Swedish Academy of Sciences observed that their work inspired "the explosive development in which hundreds of laboratories the world over commenced work on similar material." [It should be observed that Ching-Wu Chu and colleagues (University of Houston) did announce in February 1987 that a related class of ceramics (a certain form of yttrium barium copper oxide) remained superconducting up to 94 K, proclaiming that to be the first superconductor which could be cooled by liquid nitrogen (bp = 77 K) instead of requiring helium.] That announcement in itself also precipitted a "rush" of researchers to the ceramics.

Technological Fallout of Superconductivity Research

While in the course of finding viable superconductors for commercial applications, researchers have produced valuable ancillary information.

Quantization of Energy. In a scholarly paper, D. G. McDonald (U.S. National Institute of Standards and Technology, Boulder, Colorado) observes, "Ideas about quantized energy levels originated in atomic physics, but research in superconductivity has led to unparalleled precision in the measurement of energy levels. Microscopic things can be identical; macroscopic things cannot. This proposition is so imbued in the minds of physicists that it is interesting to see that it is false in the following sense. In the past, physicists believed that only atoms and molecules could have identical states of energy, but recent experiments have shown that much larger bodies, superconductors in macroscopic quantum states, have equally well-defined energies." In the paper, McDonald uses the novelty of the Josephson effect to illustrate the primary point of the technical paper. See McDonald reference listed.

Structural Chemistry. In an enlightening paper, R. J. Cava (AT&T Bell Laboratories) asserts, "The discovery of high-temperature superconductivity in oxides based on copper and rare and alkaline earths at first caught the solid-state physics and materials science communities completely by surprise. Since the earliest 30 to 40 K superconductors based on $La_{2-x}(Ca,Sr,Ba)_xCuO_4$, many new superconducting copper oxides have been discovered, with ever-increasing chemical and structural complexity. The current record transition temperature is held[1] by $Tl_2Ba_2Ca_2Cu_3O_{10}$, a material whose processing requires the stoichiometric control of five elements, each with considerably different chemical characteristics." In the Cava paper, the crystal structures of the known copper oxide superconductors are described, with particular emphasis on the manner in which they fall into structural families. The local charge picture—a framework for understanding the influence of chemical composition, stoichiometry, and doping on the electrical properties of complex structures—is also described.

This probing of complex and previously unattended solid materials typifies technical fallout from superconductor research.

Impact on Materials Processing and Chemical Engineering. In an interesting paper, R. Kumar (Indian Institute of Science, Bangalore) points out how processing considerations for achieving high-temperature superconductors has introduced new process engineering problems not contemplated heretofore. In a paper (reference listed), Kumar ob-

[1]As of 1990.

serves that processes involve multicomponent solid-solid reactions; mixing of fine powders; simultaneous precipitation of many ions from solutions, emulsions, microemulsions, and liquid membranes; the flow of cohesive powders with and without binders; the flow of thin films over partially wetted particles; grain boundary growth and composition; quick evaporation using pulsed lasers; mixing of molecules during their flight paths and the influence of oxygen jets; deposition of particles on substrates; and other relatively unfamiliar processing techniques.

Additional Reading

Amato, I.: "Finally, a Hotter Superconductor," *Science*, 755 (May 7, 1993).

Beardsley, T. M.: "Unsuperconductivity," *Sci. Amer.*, 22D (April 1989).

Bishop, D. J., Gammel, P. L., and D. A. Huse: "Resistance in High-Temperature Superconductors," *Sci. Amer.*, 48 (February 1993).

Brosha, E. L., et al.: "Metastability of Superconducting Compounds in the Y-Ba-Cu-O System," *Science*, 196 (April 9, 1993).

Caruana, C. M.: "Superconductivity: The Near and Long Term Outlook," *Chem. Eng. Progress*, 72 (May 1988).

Cava, R. J.: "Structural Chemistry and the Local Charge Picture of Copper Oxide Superconductors," *Science*, 656 (February 9, 1990).

Conradson, S. D., Raistrick, I. D., and A. R. Bishop: "Axial-Oxygen-Centered Lattice Instabilities and High-Temperature Superconductivity," *Science*, 1394 (June 15, 1990).

Erwin, S. C., and W. E. Pickett: "Theoretical Fermi-Surface Properties and Superconducting Parameters for K_3C_{60}," *Science*, 842 (November 8, 1991).

Fisk, Z., et al.: "Heavy-Electron Metals: New Highly Correlated States of Matter," *Science*, 33 (January 1, 1988).

Fisk, Z., and G. Aeppli: "Superstructures and Superconductivity," *Science*, 38 (April 2, 1993).

Foner, S., and T. P. Orlando: "Superconductors: The Long Road Ahead," *Technology Review (MIT)*, 36 (February 1988).

Gabelle, T. H., and J. K. Hulm: "Superconductivity—The State that Came in from the Cold," *Science*, 367 (January 22, 1988).

Haroche, S., and J-M Raimond: "Cavity Quantum Electrodynamics," *Sci. Amer.*, 54 (April 1993).

Hazen, R. M.: "Perovskites," *Sci. Amer.*, 74 (June 1988).

Iqbal, Z., et al.: "Superconductivity at 45 K in Rb/Tl Codoped C_{60} and C_{60}/C_{70} Mixtures," *Science*, 826 (November 8, 1991).

Ishigiuro, T., and K. Yamaji: "Organic Superconductors," Springer-Verlag, New York, 1990.

Kumar, R.: "Chemical Engineering and the Development of Hot Superconductors," *Chem. Eng. Progress*, 17 (April 1990).

Laughlin, R. B.: "The Relationship Between High-Temperature Superconductivity and the Fractional Quantum Hall Effect," *Science*, 525 (October 28, 1988).

Little, W. A.: "Experimental Constraints on Theories of High-Transition Temperature Superconductors," *Science*, 1390 (December 9, 1988).

Luss, D., et al.: "Processing High-Temperature Superconductors," *Chem. Eng. Progress*, 40 (September 1989).

McDonald, D. G.: "Superconductivity and the Quantization of Energy," *Science*, 177 (January 12, 1990).

Murphy, D. W., et al.: "Processing Techniques for the 93 K Superconductor $Ba_2YCu_3O_7$," *Science*, 922 (August 19, 1988).

Pool, R.: "Superconductor Patents: Four Groups Duke It Out," *Science*, 931 (September 1, 1989).

Pool, R.: "Superconductivity Stars React to the Market," *Science*, 373 (January 25, 1991).

Ross, P., and R. Ruthen: "Squeezed Hydrogen Forms Metal with Superconducting Potential," *Sci. Amer.*, 26 (November 1989).

Shumay, W. C., Jr.: "Superconductor Materials Engineering," *Advanced Materials & Processes*, 49 (November 1988).

Sleight, A. W.: "Chemistry of High-Temperature Superconductors," *Science*, 1519 (December 16, 1988).

Staff: "Trying to Cooperate in Order to Compete," *Technology Review (MIT)*, 13 (February/March 1991).

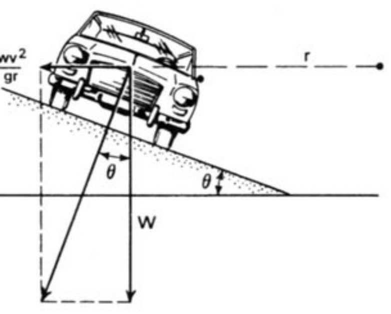
Stix, G.: "Superconducting SQUIDS," *Sci. Amer.*, 112 (March 1991).

Sun, J. Z., et al.: "Elimination of Current Dissipation in High Transition Temperature Superconductors," *Science*, 307 (January 19, 1990).

Wolsky, A. M., Giese, R. F., and E. J. Daniels: "The New Superconductors: Prospects for Applications," *Sci. Amer.*, 60 (February 1989).

SUPERCOOLING. The cooling of a liquid below its freezing point without the separation of the solid phase. This is a condition of metastable equilibrium, as is shown by solidification of the supercooled liquid upon the addition of the solid phase, or the application of certain stresses, or simply upon prolonged standing.

SUPERELEVATION. When the plane of a roadway is tilted on a curve (commonly known as banked), it is said to be superelevated. The purpose of superelevation is to permit a vehicle to round a curve on a roadway at high speed without danger of overturning or skidding. The superelevation can be made so that the resultant of dead weight and centrifugal force passes through the vertical plane of symmetry of the vehicle. In this condition, no side sway would be felt by the occupants. However, the superelevation necessary to accomplish this is different for each vehicle speed, so that it is apparent that the superelevation of a highway presupposes an average vehicle speed. The same is true of railways, although the variation of speeds with which the trains round curves is less than in the case of highway traffic.



Demonstration of superelevation principle applied to highway.

To illustrate how the superelevation depends upon vehicle speed, let it be assumed that an automobile approaches a curve on a highway at a speed of V (feet per second). If the radius of the turn is r, the centrifugal acceleration is V^2/r. Furthermore, assume that the weight is W pounds. While negotiating the curve, the car is subject to two forces, one, the weight vertically downward, the other, centrifugal force acting horizontally away from the center of curvature, and having a magnitude WV^2/gr. The surface of the road must be perpendicular to the resultant for no "side sway." If superelevation is given as the angle of bank (see figure), the angle of superelevation θ has a tangent equal to centrifugal force divided by weight. This tangent is V^2/gr, demonstrating that the superelevation must be made with respect to the radius of curvature and the velocity of the vehicle. It is independent of the dimensions and weight of the vehicle.

See also **Spiral Curve.**

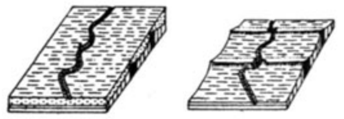
SUPERFINISHING. An abrasive process for removing smear metal, scratches and ridges produced by machining and grinding operations, and other surface irregularities from parts that are to have a highly finished surface. The process resembles lapping in that a lubricated abrasive stone is applied to the surface at comparatively low speeds and light pressures. A superfinishing head whose base is attached to the cross-slide of an engine lathe may be used. The base supports two vertical cylindrical guides on which the head proper may be manually adjusted to the work by a hand-operated lever. The abrasive stone is carried in a vertical slide which is subjected to the action of a spring for applying pressure to the stone. The stone pressure may be regulated to suit the requirements of the work by turning the screw at the top of the slideway.

SUPERFLUIDITY. The term used to describe a property of condensed matter in which a resistance-less flow of current occurs. The mass-four isotope of helium in the liquid state, plus over 20 metallic elements, are known to exhibit this phenomenon. In the case of liquid helium, these currents are hydrodynamic. For the metallic elements, they consist of electron streams. The effect occurs only at very low temperatures in the vicinity of the absolute zero (−273.16°C or 0K). In the case of helium, the maximum temperature at which the effect occurs is about 2.2K. For metals, the highest temperature is in the vicinity of 20K.

If one of the metals (commonly referred to as superconductors) is cast in the form of a ring and an external magnetic field is applied perpendicularly to its plane and then removed, a current will flow round the ring induced by Faraday induction. This current will produce a magnetic field, proportional to the current, and the size of the current may be observed by measuring this field. Were the ring (e.g., one made of lead) at a temperature above 7.2K, this current and field would decay to zero in a fraction of a second. But with the metal at a temperature below 7.2K before the external field is removed, this current shows no sign of decay even when observations extend over a period of a year. As a result of such measurements, it has been estimated that it would require 10^{99} years for the supercurrent to decay. Such persistent or "frictionless" currents in superconductors were observed in the early 1900s—hence they are not a recent discovery.

In the case of liquid helium, these currents are hydrodynamic, i.e., they consist of streams of neutral (uncharged) helium atoms flowing in rings. Since, unlike electrons, the helium atoms carry no charge, there is no resulting magnetic field. This makes such currents much more difficult to create and detect. Nevertheless, as a result of research carried out in England and the United States during the late 1950s and early 1960s, the existence of supercurrents in liquid helium has been established.

SUPERIMPOSED RIVER VALLEY. A river valley that is independent of present structural control may be described as either superimposed or antecedent. In the former case it is implied that the river has been able to maintain its course across resistant structures such as ridges, because it started as a consequent stream and has been "let down" on the underlying or nonconformable structure.


Diagrams illustrating the development of a superimposed river valley.

SUPERNOVA 1987A AND 1993J. What is there about an event that occurred 160,000 years ago that so fires the imagination and excites scientists into a veritable frenzy of observation? Supernovae are the results of violent explosions and are perhaps the most spectacular outbursts of energy we can ever see. A nova is an explosive variable star that brightens by a factor of thousands within hours or days and then fades to its previous brightness within months. The name "nova" is shortened from "stella nova," the Latin for "new star." It was once believed that these were really new stars, but we now know that they are actually old stars which suddenly explode and become much brighter. If the star was too faint to see before the explosive brightening, then it appeared as if it were a "new" star. Observations show that novae occur in close binary systems, one of whose components is a white dwarf star. White dwarf stars represent a final stage in the life history of a star, when it has consumed its nuclear fuel and has shrunk to the dimensions of the earth. In such a close binary system, mass transfer onto the white dwarf can occur when the companion star begins to expand as a result of its evolution. Detailed numerical calculations (in lieu of experiments!) show that when a critical amount of mass has been dumped onto the white dwarf star, explosive nuclear fusion occurs and the outer layer of accumulated matter is blown off.

A supernova is superficially similar to a nova in that it suddenly increases in brightness. A supernova, however, increases in brightness by factors of many millions. Hence the name *super*nova. Careful searches through historical records of visual observations from around the world suggest that supernovae were observed in AD 185, 393, 1006, 1054, 1181, 1572 and 1604.

The 1054 AD event left a spectacular remnant, widely known today as the Crab Nebula, and an associated pulsar. The events of 1572 AD and 1604 AD were observed by the famous astronomers Tycho Brahe and Johannes Kepler respectively and occurred before the invention of the telescope. No new supernova has been observed in our Milky Way since 1604 AD and thus our knowledge is almost entirely based on observations of supernovae in external galaxies. Because most external galaxies are so distant, supernovae in them appear quite faint to us, in spite of their intrinsic brightness, which at maximum light can rival that of billions of stars. Astronomers have long hoped for a bright, nearby supernova that would permit more detailed observations.

Additional information on novas and supernovas can be found in article on **Nova and Supernova.**

Supernova 1987A

On the night of 23/24 February 1987, Ian Shelton, working at the University of Toronto Las Campanas Station in northern Chile, discovered the brightest supernova seen since 1604. Not surprisingly, several other people, in Australia, Chile and New Zealand, saw it that same night, when it had a visual magnitude of about five. The official designation is Supernova 1987A (the letter A indicating that it is the first supernova discovered in 1987) in the Large Magellanic Cloud, because it appears to belong to this nearby companion galaxy to our own Milky Way. The Large and Small Magellanic Clouds are the nearest external galaxies and are named after the famous Portuguese navigator Ferdinand Magellen who travelled around the world in the early 1500s. The Magellanic Clouds are visible only from the earth's southern hemisphere and appear to the unaided eyes as illuminated clouds. "Nearby" is a relative concept, of course. The Large Magellanic Cloud is approximately 160,000 light years away, which means that what Shelton saw that night actually happened some 160,000 years ago.

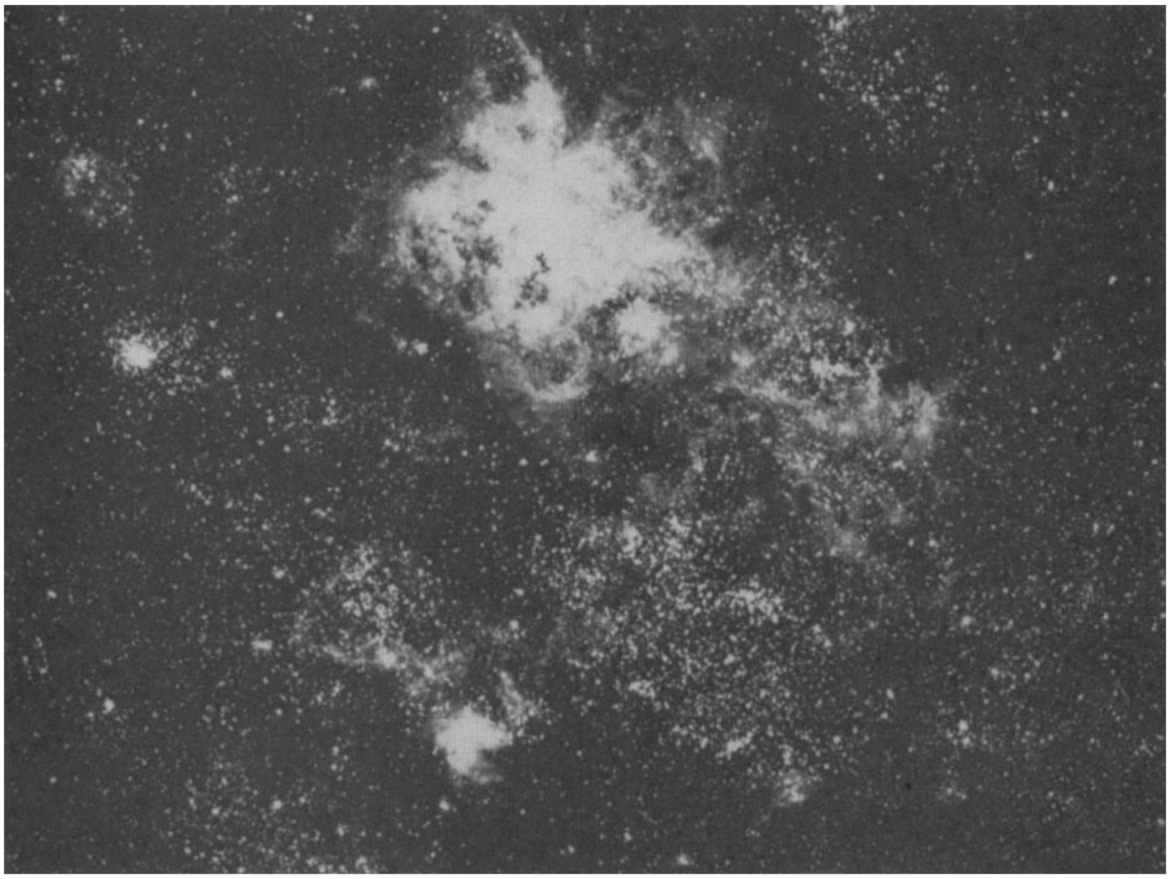
Within hours, astronomers all over the world were notified of Shelton's discovery. Astronomy is not an experimental science, and so astronomers must make the most of opportunities offered by unscheduled "experiments" which occur in the universe around us. For this purpose, the International Astronomical Union operates a Central Bureau for Astronomical Telegrams which sends out telegrams and circulars (which are distributed via postal and electronic mail services). All the suitably located telescopes were thus able to join quickly in gathering data about this serendipitous discovery, thus ensuring that Shelton's supernova already had become the best observed supernova ever. See accompanying illustration.

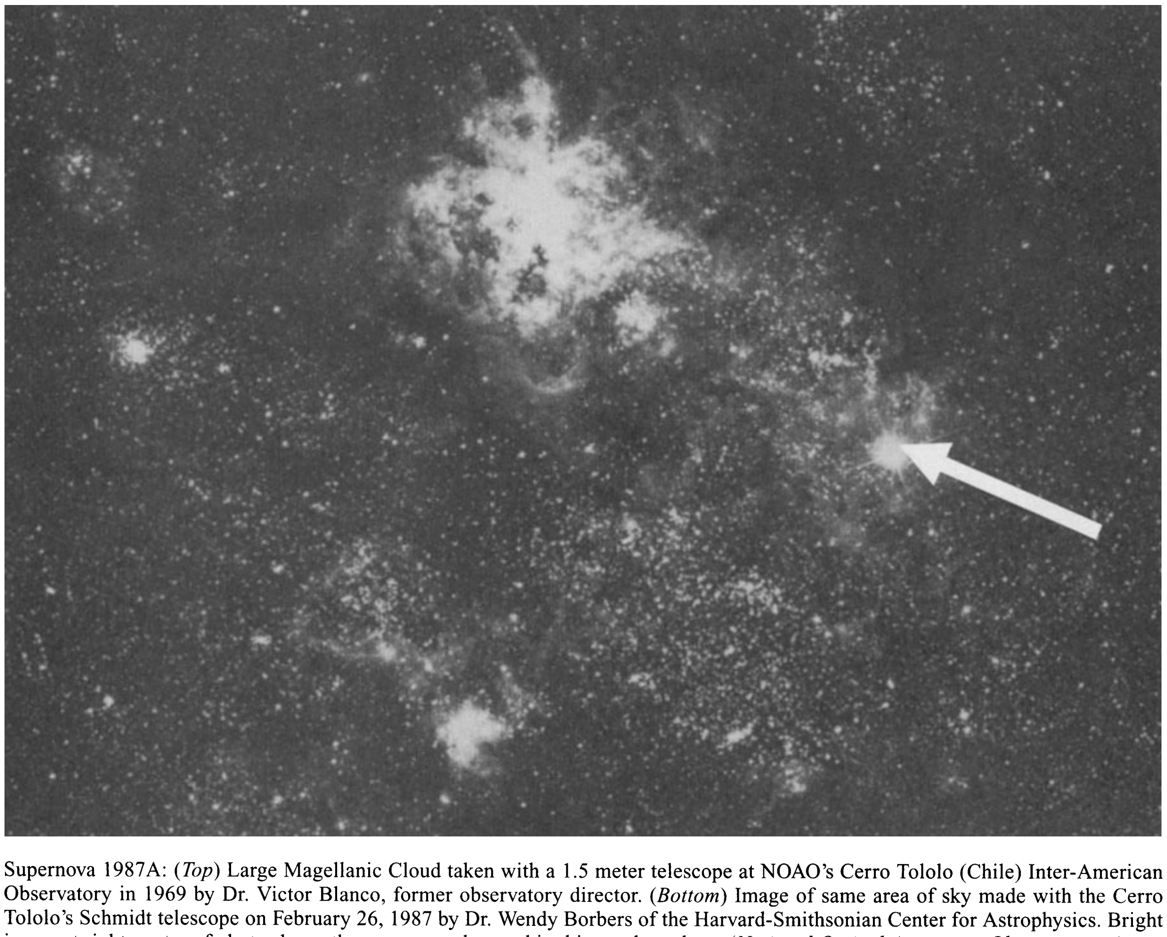
Peter Pesch, Chairman, Astronomy Department, Case Western Reserve University, Cleveland, Ohio.

EDITOR'S NOTES: At the end of 1987, Supernova 1987A continued to attract scientific attention. By December 1987, the object had faded at a steady 0.01 magnitude rate per day and was by then observable only instrumentally. It is regarded as the closest (160,000 light-years from Earth) and the most studied supernova in nearly 400 years (and, in fact, in history because of the unavailability of adequate instrumentation centuries ago). For quite some time after the event, observations, including the detection of neutrinos and the long-term falloff in luminosity aided in authenticating some theoretical predictions. During the investigation, a few intriguing problems surfaced, one of these being the blue color of the progenitor star, which was dimmer by a factor of 10 than what had been predicted. As one scientist observed, "We started out looking at the most dramatic event in the universe, and now we're arguing about what seemed like the most pedestrian thing—hydrogen burning and stellar evolution." Indeed, the event triggered a detailed reexamination of the former knowledge of stars.

Traditional theory provides the following approximate scenario for Sandu-leak—the star, losing sufficient thermonuclear fusion to prevent collapse, could no longer support its own weight and lost to "gravity." Hydrogen was converted to helium; the temperature and pressure in the core of the star progressively increased. The most tightly bound elements, oxygen, silicon, and ultimately iron, were involved, in a situation about to turn critical. The steadily growing core passed a threshold of about 1.5 solar masses; a multibillion-degree iron plasma underwent a phase transition. Thence, the iron nuclei commenced to "boil," disintegrating into helium nuclei, thus eliminating the supporting core. The core gave away and a supernova event occurred.

The gravitational potential energy of the collapsing core radiated neutrinos. The gravitational potential energy given up by the collapse is estimated at 3×10^{53} ergs (equivalent to about that required to convert about 10% of the mass of the sun into energy). Astronomers detected a neutrino pulse from the object at the large proton-decay detector near Cleveland, Ohio, as well as by a similar detector in Japan. (A





Supernova 1987A: (*Top*) Large Magellanic Cloud taken with a 1.5 meter telescope at NOAO's Cerro Tololo (Chile) Inter-American Observatory in 1969 by Dr. Victor Blanco, former observatory director. (*Bottom*) Image of same area of sky made with the Cerro Tololo's Schmidt telescope on February 26, 1987 by Dr. Wendy Borbers of the Harvard-Smithsonian Center for Astrophysics. Bright image at right-center of photo shows the supernova detected in this nearby galaxy. (*National Optical Astronomy Observatories.*)

prior burst of neutrino events observed at the Mount Blanc proton decay detector in Europe about 4 hours earlier is now believed to be spurious.)

Stanford Woosley (University of California at Santa Cruz) has observed that the emergence of cobalt-56 is yet another confirmation of standard theory. The resemblance to radioactive decay is no accident. Since June 1987, the supernova has been shining by the light of cobalt-56.

In February 1989, astronomers (University of California, Berkeley) suggested that they caught a glimpse of the ultra dense object at its heart, namely, a pulsar spinning furiously, almost 2000 times per second—so fast that it actually had broken apart. See Waldrop reference. However, also as reported by Waldrop, in late March 1989, the pulsar no longer was in view. As pointed out by one investigator, "Extensive computer analysis of the data showed that, on the night of 18 January (1989), the supernova was flickering ever so faintly at 1968.629 times per second, or more than twice as fast as any other pulsar ever seen. The rate gently rose and fell as though the pulsar were being tugged back and forth by the gravity of some kind of companion object." The foregoing frenzy of observations was disproved by a team at the Cerro Tololo Inter-American Observatory in Chile when it was announced that "the pulsar-like signals actually came from a television camera used to transmit images from the telescope to a monitor."

Supernova 1993J

An amateur astronomer, Francisco Garcia (Lugo, Spain), first reported this supernova on the night of March 28, 1993. On the following night, astronomers at the University of California, Berkeley, confirmed that Garcia had discovered a new supernova and, in fact, the brightest to shine in the Northern Hemisphere since 1937. Although supernova 1987A was far brighter, it could be seen only from the Southern Hemisphere. Professional astronomers consider 1993J to be in the same class as 1987A. The suspect object is believed to be a red supergiant, a larger and cooler star than the blue supergiant that exploded in 1987A. The object now is the target of numerous astronomical instruments.

Additional Reading

Horgan, J.: "Supernova 1987 Confirms and Contradicts Theories," *Sci. Amer.*, 26 (March 1988).
Horgan, J.: "A Remarkable 'Pulsar' Was Just a Flash in the Pan," *Sci. Amer.*, 30 (May 1990).
Horgan, J.: "Astronomers May Have Detected Supernova 1987A's Core," *Sci. Amer.*, 22D (April 1989).
Jayawardhana, R.: "A New Supernova in the Northern Sky," *Science*, 163 (April 9, 1993).
News: "Sighting of a Supernova," *Science*, **235**, 1143 (1987).
News: "The Supernova 1987A Shows a Mind of Its Own—and a Burst of Neutrinos," *Science*, **235**, 1322–1323 (1987).
News: "Supernova Neutrinos," *Sci. Amer.*, 18–19 (June 1987).
Waldrop, M. M.: "Supernova 1987A: Facts and Fancies," *Science*, 460 (January 29, 1988).
Waldrop, M. M.: "The Supernova 1987A Pulsar: Found?" *Science*, 892 (February 17, 1989).
Waldrop, M. M.: "Pulsar, Pulsar, Where Art Thou, Pulsar?" *Science*, 1553 (March 24, 1989).

SUPERPOSITION (Law of). The fundamental law in stratigraphy and historical geology stating that underlying strata are older than overlying strata unless the formations have been inverted by folding or by low-angle thrusts.

SUPERPOSITION (Nernst Principle of). The potential difference between junctions in similar pairs of solutions which have the same ratio of concentrations are the same even if the absolute concentrations are different, e.g., the same potential difference exists between normal solutions of HCl and KCl as exists between tenth-normal solutions of HCl and KCl.

SUPERPOSITION (Principle of). If a physical system is acted on by a number of independent influences, the resultant influence is the sum (vector or algebraic as circumstances dictate) of the individual influences. The principle takes on many specific forms depending on the nature of the system and the influence in question. For example, when two forces act simultaneously on a particle the resultant force is the vector sum of the two. Another example is provided by the small oscillations of a system about a state of equilibrium. Thus the total displacement of a vibrating string is the algebraic sum of all its various harmonic modes of oscillation which add without interfering with each other. The principle is validated in this case by the fact that the wave equation governing the oscillations is linear. Superposition does not apply to nonlinear systems.

The principle can also be applied to quantum mechanics. Here it is exemplified by the postulate that any state function of a given quantum mechanical system corresponding to a given observable (e.g., the energy) can be expressed as a linear expansion of the eigenstates of the system for the same observable.

SUPERSATURATED VAPOR. A vapor that remains dry, although its heat content is less than that of dry and saturated vapor at the pressure. Supersaturation is an unstable condition, and is found in the steam emerging from the nozzles of a steam turbine. The abnormality of the phenomenon is similar to that of supercooling. Supersaturation of the steam probably results from the very rapid expansion of steam in the nozzle, permitting the traverse of a short distance before the condensation of moisture is completed. At a certain definite point, however, known as the Williams limit, the supersaturation vanishes, and the steam regains the wet state which would be normal in view of the pressure and the heat content. Supersaturation of vapor is impossible in the presence of numerous charged ions or dust particles.

SUPERSATURATION (Chemical). The condition existing in a solution when it contains more solute than is needed to cause saturation. Thermodynamically, this type of supersaturation is closely allied to supersaturation of a vapor, since the solute cannot crystallize out in solutions free from impurities or seed crystals of the solute. See **Supersaturated Vapor.**

SUPERSONIC AERODYNAMICS. In the entry on **Aerodynamics,** the topic is generally covered from the standpoint of subsonic flows. World War II fighters were the first aircraft to attain speeds which produced, on critical points of the airplane, local flow velocities of Mach 1.0 and higher, although the airplane speed was less than the speed of sound. Only in the last couple of decades has aerodynamic experience passed beyond a limited theoretical knowledge. But since the early 1950s, a considerable bank of knowledge on supersonic flight has been collected.

For flight at low speeds, below about 300 miles per hour, air acts as an incompressible fluid. However, as velocity increases, air density changes about the airplane and this effect becomes increasingly important as speeds are increased. When flow velocities reach sonic speeds at some point on an airplane, the airplane's drag begins to increase at a rate much greater than that indicated by subsonic aerodynamic theory; subsonic flow principles are invalid at all speeds above this point.

Certain new definitions and concepts are necessary in dealing with air as a compressible fluid and with supersonic speeds:

Mach number is the ratio of the speed of motion to the speed of sound in air. The term comes from an Austrian physicist, Ernst Mach. An airplane traveling at the speed of sound is traveling at Mach 1.0.

A *shock wave* (compression wave) occurs in supersonic flow when the air must enter a volume smaller than the volume it has been occupying.

An *expansion wave* occurs in supersonic flow when the air is entering a larger volume than it has been occupying.

Supersonic Flow Characteristics. When an airplane flies at subsonic speeds, the air ahead is "warned" of the airplane's coming by a pressure change transmitted ahead of the airplane at the speed of sound. Because of this warning, the air begins to move aside before the airplane arrives and is prepared to let it pass easily. If the airplane travels at supersonic speeds, the air ahead receives no advance warning of the airplane's approach because the airplane is outspeeding its own pressure waves. Sound pressure changes are felt only within a cone-shaped region behind the nose of the airplane. Since the air is unprepared for the airplane's arrival, it must move to one side abruptly to let the airplane pass. This sudden displacement of the air is accomplished through a "shock wave." See Fig. 1.

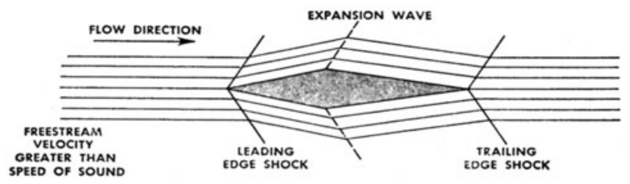


Fig. 1. Typical supersonic flow pattern.

The water-wave analogy furnishes a good physical picture of the subsonic "warning" system and supersonic "shock" formation. If one drops pebbles into a smooth pond of water, one each second, from the same point, each pebble will produce a water-wave moving outward with constantly increasing radius, as shown in Fig. 2. This is similar to the pattern of sound waves produced by an airplane sitting on the runway before take-off. Even though one cannot see the airplane, its presence is signaled by these outward rolling waves of engine noise.

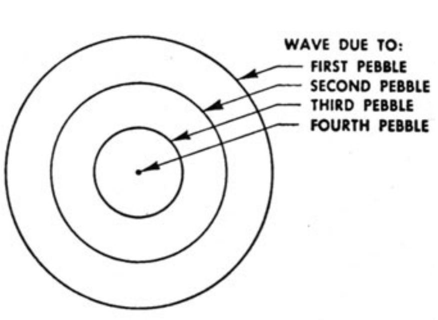


Fig. 2. Supersonic aerodynamics.

Now suppose we move slowly over the pond dropping pebbles at regular intervals. The picture of the waves is changed to that shown in Fig. 3. Each pebble still produces a circular wave, but the circles are crowded together on the side toward which we are moving; the center of each succeeding circle is displaced from the preceding one by a distance proportional to the speed at which we are traveling over the water. The wave pattern is identical to the pattern of sound waves around an airplane flying at subsonic speeds. The air ahead of the airplane is warned of the imminent arrival of the airplane and the warning time decreases with increasing airplane speed. The warning time is zero when the airplane is flying at exactly sonic speed. The corresponding water wave pattern is shown in Fig. 4. If we move across the water more rapidly than the water-wave speed, the wave pattern is markedly different from the previously illustrated patterns. The smaller circles are no longer completely inside the next larger ones. Now all the circles are included within a wedge-shaped region as shown in Fig. 5. This is similar to the sound wave pattern for an airplane flying at supersonic speed. The airplane is, in fact, a continuous disturbance in the air rather than an intermittent one, such as the pebbles falling regularly into the pond.

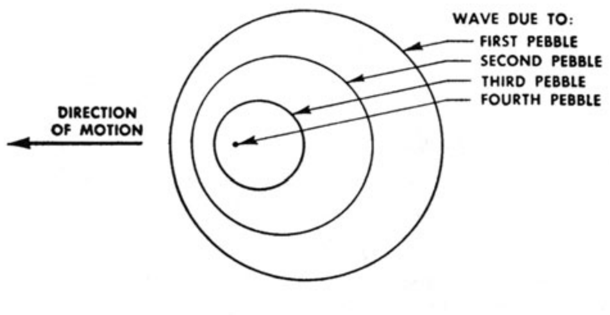


Fig. 3. Waves for motion at subsonic speed.

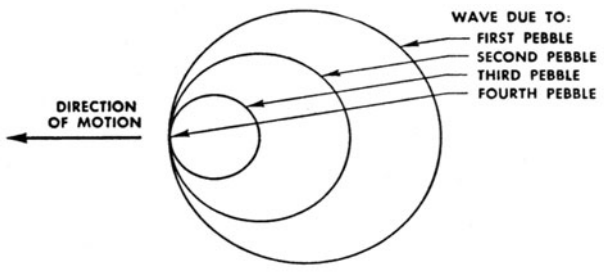


Fig. 4. Waves simulating Mach 1.0 motion.

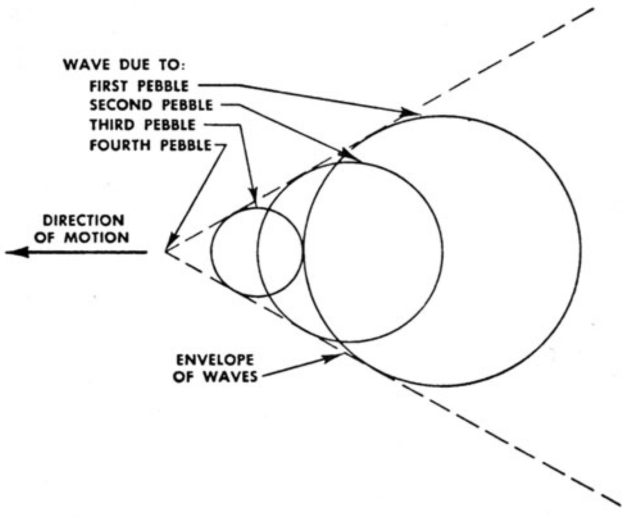


Fig. 5. Waves simulating supersonic motion.

Therefore, instead of several circles, there is an envelope surface of countless circles. The wedge on the surface of the pond looks like a section through the cone formed by an airplane in the air. Figure 5 indicates that there is considerable overrunning and interference between the wave circles; we might suspect from this that such interference will change the envelope shape for an actual airplane. This is, in fact, the case.

If the airplane is very streamlined and has a long, sharply pointed nose, then the air is not required to move a great distance suddenly in allowing the airplane to pass. In this case, the interference between sound waves is slight; the envelope is defined by the Mach angle, μ. Figure 6 shows that the Mach angle is the angle whose sine is the speed of sound divided by body velocity, or $1/V$. Thus, the Mach angle is 90° at a Mach number of 1.0; 30° at a Mach number of 2.0; and 10° at a Mach number of 5.75, for example. This envelope is called a Mach line in two dimensions or a Mach cone in three dimensions.

Types of Supersonic Waves. It is now apparent that waves are formed about any disturbance in a supersonic stream of air. The type of wave formed depends upon the nature of the disturbing influence. In our case, an airplane is the disturbing influence; its shape determines the location and characteristics of the waves formed. A wave of some type

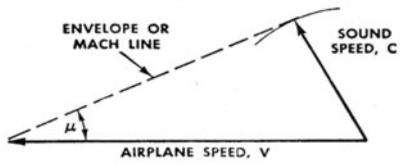


Fig. 6. Mach angle.

will exist whenever the air is required to change direction. The wave caused by a slight disturbance was defined as a Mach line. Air passing through a Mach line undergoes an infinitesimally small amount of temperature increase, pressure increase, and decrease in velocity. The Mach line envelope due to a long, slender, sharply-pointed body is conical in shape and is called a Mach cone. See Fig. 7. The envelope for a very thin wing is a wedge at center span bounded by a Mach cone at each tip, as shown in Fig. 8. The apex angle of these wedges and cones is the Mach angle, μ.

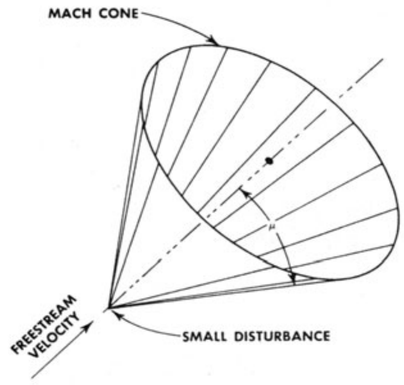


Fig. 7. Conical shock wave.

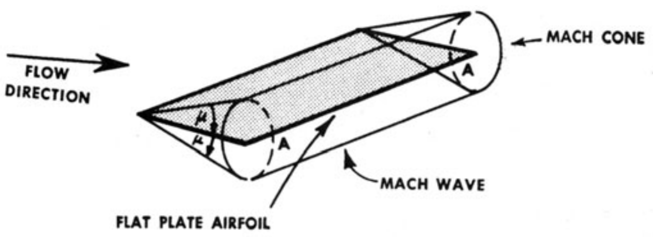


Fig. 8. Shock wave and Mach cones.

Mach waves are associated only with very small or gradual changes in the flow direction of the passing air. The bodies which are small enough to produce Mach waves are too slender to be incorporated on an actual airplane. To be practical, an airplane must be too blunt and thick for Mach waves to form; instead, a blunt body causes shock waves, called shocks.

These shocks are formed by the interference of sound waves mentioned earlier in the water wave discussion. Shocks are like Mach lines in that the pressure and temperature of the air passing through are abruptly increased and the air velocity is decreased. However, the magnitude of these changes through a shock is many times greater than the magnitude of these changes through a Mach line.

The difference in the magnitude of these changes is the essential difference between a Mach line and a shock wave. Since the drag of an object is dependent upon the pressure on its surface, the drag caused by a shock is very high compared to that caused by a Mach wave on the same body. Fundamentally, a Mach wave may be thought of as a shock of negligible strength, a shock through which air undergoes the smallest pressure, temperature, and velocity changes. The magnitude of the changes in these properties is used to measure the strength of a shock. The strength of a shock is dependent upon its angle with the freestream and the freestream Mach number. Strong shocks are associated with high drag. The strongest shocks are normal shocks, so-called because they stand at right angles to the freestream. All shocks standing at an angle of less than 90° to the freestream are called oblique shocks. Figure 9 shows examples of these two general cases.

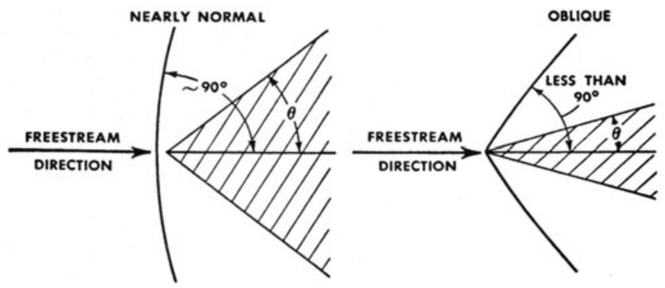


Fig. 9. Types of shock waves.

The Mach line and shock wave are compression waves. There is also the expansion wave or fan, which has characteristics opposite to the compression wave. In passing through an expansion wave, air velocity increases, while temperature and pressure are reduced. Expansion waves occur where bodies begin to narrow, making more space available for the passing air to occupy. Figure 10 illustrates a typical expansion wave. Since compression and expansion waves are opposite in nature, they tend to cancel each other when they intersect, and the shock's strength is reduced accordingly. Figure 11 shows a complete wave pattern on a double-wedge airfoil. The airfoil is at zero angle of attack. Shocks are formed at the leading and trailing edges while expansion fans occur at the surface angles or discontinuities. To obtain lift, the airfoil would have to have some finite angle of attack; the resultant wave pattern is shown in Fig. 12.

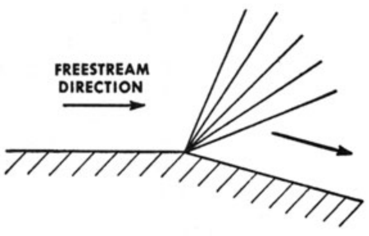


Fig. 10. Expansion wave.

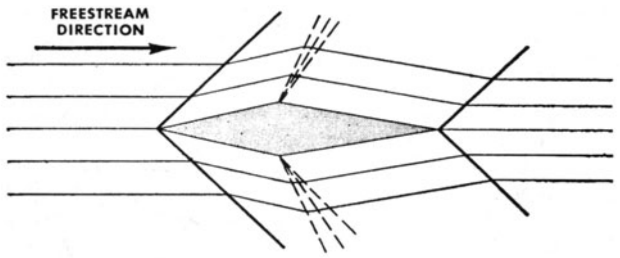


Fig. 11. Wave pattern on double-wedge airfoil where angle of attack is zero degrees.

The oblique shocks at the upper surface of the nose and lower surface of the rail are replaced by expansion fans in Fig. 12. This is a graphic illustration of the effect of slope change on wave formation. In one case, the air is pushed away to allow the airfoil to pass; in the other, the air has room to expand after passing over the airfoil nose, but undergoes high compression at the lower surface nose.

Airfoil Characteristics. The double-wedge airfoil used to illustrate wave patterns is convenient to study because the flow changes in direction only at four definite points. Another typical airfoil section might

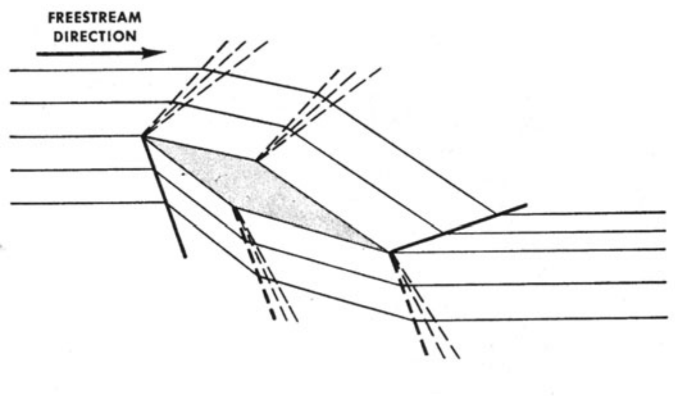


Fig. 12. Wave pattern on a double-wedge airfoil at angle of attack.

be as shown in Fig. 13. Here there are leading- and trailing-edge shocks, but the expansion is continuous over the entire surface between. The expansion waves intersect the leading-edge shock and progressively weaken it, thus making it a curved shock. The local pressure coefficient along the chord is plotted as a part of Fig. 13. The pressure coefficient is proportional to the local slope. The local slope is influenced by airfoil shape and angle of attack as well as by freestream Mach number. The upper and lower surfaces produce about equal amounts of lifting pressure. This represents a departure from the subsonic case where most of the lifting pressure comes from the upper surface.

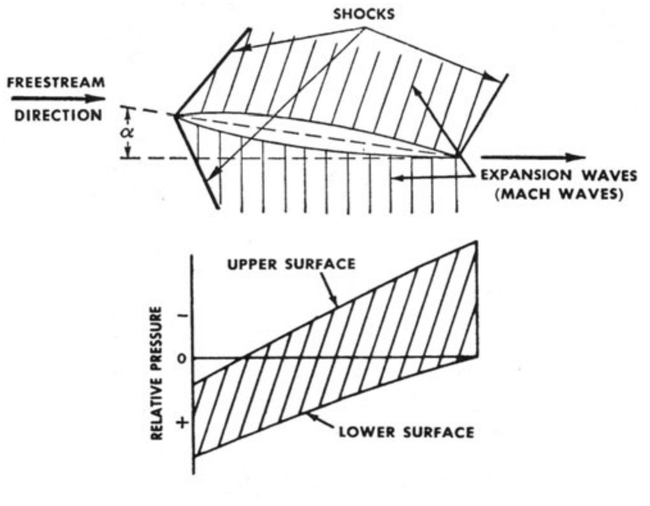


Fig. 13. Pressure distribution of an airfoil.

For supersonic airfoils, the center of pressure is very close to the half-chord point. This is a marked departure from the quarter-chord location of the center of pressure in subsonic flight.

The lift coefficient for supersonic airfoils of infinite span at moderate angle of attack and Mach number is

$$C_L = \frac{4\alpha}{\sqrt{M^2-1}}$$

where C_L = lift coefficient
α = wing angle of attack, radians
M = freestream Mach number

The expression for lift $(C_L(\rho V^2/2)S)$ is still applicable, and

$$L = \left(\frac{4\alpha}{\sqrt{M^2-1)}}\right)\left(\frac{\rho V^2}{2}S\right) = \frac{2\alpha\rho V^2 S}{\sqrt{M^2-1)}}$$

where V = velocity, feet per second.
S = wing area, square feet.
$\frac{\rho V^2}{2}$ = dynamic pressure, pounds per square foot
(ρ = air density in slugs per cubic foot).

The drag coefficient of airfoils in supersonic flight is composed of several equally important parts. These are pressure drag, friction drag, and drag due to lift. Pressure drag is the drag due to pressure distribution over the wing at the angle of zero lift and does not vary appreciably. It is also referred to as wave drag, and is similar to subsonic form drag. The approximately pressure drag coefficient is

$$C_D = \frac{5.33(t/c)}{\sqrt{M^2-1}}$$

where C_D = drag coefficient
t/c = thickness ratio

The pressure drag coefficient increases with increasing wing thickness and decreases with increasing Mach number. Consequently, supersonic airfoils are thin sections.

Friction drag results from the viscosity of the air. This tendency of the air to cling to a surface and to itself is felt with supersonic flow just as with subsonic.

Drag due to lift is the component of the resultant pressure force which acts in the drag direction. It too has an exact counterpart in subsonic flow. Drag due to lift is

$$C_D = \frac{4\alpha^2}{\sqrt{M^2-1}}$$

The term induced drag used for subsonic flow is not applied to supersonic drags. Induced drag in the subsonic case is due to tip losses. Tip losses occur in supersonic flow, but not in the same proportions as in subsonic flow.

Wing (Finite Span) Characteristics. It should be recalled that in subsonic flow, a wing of finite span experiences a three-dimensional flow which includes tip vortices, a downwash field, and induced velocities locally along the wing surface. See **Aerodynamics.** This is not true in supersonic flow. In Fig. 14, note that the pressure along the wing between the tip Mach cones is the same as for an airfoil of infinite length. Vortices produced within the tip Mach cones reduce the pressure from the airfoil value to zero at the tip, with the average lifting pressure in the tip region one-half the airfoil value. With a low aspect ratio, the absence of any airfoil circulation, as in subsonic flow, results in a marked decrease in wing average lifting pressure, and lift coefficient.

The supersonic drag due to lift depends upon the airfoil and angle of attack, while the subsonic induced drag is a function of lift coefficient and aspect ratio.

If a wing with a planform other than rectangular is used, tip losses can be eliminated. The delta, or triangular, wing planform accom-

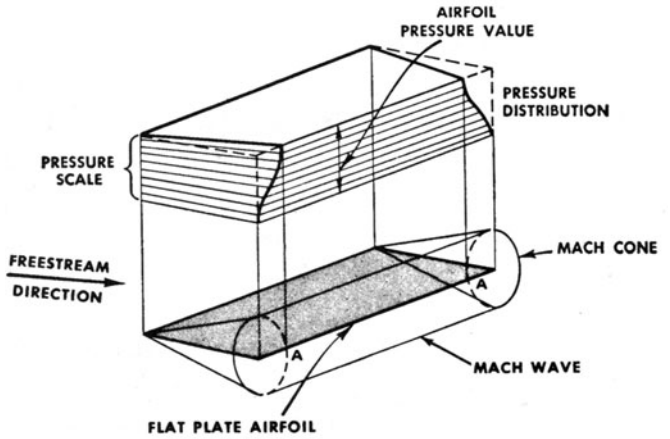


Fig. 14. Supersonic pressure distribution on a rectangular wing.

plishes this and can be illustrated by the two possible pressure patterns over a delta wing, depending on the relationship between freestream Mach number and wing leading-edge sweep.

In this case (Fig. 15), the components of velocity perpendicular to the leading edge are subsonic, even though the freestream flow is supersonic. The lifting pressure is maximum along the leading edge and decreases rapidly toward the center of the wing. The average lift coefficient is less than would be obtained by a similar airfoil in subsonic flow.

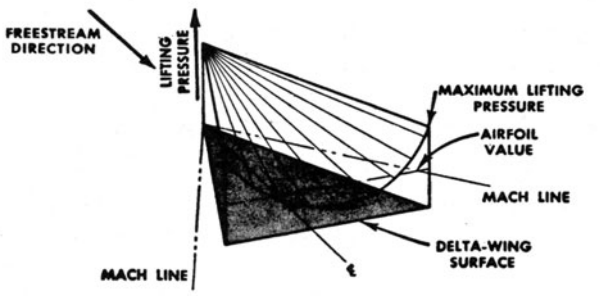


Fig. 15. Pressure distribution on a delta wing (leading edge inside the Mach line).

In the other pattern, the wing leading edge lies ahead of the tip Mach cone. Figure 16 illustrates that the highest lifting pressure still exists along the wing leading edge. In this case, however, the lifting pressure remains constant at this peak value in the region between the leading edge and the Mach cone. Inside the Mach cone, the lifting pressure again decreases, but the wing's average lift coefficient is as high as can be obtained with an airfoil of a similar cross section. When the leading edge is outside, the pressure drag coefficient is lower than airfoil pressure drag. Wing pressure drag reaches a maximum when the Mach cone lies along the leading edge.

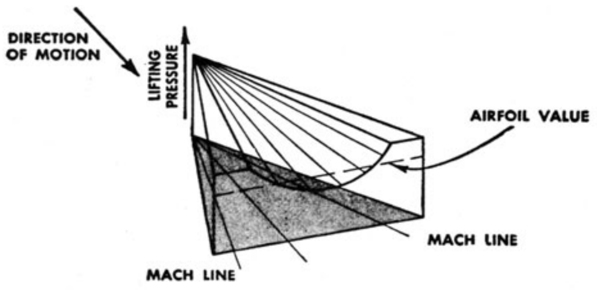


Fig. 16. Pressure distribution on a delta wing (leading edge outside the Mach line).

Consider the effects of modifying the delta planform. If area is added at the trailing edge to make a diamond planform, it is being added where the local lift coefficient is low. In this case, the average wing lift coefficient is less than that obtainable with a delta planform. Conversely, cutting out area to given an arrow planform will increase the average lift coefficient.

Body Characteristics. This discussion pertains only to bodies of revolution. A body of revolution is one whose cross section perpendicular to its longitudinal axis is always circular (Fig. 17). Airplane fuselages are generally as near circular in cross section as volume requirements will permit. Studies and tests indicate that a parabolic longitudinal shape is desirable. The pressure distribution on a body of revolution at zero angle of attack shows a positive value at the nose lower than for an airfoil of the same "seminose angle." See Fig. 17. Following this, the air finds more room in which to expand than in the case of the airfoil; it can fill the "ring" all around the body. The pressure drops so rapidly that the pressure coefficient returns to zero before the body slope has returned to zero, i.e., before the body contour becomes parallel to the longitudinal axis of the body.

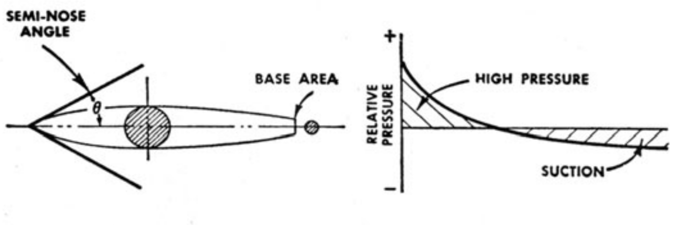


Fig. 17. Pressure distribution on a body of revolution.

The expansion continues over the aft part of the body and the pressure coefficient becomes more negative. However, the largest negative value of the pressure coefficient is limited by the occurrences of a complete expansion to a vacuum. Usually, the positive pressure coefficient at the nose is larger than this maximum negative value.

The pressure drag of bodies of revolution depends upon their shape, angle of attack, and flight Mach number. Angles of attack other than zero usually do not cause a large drag increase, nor will they cause the body to produce much lift. Drag force increases with increasing Mach number just as does wing lift force. Body shape has a strong influence on body drag. The longitudinal lines should be parabolic; the exact equation of the lines is a function of Mach number. A good fineness ratio for low drag is a length to maximum diameter ratio of 8 to 12. Many bodies are designed so that the pressure drag coefficient reaches its peak at very low supersonic Mach numbers, then drops rapidly until it begins to approach a minimum value at about Mach 2. The center of pressure of such bodies will be close to the half-length point if the tail is pointed. As base area increases, the center of pressure moves forward.

Wing-Body Combinations. In consideration of subsonic drag, it is apparent that a factor is interference drag, resulting from the effect of angles between surfaces, such as wing and fuselage. This drag is reduced by the use of smooth fairings, thus avoiding sharp corners—in effect it is corrected "locally" at the point it occurs.

In supersonic airplanes, the interference problem is much more critical and cannot be solved locally. It has been stated that the ideal streamline shape is a body of revolution having a longitudinal parabolic curve, that is, if the cross-sectional areas of the ideal body, taken at even increments along its axis, were plotted, the result would be a parabolic curve. Studies in the supersonic area indicate that in supersonic aircraft, the parabolic cross-sectional area distribution from nose to tail must be based on the complete airplane cross section, not just the fuselage cross section. The so-called "Coke-bottle" shape of the fuselage of some supersonic airplanes is the result of the application of this principle.

Supersonic Stability and Control. The analysis of the stability and control characteristics of an airplane capable of flight at supersonic speeds is essentially the same as that used for subsonic aircraft. A supersonic airplane must be statically and dynamically stable; it must have a control system that gives the pilot accurate, safe control of the airplane throughout its flight regime. However, the magnitude of control forces and moments and the changes in these forces and moments resulting from a displacement of the airplane about one or more of the stability axes are much greater at supersonic speeds than at subsonic speeds. Consequently, supersonic airplanes have power control systems to enable the pilot to move the control surfaces. The power control systems used on these aircraft can be designed so that the pilot has the same stick feel for a maneuver of given severity at all speeds despite the variation of control force with speed changes. In general, these airplanes also need 3-axis stability-augmentation systems to provide satisfactory dynamic stability.

Aerodynamic Heating. This results from the conversion of kinetic energy to thermal energy. At the surface of an aircraft, the air is slowed to zero velocity. This means that freestream kinetic energy has been converted to thermal energy. The resultant stagnation temperature is a function of freestream temperature and Mach number. Most of the aircraft surface does not reach this stagnation temperature because the conversion of kinetic energy to thermal energy is not 100% efficient. The maximum resultant skin temperature, known as adiabatic wall temperature, is 85 to 95% as high as the stagnation temperature. Radiation reduces skin temperature to less than adiabatic wall temperature; this effect increases as altitude and Mach number increase. Adiabatic wall temperature minus heat loss to radiation gives equilibrium temperature; this is the temperature to which an aircraft is designed. Typical stagnation temperatures are about 260°F (127°C) at Mach 2.0; and 1,550°F (843°C) at Mach 5. These temperatures serve to explain the structural and airconditioning problems involved in the design of supersonic airplanes.

Noise and Supersonic "Boom." The high-thrust jet engines used to power modern aircraft produce a sound level which tends to exceed human tolerance. Users of these aircraft equip their ground crews and mechanics with ear plugs to prevent physical and mental damage due to this noise. In addition to their effect on people, these sounds, or pressure waves, must be considered in the design of any part of the aircraft on which they impinge.

Supersonic flight is sufficiently commonplace to the public, but an associated phenomenon, known as the sonic boom or supersonic bang, has not met with the same public acceptance. A boom will occur whenever a shock wave (pressure wave) emanating from an aircraft flying supersonically reaches an observer. According to the water-wave analogy given earlier, a supersonic boom is similar to a water wave moving past a floating leaf and causing the leaf to bob up and down.

There are many shocks emanating from an aircraft flying supersonically, but these usually interact and coalesce to form two main shocks—one from the nose and one from the aft end of the aircraft. For this reason, the shape of an aircraft has little or no influence on the number of shocks reaching an observer. The bow and tail shock waves gradually diverge as they extend outward from the aircraft. This divergence is due to a slight difference in propagation velocities of the two waves. The observer will hear two bangs or booms if the time interval between the passing of the two waves is of the order of 0.10 second or greater. If this time interval is much less than this figure, as would be experienced during a low-level pass of a small aircraft, the ear could not discriminate between the two shock waves and would hear only one.

The loudness of the boom is a function of the distance between airplane flight path and observer, Mach number, aircraft size and shape, atmospheric pressure, temperature, and winds. The factor having the strongest influence on loudness is the distance between observer and airplane flight path. The loudness, as an increase in pressure above atmospheric, is inversely proportional to the three-quarter power of this distance.

Atmospheric temperature and wind gradients cause a bending or refraction of shock waves. Other atmospheric conditions, such as clouds, cause a diffusion of the waves. Hence in some cases the shockwave pattern from the airplane may become so distorted and attenuated that the bang is not even heard on the ground. Prediction of these effects under actual flight conditions is extremely difficult, though by precise control of airplane speed and flight path, shock waves may be focused at a point in space or on the ground. Such conditions can produce a supersonic boom on the order of ten times the normal intensity.

Several means of avoiding sonic boom from all causes have been tested. One, based on extremely accurate prediction of temperature and wind at the time of take-off, would adjust the angle of climb for subsonic speed. A two-degree error in temperature prediction, however, could negate a good plan and possibly the resultant noise would be worse than were no effort made to control it. Changing the climb angle, of course, also could adversely affect engine performance and fuel consumption. Another method advocates deceleration to subsonic speed before descent from cruise level, solving only one phase of the problem. Another plan embraces routing supersonic aircraft over unpopulated areas. Flying at greater altitudes is an alternative. However, the result of higher altitudes might decrease the intensity of the shock wave at the ground, but it would also spread the area affected. As of the early 1980s, the best approach appears to be a combination of very careful operational control at low altitudes and cruise altitudes sufficiently high to prevent or reduce the intensity effect of shock waves reaching the ground.

See also **Aerodynamics;** and **Airplane.**

SUPERVISORY CONTROL. In control system terminology, a control action in which the control loops operate independently, subject to intermittent corrective action, e.g., setpoint changes from an external source. Generally, a supervisory control system is assumed to incorporate a computer which provides broad, possibly relatively infrequent commands to a control system. The latter continues to perform its usual control functions during those periods between commands provided by the supervisory computer. In this sense, supervisory control is superimposed over the fundamental, second-to-second control system. A supervisory control computer thus replaces or assists the human operator in making periodic adjustments to a control system in an effort to better realize process objectives. The computer in a supervisory control system decides according to the information which it receives, assisted by certain programmed calculations or logic procedures, what adjustments in the control system should be made. Where the computer is completely *off-time*, the human operator will make the control adjustments manually based upon the computer inputs he receives. In other configurations, the computer may be interconnected with the process control system on a periodic or continuous basis. There are several shades of difference between a purely off-line and a purely on-line situation.

SUPPRESSOR (Electrical Noise). An element or device used in electric or electronic components or circuits to prevent or reduce undesired actions or currents. See also **Static (Communication).**

SUPPURATION. The formation and/or discharge of pus (leukocytes, serum, microorganisms, and necrotic debris)—as from a pustule (abscess, bubo, pimple, or other type of sore).

SURFACE. The locus of points satisfying an equation in three variables $f(x, y, z) = 0$. If u, v are parameters, the equation $x = \phi_1(u, v)$, $y = \phi_2(u, v)$, $z = \phi_3(u, v)$ are often useful. Figures which lie on the surface of a plane are studied in plane geometry. They include triangles, quadrilaterals, polygons, and circles. The differential properties of surfaces are a generalization of those which occur for plane and space curves but they become fairly complicated. They can be studied by means of differential calculus and differential geometry but tensor methods are especially suitable. A ruled surface is one that can be generated by the motion of a straight line called the generatrix. Any quadric surface can be produced in this way but the generatrix is real only for the cone, cylinder, hyperboloid of one sheet, and the hyperbolic paraboloid. In the other possible cases, the generatrix is imaginary. A surface of revolution results when a plane curve is revolved about a line lying in its plane. Special cases of each of the quadric surfaces may also be produced in this way.

Properly one should distinguish between a surface and a solid but often, for example, a sphere and a spherical surface are considered to be the same mathematical concept.

See also **Area.**

SURFACE (of Revolution). A surface which may be generated by rotating a plane curve about an axis in its plane. Sections of the surface perpendicular to the axis are circles called *parallel circles*, or *parallels*, or *lines of latitude*. Sections of the surface by planes containing the axis are called *meridan sections*, or *meridians*, or *lines of longitude*.

See also **Conic Section.**

SURFACE TENSION. Fluid surfaces exhibit certain features resembling the properties of a stretched elastic membrane; hence the term surface tension. Thus, one may lay a needle or a safety-razor blade upon the surface of water, and it will lie at rest in a shallow depression caused by its weight, much as if it were on a rubber air-cushion. A soap bubble, likewise, tends to contract, and actually creates a pressure inside, somewhat after the manner of a rubber balloon. The analogy is imperfect, however, since the tension in the rubber increases with the radius of the balloon, and the pressure inside, which would otherwise decrease, remains approximately constant; while the liquid "film tension" remains constant and the pressure in the bubble falls off as the bubble is blown.

Surface tension results from the tendency of a liquid surface to contract. It is given by the tension σ across a unit length of a line on the surface of the liquid. The surface tension of a liquid depends on the temperature; it diminishes as temperature increases and becomes 0 at the critical temperature. For water σ is 0.073 newtons/meter at 20°C, and for mercury, it is 0.47 newtons/meter at 18°C.

Surface tension is intimately connected with capillarity, that is, rise or depression of liquid inside a tube of small bore when the tube is dipped into the liquid. Another factor which is related to this phenomenon is the angle of contact. If a liquid is in contact with a solid and with air along a line, the angle θ between the solid-liquid interface and the liquid-air interface is called the angle of contact. See Fig. 1. If $\theta = 0$, the liquid is said to wet the tube thoroughly. If θ is less than 90°, the liquid rises in the capillary; and if more than 90°, the liquid does not wet the solid, but is depressed in the tube. For mercury on glass, the angle of contact is 140°, so that mercury is depressed when a glass capillary is dipped into mercury. The rise h of the liquid in the capillary is given by $h = 2\sigma \cos \theta / r\rho g$, where r is the radius of the tube, ρ the density of the liquid, and g is the acceleration due to gravity.

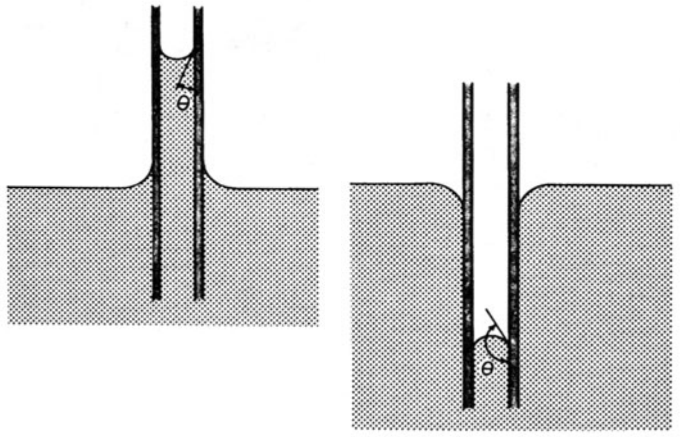


Fig. 1. Interrelationship between surface tension and capillarity: (*Left*) Case where angle theta is less than 90° (water); (*Right*) case where angle theta is greater than 90° (mercury).

Surface tension can be explained on the basis of molecular theory. If the surface area of liquid is expanded, some of the molecules inside the liquid rise to the surface. Because a molecule inside a mass of liquid is under the forces of the surrounding molecules, while a molecule on the surface is only partly surrounded by other molecules, work is necessary to bring molecules from the inside to the surface. This indicates that force must be applied along the surface in order to increase the area of the surface. This force appears as tension on the surface and when expressed as tension per unit length of a line lying on the surface, it is called the surface tension of the liquid.

The molecular theory of surface tension was dealt with by Laplace (1749–1827). But, as a result of the clarification of the nature of intermolecular forces by quantum mechanics and of the more recent developments in the study of molecular distribution in liquids, the nature and value of surface tension have been better understood from a molecular viewpoint. Surface tension is closely associated with a sudden, but continuous change in the density from the value for bulk liquid to the value for the gaseous state in traversing the surface. See Fig. 2. As a result of this inhomogeneity, the stress across a strip parallel to the boundary—ρ_N per unit area—is different from that across a strip perpendicular to the boundary—ρ_T per unit area. This is in contrast with the case of homogeneous fluid in which the stress across any elementary plane has the same value regardless of the direction of the plane.

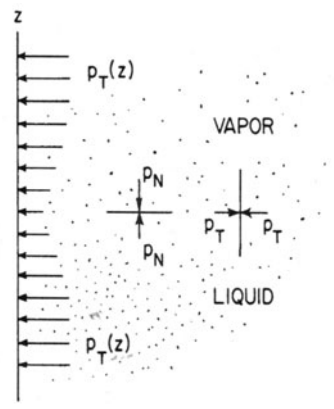


Fig. 2. Stress relationships in surface tension.

The stress ρ_T is a function of the coordinate z, the z-axis being taken normal to the surface and directed from liquid to vapor. The stress ρ_N is constant throughout the liquid and the vapor. Figure 2 shows the stress ρ_N and ρ_T. The stress $\rho_T(z)$ as a function of z is also shown on the left side of the figure.

SURFACE TENSION (Gibbs Formula). The total differential of the surface tension γ in the variables temperature T, and chemical potentials μ, is

$$d\gamma = -s^a\, dT - \sum_i \Gamma_i\, d\mu_i \qquad (1)$$

where s^a is the entropy per unit area of the surface phase and Γ_i the adsorption of component i.

This formula is the analog for a surface phase, of the Gibbs-Duhem equation for a bulk phase.

Another basic formula also due to Gibbs which relates the surface tension to the thermodynamic functions of the surface phase is

$$\gamma = A^a - \sum_i \Gamma_i \mu_i \qquad (2)$$

where A^a is the Helmholtz free energy (see **Free Energy (2)**) per unit area of the surface phase. This expression is the analog of the relation between the Gibbs free energy and the chemical potentials

$$G = \sum_i n_i \mu_i \qquad (3)$$

valid for a bulk phase.

SURGE. In electric terminology, a surge is an oscillation of relatively great magnitude set up by an electric discharge in a line or system. In meteorology, a surge is a general change in barometric pressure apparently superposed upon cyclonic and normal diurnal changes.

SURGEONFISHES (*Osteichthyes*). Of the suborder *Acanthuroidea*, family *Acanthuridae*, the surgeonfishes are well named because of the presence of a sharp, knifelike structure on the sides of the caudal peduncle. This is located just ahead of the tail. The fisherman unless careful can receive a serious cut from this structure. There are some 100 species of herbivorous surgeonfishes, the majority of which do not exceed 20 inches (51 centimeters) in length. They are considered food fishes, but the *Acanthurus triostegus* (Indo-Pacific convict fish) has been implicated in tropical fish poisoning instances. The general food supply of the surgeonfish is algae gathered from rocks and coral. Few of the most commonly seen species (*Ctenochaetus strigosus* and *C. striatus*) do well in aquariums.

SURGE TANK OR VESSEL. A liquid-holding chamber used to adsorb irregularities in flow. A surge tank may be used in a process where the total amount of fluid flowing around the closed cycle is constant, but where the volume passing one point in the cycle varies from that at another.

SUTURE. 1. The line of union of the adjacent flat bones making up the skull. 2. The surgical sewing-up of a wound or incision. Suture material is generally classified either as absorbable or nonabsorbable. The absorbable type is often made of either plain or chromic catgut. Nonabsorbable material is silk, linen, fine wire, or strands of synthetic composition such as nylon.

A fascial suture is one fashioned of a strip of fascia which is usually removed from the thigh where it covers the external muscles. Such sutures are often described as living sutures as they act similarly to a graft. They are used principally in the repair of large herniae where the tissues are weak.

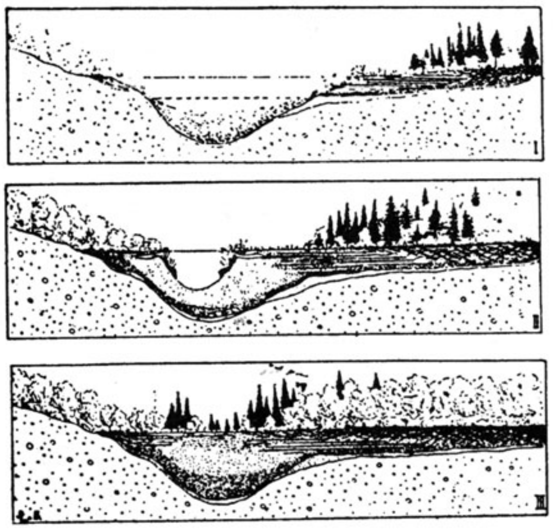
SWALLOW (*Aves, Passeriformes*). An insect-eating bird with a short wide beak, long and relatively narrow wings, weak feet and legs, and usually a forked tail. The distribution of the swallows is worldwide. They constitute the well-marked family *Hirundinidae*.

The swallows' nests are built in burrows, holes in trees, and about human dwellings, hence some of the species are familiar friends. In North America the purple martin, *Progne subis*, is widely known from the large colonies that nest in bird houses year after year. Among the true swallows of this continent the barn swallow, *Hirundo erythrogaster*, is among the most beautiful and is undoubtedly the most widely known. The bank, *Riparia riparia*, cliff, *Petrochelidon albifrons*, and rough-winged, *Stelgidopteryx ruficollis*, swallows are also widely distributed and locally common, though less beautiful than the tree swallow and the western violet-green swallow, *Tachycineta thalassina*. The glossy blue and green shades of the upper parts of the last two species are very striking, but both species are found in wild areas, and are thus less familiar than those mentioned above.

SWALLOW FLOAT (or Swallow Plinger). An electronic device that can be set to float in the seas at a predetermined depth from which it emits sounds that can be detected by appropriate devices at distances up to several miles.

SWALLOWTAIL (*Insecta, Lepidoptera*). A large butterfly, usually with slender tails extending from the hinder angles of the hind wings. The many species of swallowtails belong to the family *Papilionidae*. They are widely distributed in the tropical and temperate zones, especially in the former where some are very beautiful and brilliantly colored. Twenty-one species occur in North America. Most of them are yellow with black markings or vice versa but the common pawpaw swallowtail of the eastern and southern states is greenish-white with black bands and some red marks. Some of our species have metallic blue or green scales on the hind wings. Although some species lack the tails, they are also called swallowtails by association with the typical forms.

SWAMP. Where the flatness of the land, the presence of impervious soils or bedrock, or abnormal amounts of plant material obstruct or entirely prevent the normal drainage of an area, an excess of moisture will accumulate to the point of saturation and a swamp will come into existence. While most swamps are level this is not a necessary condition, for hillside swamps are by no means uncommon, due to a constant supply of percolating groundwater which maintains the swampy condition.



Origin and development of a peat bog, illustrating the successive stages in filling of a pond by the growth of the peat bog. (*After Dachnowski.*)

Lake basins are occasionally filled with vegetation and sediment, thus becoming swamps; these are frequently referred to as muskegs, a word of American Indian origin. Swamps may be formed on the flood plains of rivers as well as upon their deltas; they are characteristic of the flat ill-drained areas of the Atlantic Coastal Plain, examples of which are the Great Dismal Swamp which covers about 2,000 square miles (5,180 square kilometers) in the states of Virginia and North Carolina, and the Everglades of Florida, covering about 4,000 square miles (10,360 square kilometers).

Coastal saltwater swamps may develop in the zone between high and low tides or extend up river estuaries; examples of these are common along the Atlantic and Gulf coasts of the United States. In certain northern latitudes swamps develop into peat bogs. Peat bogs are an important source of fuel in Northern Europe, and also serve as an interesting illustration of the origin of coal, as exemplified in the peat, lignite, bituminous coal series. The accompanying figure illustrates the formation of a peat bog.

SWEEPBACK (Aircraft Wing). The acute angle between a line perpendicular to the plane of symmetry and the locus spanwise of the aerodynamic centers of the airfoils comprising the wing. For airplanes operating at subsonic speeds, sweepback may be employed to obtain the proper relationship between the center of gravity of the airplane and the aerodynamic center of the wing; for designs operating at or above sonic velocities, sweepback is employed to reduce the drag of the wing and the sweepback angle is equal to or greater than the Mach angle.

Sweepback has the same effect as incorporating a positive dihedral angle in the wing design for producing lateral stability. Since the sweepback for very high-speed aircraft is high (30 to 60 degrees) the excess lateral stability due to sweepback has to be counteracted by a negative dihedral, thereby producing a drooping or "tired" look to the wings. See also **Aerodynamics;** and **Airplane.**

SWEEP CIRCUIT. A circuit utilizing the transient voltage produced in a resistor-capacitor combination which furnishes the voltage or current for deflecting the electron beam in a cathode ray tube (see **Cathode-Ray Tube**). Figure 1a shows a simple but widely used sweep circuit. The capacitor is charged through the resistance *R* until the plate to cathode voltage of the thyratron reaches the breakdown value, at which time it begins to conduct and permits the capacitor to discharge very rapidly.

The voltage across the capacitor thus varies between the breakdown and extinction voltages of the thyratron. The transition from the lower

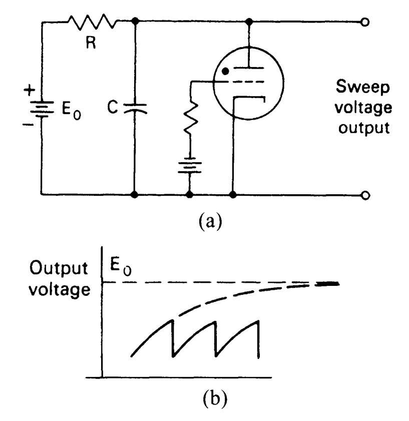


Fig. 1. (a) Sweep circuit using thyratron; (b) voltage waveform at output of a sweep circuit.

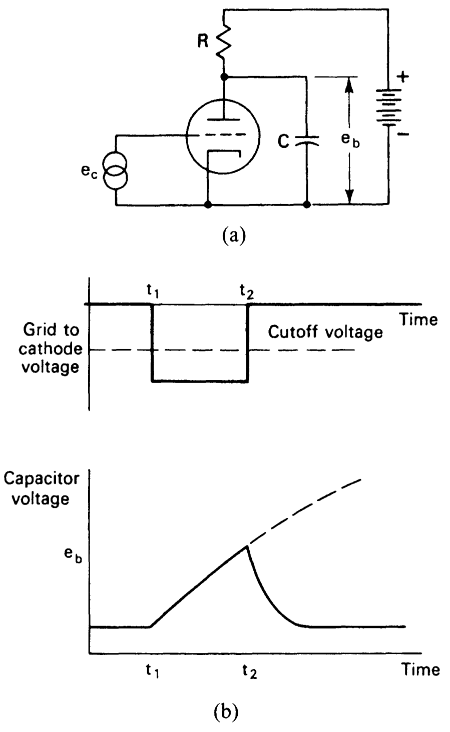


Fig. 3. (a) Typical sweep generator circuit; (b) input and output voltage waveforms for sweep circuits.

to the higher value is along an exponential path determined by R, C, and E_0. The process is then repeated and the resultant voltage output is as shown. Although this circuit is free-running, i.e., does not require an external signal to produce the sweep voltage, the output voltage may be synchronized with another signal by the introduction of a synchronizing signal at the grid. Virtually the same properties are obtained from the circuit of Fig. 2 in which a unijunction transistor is employed. Capacitor C is charged through resistor R until the emitter voltage of the unijunction transistor reaches the peak point value, which results in the emitter junction becoming forward biased, with a corresponding sharp reduction in the dynamic resistance between the emitter and the base connected to the negative side of the battery. Capacitor C then discharges through this low impedance emitter-to-base path until the emitter voltage becomes so low that the emitter junction fails to conduct. At this time, charging of capacitor C resumes and the cycle is repeated. The voltage waveforms are the same as shown in Fig. 1a. The second transistor is an emitter follower used to minimize the effect of the circuit to which the sweep voltage is supplied on the operation of the unijunction transistor circuit.

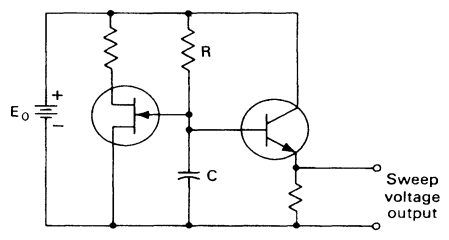


Fig. 2. Sweep circuit using unijunction transistor.

A circuit that provides a single sweep output for each input pulse is shown in Fig. 3a; the associated voltage waveforms are given in Fig. 3b. Prior to time t_1, the triode conducts at zero bias, and the voltage across capacitor C is only a small fraction of the battery voltage because of the voltage drop in the large resistor R. At time t_1, the grid to cathode voltage is made negative enough to cut off the flow of plate current. This condition is maintained until time t_2. During the interval of time that the tube is cut off, the capacitor C is charged from the battery through the resistor R. The voltage across the capacitor increases along the exponential curve associated with the transient in a resistor-capacitor circuit. Ultimately, e_b would become equal to the battery voltage. The exponential rise is not permitted to continue after time t_2, however, for when the tube is returned to zero bias at that time, the capacitor discharges through the tube and its voltage rapidly assumes the value which it had at the start of the charging process. Many refinements, including feedback, may be added to the circuit to increase the degree of linearity of the voltage change between t_1 and t_2.

SWEETENERS. Drawings in Egyptian tombs depicting beekeeping practices and honey production attest that the demand for sweet-tasting substances dates back to 2600 B.C. Sugar consumption varies considerably from one country to the next as shown by the accompanying table.

In terms of sugar consumption in the United States, until the early 1940s, sucrose from sugarcane and sugar beets accounted for a very high volume of the fundamental sweeteners. Since that time there has been continuously increasing consumption of corn sweetners and other caloric sweeteners, notably high-fructose corn syrup (HFCS), and, of course, a marked impact on sucrose consumption occurred with the introduction of artificial sweeteners, particularly of saccharin and aspartame. Sweeteners fall into two general categories—*nutritive* and *nonnutritive*.

Nutritive Sweeteners

In addition to their sweetening power, nutritive sweeteners are effective preservatives in numerous foods. Sweeteners tie up water, essential

SUGAR CONSUMPTION PER CAPITA PER YEAR

Country	Refined Sugar	
	Pounds	Kilograms
Israel	150	68.0
Bulgaria	130	59.0
Australia	119	54.0
New Zealand	110	49.9
Costa Rica	108	49.0
Cuba	107	48.5
Switzerland	106	48.1
United States	102	46.3
Hungary	99	44.9
Iceland	98	44.9
Poland	95	43.1
Sweden	94	43.1
Austria	92	41.7
Czechoslovakia	92	41.7
European Economic Community	89	40.4
Norway	87	39.5

SOURCE: International Sugar Organization.

for microorganism growth, thus preventing or inhibiting spoilage. Nutritive sweeteners also serve as food for yeasts and other fermenting agents, so important in many processes, including baking. The principal functional properties of sucrose are (1) browning reactions, (2) fermentability, (3) flavor enhancement, (4) freezing-point depression, (5) nutritive solids source, (6) osmotic pressure, (7) sweetness, (8) texture tenderizer, and (9) viscosity/bodying agent.

Among the principal *natural sugars* are fructose, glucose (also called dextrose), honey, invert sugar, lactose, maltose, raffinose and stachyose, sucrose, sugar alcohols, and xylitol.

Dextrose Equivalent. A means for comparing one sugar with another. The total amount of reducing sugars, expressed as dextrose (glucose), that is present in a given sugar syrup is calculated as a percentage of the total dry substance. More technically, the dextrose equivalent (DE) is the number of reducing ends of sugar that will react with copper. The DE can be measured in several ways.

Fructose. Also called *levulose* or *fruit sugar*, $C_6H_{12}O_6$. It is the sweetest of the common sugars, being from 1.1 to 2.0 times as sweet as sucrose. Fructose is generally found in fruits and honey. An apple is 4% sucrose, 6% fructose, and 1% glucose (by weight). A grape (*Vitis labrusca*) is about 2% sucrose, 8% fructose, 7% glucose, and 2% maltose (by weight) (Shallenberger, 1974). See also **Apple;** and **Grape.** Commercially processed fructose is available as white crystals, soluble in water, alcohol, and ether, with a melting point between 103 and 105°C (217.4 and 221°F) (decomposition). Fructose can be derived by the hydrolysis of inulin; by the hydrolysis of beet sugar followed by lime separation; and from cornstarch by enzymic or microbial action.

Dry crystalline fructose is reported to have a sweetness level of 180 on a scale in which sucrose is represented at 100 (Andres, 1977). In cool, weak solutions and at lower pH, sweetness value is reported to be 140–150. At neutral pH or higher temperatures, the sweetness level drops, and at 50°C (122°F) sweetness equals that of a corresponding sucrose solution. A synergistic sweetness effect is reported between sucrose and fructose. A 40–60% fructose/sucrose mixture in a 100% water solution is sweeter than either component under comparable conditions (Unpublished report, University of Helsinki, 1972).

Glucose. Also known as *grape sugar* or *dextrose*, this is the main compound into which other sugars and carbohydrates are converted in the human body and thus is the major sugar found in blood. Glucose is naturally present in many fruits and is the basic "repeating" unit of the starches found in many vegetables, such as potato. Purified glucose takes the form of colorless crystals or white granular powder, odorless, with a sweet taste. Soluble in water, slightly soluble in alcohol. Melting point is 146°C (294.8°F). Glucose finds many uses—confectionery, infant foods, brewing and winemaking, caramel coloring, baking, and canning. Glucose is derived from the hydrolysis of corn starch with acids or enzymes. Glucose is a component of invert sugar and glucose syrup. Glucose was first obtained (1974) from cellulose by enzyme hydrolysis.

Corn (maize) syrup is a sweetener derived from corn starch by a process that was first commercialized in the 1920s. Corn syrup is composed of glucose and a variety of sugars described as the "maltose series of oligosaccharides." These syrups are not as sweet as sucrose, but are very often used in conjunction with sugar in confections and other food products.

Five types of corn sweeteners are commercially available: (1) *Corn syrup* (glucose syrup), with a DE of 20 or more, is a purified and concentrated aqueous solution of mono-, di-, and oligosaccharides. High fructose corn syrup (HFCS) is prepared by enzymatically converting glucose to fructose with glucose isomerase. (2) *Maltodextrin*, concentrated solutions or dried powders of disaccharides, characterized with a DE of less than 20. The manufacturing process is similar to that of corn syrup except that the conversion process is stopped at an earlier stage. (3) *Dried corn syrup* is a granular, crystalline, or powder product, from which a portion of the water has been removed. (4) *Dextrose monohydrate* is a purified and crystallized form of D-glucose, and contains one molecule of water of crystallization per molecule of D-glucose. (5) *Dextrose anhydrous* is primarily D-glucose with no water of crystallization.

Galactose. A monosaccharide commonly occurring in milk sugar or lactose. Formula, $C_6H_{12}O_6$.

Honey. A natural syrup which varies in composition and flavor, depending upon the plant source from which the nectar was collected by the honeybee, the amount of processing, and the duration of storage. The principal sugars contained in honey are fructose and glucose, the same components as in table sugar. There are minute amounts of vitamins and minerals in honey, but these are not usually considered in terms of calculating minimum requirements. See also the entry on **Honey.**

Invert Sugar. A mixture of 50% glucose and 50% fructose obtained by the hydrolysis of sucrose. Invert sugar absorbs water readily, and is usually only handled as a syrup. Because of its fructose content, invert sugar is levorotatory in solution, and sweeter than sucrose. Invert sugar is often incorporated in products where loss of water must be minimized. Commercially, invert sugar is obtained from the inversion of a 96% cane sugar solution. This sugar is used in various foods, in the brewing industry, confectionery field, and in tobacco curing.

Lactose. *Milk sugar* or *saccharum lactis*, $C_{12}H_{22}O_{11} \cdot H_2O$. Purified lactose is a white, hard, crystalline mass or white powder with a sweet taste, odorless. It is stable in air, soluble in water, insoluble in ether and chloroform, very slightly soluble in alcohol. The compound decomposes at 203.5°C (398.3°F). Lactose is derived from whey, by concentration and crystallization. Cow's milk contains about 5% lactose. Because of its relative lack of sweetening power, lactose is not considered a sweetener in the usual sense. It is used as a bulking agent in numerous food products. Lactose can be used effectively as a carrier for artificial sweeteners to give a free-flowing powder that is easily handled. There has been interest in the hydrolysis of lactose into glucose and galactose, both enzymatically and chemically. It has been reported that glucose and galactose are known to be sweeter than lactose itself. The relative sweetness of sugars is not a constant relationship, but depends upon many factors, including pH, temperature, and presence of other constituents. Mixtures of sugars can make a different sweetness impression than that of individual sugars alone. Synergistic sweetness often results from a combination of sugars.

Maltose. Also known as *malt sugar*, maltose is a product of the fermentation of starches by enzymes or yeast. Barley malt, which is used as an adjunct in brewing, enhances the flavor and color of beer because of its maltose content. Maltose also is formed by yeast during breadmaking. Maltose is the most common reducing disaccharide, $C_{12}H_{22}O_{11} \cdot H_2O$, composed of two molecules of glucose. It is found in starch and glycogen. Purified maltose takes the form of colorless crystals, melting point, 102–103°C (215.6–217.4°F). Soluble in alcohol; insoluble in ether. Combustible. Maltose is used as a nutrient, sweetener, and culture medium.

Raffinose and Stachyose. These are sugars found in significant amounts in some foods, such as beans. These sugars are not digested in the stomach and upper intestine as are other disaccharides. They are fermented by bacteria in the lower digestive tract, producing gases and sometimes causing discomfort from flatulence. Raffinose is a trisaccharide composed of one molecule each of D(+)-galactose, D(+)-glucose, and D(−)-fructose, $C_{18}H_{32}O_{16} \cdot 5H_2O$. Raffinose is sometimes used in the preparation of other saccharides.

Sucrose. Table sugar, also known as *saccharose*. Sucrose is a disaccharide, composed of two simple sugars, glucose and fructose, chemically bound together, $C_{12}H_{22}O_{11}$. Hard, white, dry crystals, lumps, or powder. Sweet taste, odorless. Soluble in water; slightly soluble in alcohol. Solutions are neutral to litmus. Decomposes in range of 160–186°C (320–366.8°F). Combustible. Optical rotation = +33.6°. Derived from sugarcane or sugar beets and also obtainable from sorghum. Sucrose is the most abundant free sugar in the plant kingdom and has been used since antiquity (Mead and Chem, 1977).

Raw sugar, centrifugal sugar, refined sugar, and molasses are described in the entry on **Sugarcane.** *Turbinado sugar* is raw sugar which has been refined to remove impurities and most of the molasses. It is edible when produced under sanitary conditions and has a molasses flavor. *Brown sugar* consists of sucrose crystals covered with a film of molasses syrup that give the characteristic color and flavor. The sucrose content varies from 91 to 96%. *Confectioner's* or *powdered sugar* is another form of sucrose made by grinding the sugar crystals. It is usually mixed with about 3% starch to prevent clumping. It is used for household baking, canning, and table use, or industrially where rapid solution in cold liquids is desirable.

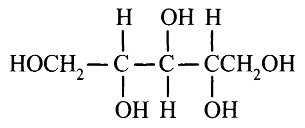
Sugar Alcohols. These are polyols, chemically reduced carbohydrates. Important in this group are *sorbitol, mannitol, maltitol*, and *xylitol*. Xylitol is described later.

Polyols are frequently used sugar substitutes and are particularly suited to situations where their different sensory and functional properties are attractive. In addition to sweetness, some of the polyols have other useful properties. For example, although it contains the same number of calories/gram as other sweeteners, sorbitol is absorbed more slowly from the digestive tract than is sucrose. It is, therefore, useful in making foods intended for special diets. When consumed in large quantities (1–2 oz; 25059 g)/day, sorbitol can have a laxative effect, apparently because of its comparatively slow intestinal absorption.

When sugar alcohols are ingested, the body converts them first to fructose, which does not require insulin to facilitate its entry into the cells. For this reason, ingesting these sweeteners (including fructose itself) does not cause the immediate increase in blood sugar level which occurs upon eating glucose or sucrose. Within the body, however, the fructose is rapidly converted to other compounds, which *do* require insulin in their metabolism. One effect of this stepwise metabolism is to "damp out" the peaks in blood sugar levels which occur immediately after ingesting sucrose, but which are absent after ingesting fructose, even if the eventual insulin requirements are the same. Thus, individuals with metabolic problems should not make the assumption that fruit sugars are perfectly all right to consume, but first should consult their physicians. In fact, some health scientists are dubious about pursuing the apparent claims for substituting fructose and sugar alcohols for sucrose as a major sweetener, particularly for diabetics, until more research is done on their long-range nutritional and biophysiological consequences. Research interest has also focused on these sweeteners because of their relatively low potential for causing dental caries. Studies have shown about a 30% reduction in dental caries in laboratory animals on sorbitol and mannitol diets, and virtually complete elimination of caries in the animals when on xylitol diets.

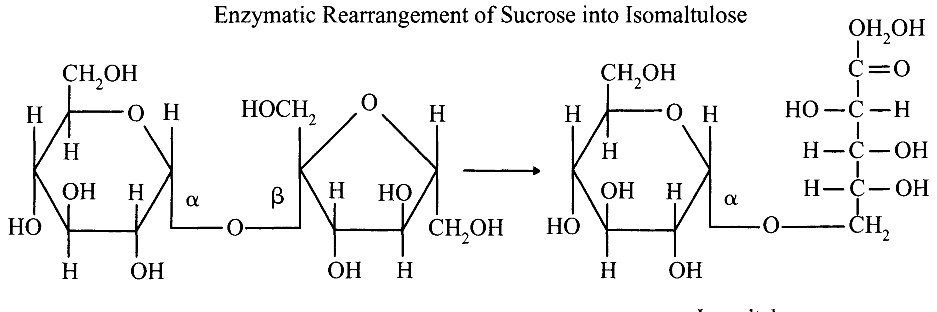
Xylitol. This is a 5-carbon sugar alcohol that occurs widely in nature—raspberries, strawberries, yellow plums, cauliflower, spinach, and many other plants. Although widely distributed in nature, it is present in low concentrations and this makes it uneconomic to extract the substance directly from plants. Thus, commercial xylitol must be produced from xylan or xylose-rich precursors through the use of chemical, enzymatic, and other bioprocessing conversions. A frequently used source has been birch tree chips. Other appropriate starting materials include beech and other hardwood chips, almond and pecan shells, cottonseed hulls, straw, cornstalks (maize), and corn cobs. The base source in the aforementioned agricultural waste materials is hemicellulose xylan. The hemicellulose is acid hydrolyzed to yield xylose which, followed by hydrogenation and chromatographic separation, yields xylitol.

Xylitol is equally as sweet as sucrose. This property is of advantage to food processors because in reformulating a product from sucrose to xylitol, approximately the same amounts of xylitol can be used. Because xylitol has a negative heat of solution, the substance cools the saliva, producing a perceived sensation of coolness, quite desirable in some food products, notably beverages. Recently, this property has been used in an iced-tea-flavored candy distributed in the European market. As of the late 1980s, 28 countries have ruled positively in terms of xylitol for use in commercial products. Xylitol has been found particularly attractive for use in chewing gum, mint and hard candies, and as a coating for pharmaceutical products. Xylitol has the structural formula shown below, with a molecular weight of 152.1. It is a crystalline, white, sweet, odorless powder, soluble in water and slightly soluble in ethanol and methanol. It has no optical activity.



Isomalt. Developed in Germany, isomalt is described as an energy-reduced bulk sweetener and marketed in Europe under the tradename *Palatinit*™. The compound is produced from sucrose in a two-step process, as shown below.

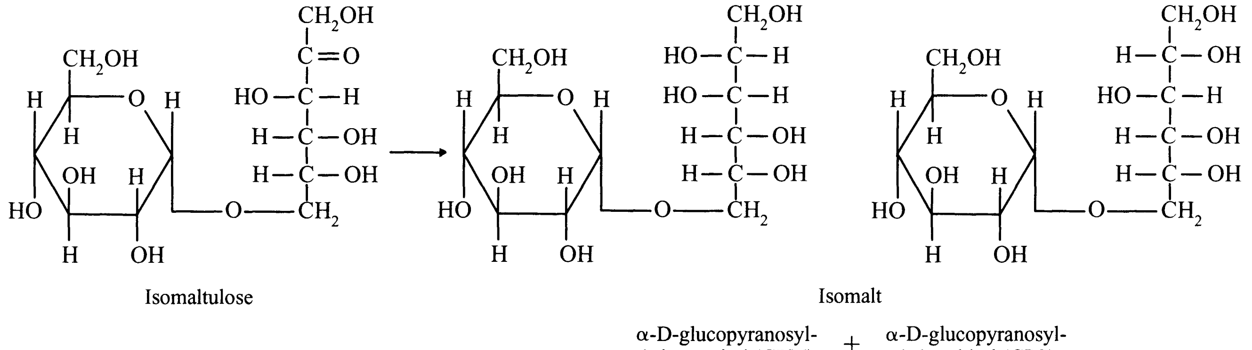
Enzymatic Rearrangement of Sucrose into Isomaltulose



Sucrose

Isomaltulose

Hydrogenation of Isomaltulose to Produce Isomalt



Isomaltulose

Isomalt

α-D-glucopyranosyl-1,6-mannitol (GPM) + α-D-glucopyranosyl-1,6-sorbitol (GPS)

In the first step, the easily hydrolyzable 1-2 glucoside linkage between the glucose and fructose moieties of sucrose are catalyzed by immobilized enzymes to produce isomaltulose, *Palatinos.*™ After crystallization, the isomaltulose is hydrogenated in a neutral aqueous solution using a nickel catalyst.

It is claimed that isomalt is odorless, white, crystalline, and sweet tasting without the accompanying taste or aftertaste. Sweetening power is from 0.45 to 0.6 that of sucrose. A synergistic effect is achieved when isomalt is combined with other artificial sweeteners and sugar substitutes. Principal applications are in confections, pan-coated goods, and chewing gum. The substance was approved for use in most European countries in 1985. Classification of isomalt as a GRAS substance was petitioned in the United States. (GRAS = generally regarded as safe.)

Aspartame. This synthetic sweetener is included with the nutritive sweeteners because it does have some caloric value (when metabolized as a protein, it releases 4 kcal/g). The relationship between sweetness of aspartame and sucrose is almost linear when plotted on a log-log scale. Aspartame is 182 times sweeter than a 2% sucrose solution, but only 43 times sweeter than a 30% solution. The clean, full sweetness of aspartame is similar to that of sucrose and complements other flavors.

The full name of aspartame is *aspartylphenylalanine*, a dipeptide that degrades to a simple amino acid. It has been reported as easily metabolized by humans. Aspartame was accidentally discovered in 1965 with the synthesis of a product for ulcer therapy. Aspartame is metabolized by the same biochemical pathway as proteins, yielding phenylalanine, aspartic acid, and methanol. Because of the byproduct phenylalanine, which some individuals are unable to metabolize, appropriate labeling is required. This is a concern for individuals with phenylketonuria (PKU). Aspartame was first approved in the United States in 1974, then banned in 1975. In July of 1981, it was approved for use in various foods, dry beverage mixes, and in tabletop sweeteners. Approval for use in carbonated beverages was granted in July 1983.

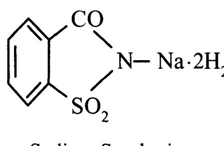
Currently, aspartame is used in tabletop sweeteners (*Equal* in the U.S.; *Egal* in Quebec, Canada; and *Canderal* in Europe and the U.K.). Aspartame currently is incorporated as the exclusive sweetening ingredient in nearly all diet soft drinks in the United States. In other countries, it may be blended with saccharin at a level close to 50% of the saccharin level. Soft-drink manufacturers have taken some measures to enhance stability by raising pH slightly and by more closely controlling the inventory for carbonated soft drinks. Notable differences in sweetness are perceived at a 40% loss in aspartame level.

Crystalline Maltitol. Classified as a bulk sweetener with taste and mouthfeel similar to sucrose, crystalline maltitol contains maltitol as the major component (88+%), with small amounts of sorbitol, maltotriitrol, and hydrogenated oligosaccharides. Its use is in tabletop sweeteners, chocolate, candy, and baked goods. Maltitol has been a major component of hydrogenated glucose syrup in the United States since 1977 and has been used in Japan since 1963. The product was introduced in Europe in 1984. Classification of crystalline maltitol as a GRAS substance was petitioned in the United States in 1986.

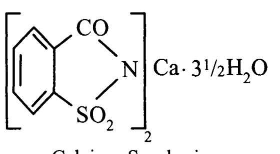
Nonnutritive Sweeteners

There are several currently used and a number of potential noncaloric sweeteners, including saccharin, cyclamate (banned in the U.S., but permitted in approximately 40 other countries), acesulfame K, monellin (from the serendipity berry), stevioside, glycyrrhizin, hernandulcin, neosugar, miraculin (from miracle fruit), and a sweetener-enhancer, thaumatin, are being investigated.

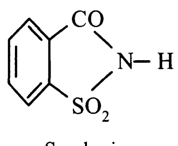
Saccharin. A noncaloric sweetener that is about 300 times as sweet as sugar. The compound is manufactured on a large scale in several countries. It is made as saccharin, sodium saccharin, and calcium saccharin, as shown by formulas below.



Sodium Saccharin



Calcium Saccharin



Saccharin

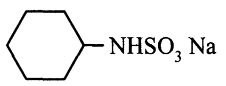
Saccharin (ortho-benzosulfimide) was discovered in 1879 by I. Remsen and C. Fahlberg when they were researching the oxidation products of toluene sulfone amide. The most common forms of saccharin are sodium and calcium saccharin, although ammonium and other salts have been prepared and used to a very limited extent. The saccharins are white, crystalline powders, with melting points between 226 and 230°C (438.8 and 446°F). Soluble in amyl acetate, ethyl acetate, benzene, and alcohol; slightly soluble in water, chloroform, and ether. Saccharin is derived from a mixture of toluenesulfonic acids. They are converted into the sodium salts, then distilled with phosphorus trichloride and chlorine to obtain the ortho-toluene sulfonyl chloride, which by means of ammonia is converted into ortho-toluenesulfamide. This is oxidized with permanganate, then treated with acid, and saccharin is crystallized out. In food formulations, saccharin is used mainly in the form of its sodium and calcium salts. Sodium bicarbonate may be added to provide improved water solubility.

Saccharin is used in conjunction with aspartame in carbonated beverages. Other uses include tabletop sweeteners, dry beverage blends, canned fruits, gelatin desserts, cooked and instant puddings, salad dressings, jams, jellies, preserves, and baked goods.

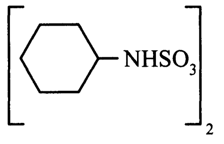
For many years, saccharin has been under investigation by a number of countries. As of the late 1980s, some questions remain unresolved.

Cyclamate. Group name for synthetic, nonnutritive sweeteners derived from cyclohexylamine or cyclamic acid. The series includes sodium, potassium, and calcium cyclamates. Cyclamates occur as white crystals, or as white crystalline powders. They are odorless and in dilute solution are about 30 times as sweet as sucrose. The purity of commercially available compounds is approximately 98%.

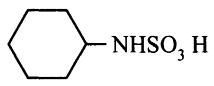
Discovered in 1937 and patented in 1940, cyclamate is a derivative of cyclohexylamine, specifically, cyclohexane sulfonic acid. The sodium salt form is normally used, but the calcium salt may be subsituted in low-sodium diets. See structural formulas below.



Sodium Cyclamate



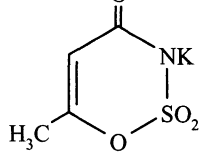
Calcium Cyclamate



Cyclamic Acid

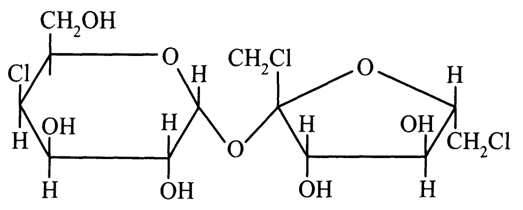
Once widely used, cyclamate was prohibited in the United States in 1970. Although used in many other countries, reapproval in the United States has not yet been established. An independent review of the possible carcinogenicity of cyclamate was conducted in April 1985 by the National Academy of Sciences/National Research Council at the request of the Food and Drug Administration. The review concluded that cyclamate itself is not a carcinogen, although it may serve as a promotor or cocarcinogen in the presence of other substances.

Acesulfame-K. This substance (potassium salt of the cyclic sulfanomide), 6-methyl-1,2,3-oxathiazine-4(3H)-1,2,2-dioxide, shown below, was developed by Karl Clauss (Hoechst Celanese Corporation, Somerville, New Jersey) in 1967. The compound is a white, odorless, crystalline substance with a sweetening power 200 times that of sucrose. A synergistic effect is produced when the substance is combined with a number of other sweeteners. The substance is calorie-free and not metabolized in the human body. Approval of the use of Acesulfame-K was given by the Food and Drug Administration (FDA) in the United States in 1983 and it is found in scores of popular retail products, including yogurt, rice pudding, and soft drinks.



Acesulfame-K

Sucralose. Developed in England during the mid-1980s, testing and evaluation commenced in 1988. The structural formula of the compound (a chlorinated disaccharide derived from sucrose) is shown below.



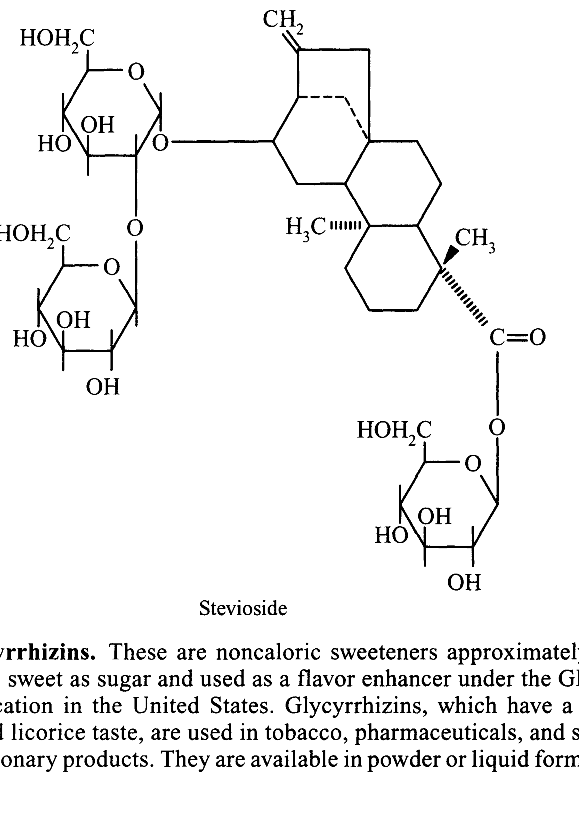
Sucralose

Sucralose is absorbed poorly in humans and other mammalian species. The small portion that is absorbed is not broken down and is quickly excreted. It has been reported that an extensive array of studies has demonstrated that sucralose is nontoxic—not carcinogenic, teratogenic, mutagenic, or caloric.

Monellin. The sweetness of this compound is claimed to be 1500 to 3000 times that of sucrose, but a different flavor profile prevails. The detection of a sweet taste is slow, commencing after a few seconds in contact with the taste buds, then gradually increasing to its peak intensity. The sweet taste can persist for up to an hour. The source is the relatively rare serendipity berry, the fruit of a noncultivated West African vine. Extraction of the sweet component is effected by treating the berry with a series of enzymes (pectinase and bromelain), followed by dialysis and chromatographic separation. The compound resulting contains the protein *monoellin* with a molecular weight of about 10,700 and composed of two nonidentical polypeptide chains of 50 and 43 amino acids. Neither of the individual chains imparts sweetness. Regulatory measures have not been instituted because of the compound's apparent instability and limited raw resources for processing. However, to date, tests with mice have shown no evidence of toxicity.

Stevioside. Derived from the roots of the herb *Stevia rebaudiana*, this compound has found limited use in Japan and a few other countries as a low-calorie sweetener having about 300 times the sweetening power of sucrose. The compound has not been investigated thoroughly by a number of countries with strong regulatory agencies and, therefore, is not on the immediate horizon for wide consideration as a sweetener.

The dried leaves of *S. rebaudiana* have been used in Paraguay for many years to sweeten bitter drinks. From 3 to 8% of the dried leaves is stevioside, which is a diterpene glycoside as shown by the formula below.



Stevioside

Glycyrrhizins. These are noncaloric sweeteners approximately 50 times as sweet as sugar and used as a flavor enhancer under the GRAS classification in the United States. Glycyrrhizins, which have a pronounced licorice taste, are used in tobacco, pharmaceuticals, and some confectionary products. They are available in powder or liquid form and with color, or as odorless, colorless products. These compounds are stable at high temperatures (132°C; 270°F) for a short time and thus can be used in bakery products. In some chocolate-based products, the sweetener has been used to replace up to 20% of the cocoa. The sweetener also has excellent foaming and emulsifying action in aqueous solutions. Typical products in which these sweeteners may have application include cake mixes, ice creams, candies, cookies, desserts, beverages, meat products, sauces, and seasonings, as well as some fruit and vegetable products. Generally available as malted and ammoniated glycyrrhizin.

The basic compound is a triterpene glycoside. It is extracted from the licorice root, of which the principal sources are China, Russia, Spain, Italy, France, Iran, Iraq, and Turkey. The roots, containing 10% moisture, are dried and shredded, after which they are extracted with aqueous ammonia, concentrated in vacuum evaporators, precipitated with sulfuric acid, and crystallized with 95% ethyl alcohol.

Hernandulcin. Tasting panels have estimated that this substance is 1000 times sweeter than sucrose, but the flavor profile is described as somewhat less pleasant than that of sucrose. Hernandulcin is derived from a plant, *Lippia dulcis* Trev, commonly known as "sweet herb" by the Aztecs as early as the 1570s. It has been categorized as noncarcinogenic, based upon standard bacterial mutagenicity tests. The economic potential is being studied.

Neosugar. This is another substance in early stages of development and testing. The compound is composed of sucrose attached in a beta(2-1) linkage to 2, 3, or 4 fructose units.

Miraculin. Rather than a sweet-tasting substance, miraculin is described as a taste-modifying substance that elicits a sweet taste to tart foods. The product has been reported as used by African cultures for over a century. The compound is derived from a shrub (*Synsepalum dulcificum*) which grows in West Africa. Miraculin is a glycoprotein with a molecular weight ranging from 42,000 to 44,000. Approval of the Food and Drug Administration has thus far been denied, awaiting further tests. A GRAS category was denied in 1974.

Thaumatin. This is a protein extracted and purified from *Thaumatococcus danielli*, a plant that is found in West Africa. The leaves of the plant have been used for many years in Africa for wrapping food during cooking. Claims have been made that thaumatin is from 2000 to 2500 times sweeter than 8–10% solution of sucrose. The final product is odorless, cream-colored and imparts a lingering licorice-like aftertaste. The substance synergizes well with monosodium glutamate (MSG) and is used in typical Japanese seasonings as well as in chewing gum, pet foods, and certain pharmaceuticals (to mask unpleasant flavor notes). Use in Japan has been approved since 1979. It is considered a GRAS substance in the United States for use in chewing gum. In this application, thaumatin extends the flavor and boosts the perceived duration of flavor. The compound is normally applied as a dust to the surface of gum. Some authorities believe that the use of thaumatin in pet foods has high potential.

Sweeteners in Formulating and Processing

In using low-calorie sweeteners in various food products, the problems are not limited to flavor, but often much more importantly involve texture, acidity, storage stability, and preservability, among others. Acceptable nonnutritively sweetened products cannot be developed by the simple substitution of artificial sweeteners for sugars. Rather, the new product must be completely reformulated from the beginning. Three examples follow.

Jams, Jellies, and Preserves. Traditional products in this category contain 65% or more soluble solids. In low-calorie analogs, soluble solids range from 15% to 20%. Under these circumstances, commonly used pectins (high methoxyl content) do not suffice. Thus, special LM (low methoxyl) pectins must be used, along with additional gelling agents, such as locust bean gum, guar gum, and other gums and mucilagenous substances, some of which may require some masking. In the absence of sugar, a preservative, such as ascorbic acid, sorbic acid, sorbate salts, propionate salts, and benzoates, usually is required to the extent of about 0.1% (weight).

Soft Drinks. In addition to providing sweetness, sugar also functions to provide mouthfeel and to stabilize the carbon dioxide of soft drinks. To contribute to mouthfeel, the use of hydrocolloids and sorbitol has been attempted with limited success. Hydrocolloids also help to some degree with the problem of carbonation retention, but the principal solutions to this problem involve avoiding all factors which contribute to carbonation loss. Thus, the requirement for very well filtered water to eliminate particulates as possible nucleation points; any substances that promote foaming must be avoided; any emulsifying agents used in connection with flavoring agents must be handled carefully to avoid foaming; carbonation should be carried out at low temperature (34°F; 1.1°C); and trace quantities of metals must be absent from the water.

Bakery Products. These foods are among the most difficult as regards the use of artificial sweeteners. A listing of the functions of sugar in baked goods beyond that of providing sweetness is indicative of these problems. Sugar contributes to texture in forming structures, in providing moist and tender crumbs by counteracting the toughening characteristics of flour, milk, and egg solids. In the emulsification process required to retain gas during leavening, sugar is an effective accessory agent. Ingredients frequently used in bakery products to compensate for the absence of sugar include carboxymethylcellulose, mannitol, sorbitol, and dextrins, but, generally, these have not been very satisfactory—either to processor or consumer. This remains a large area of challenge for the food processors and ingredient manufacturers.

Evaluating Synthetic Sweeteners. Evaluation of new sweeteners, unlike that of most functional food ingredients, is not possible using totally objective means. There are no general rules leading to structure/function relationships for all classes of sweeteners. The principal judgments must rely on human sensory panel tests. Matters of this type are described from a general standpoint in the entry on **Sensory Evaluation.** The training and administration of sensory panels for sweeteners are beyond the scope of this volume.

Additional Reading

Barndt, R. L., and G. Jackson: "Stability of Sucralose in Baked Goods," *Food Technology*, 62 (January 1990).

Bartoshuk, L. M.: "Sweetness: History, Preference, and Genetic Variability," *Food Technology*, 108 (November 1991).

Birch, G. G.: "Chemical and Biochemical Mechanisms of Sweetness," *Food Technology*, 121 (November 1991).

Farber, S. A.: "The Price of Sweetness," *Technology Review (MIT)*, 46 (January 1990).

Igoe, R. S.: "Dictionary of Food Ingredients," 2nd Edition, Van Nostrand Reinhold, New York, 1989.

Keller, W. E., et al.: "Formulation of Aspartame-Sweetened Frozen Dairy Dessert without Bulking Agents," *Food Technology*, 102 (February 1991).

Lindley, M. G.: "From Basic Research on Sweetness to the Development of Sweeteners," *Food Technology*, 134 (November 1991).

Noble, A. C., Matysiak, N. L., and S. Bonnans: "Factors Affecting the Time-Intensity Parameters of Sweetness," *Food Technology*, 128 (November 1991).

O'Mahony, M.: "Techniques and Problems in Measuring Sweet Taste," *Food Technology*, 128 (November 1991).

Pepper, T., and P. M. Olinger: "Xylitol in Sugar-Free Confections," *Food Technology*, 98 (October 1988).

Staff: "Applications of Aspartame in Baking," *Food Technology*, 56 (January 1988).

Staff: "Evaluation of Advanced Sweeteners," *Food Technology*, 60 (January 1988).

Staff: "FDA Clears Hoechst's Non-Caloric Sweetener for Use in Dry Foods," *Food Technology*, 108 (October 1988).

Wnnia, S. M.: "Modeling the Sweet Taste of Mixtures," *Food Technology*, 140 (November 1991).

Wong, D.: "Mechanism and Theory in Food Chemistry," Van Nostrand Reinhold, New York, 1989.

SWEETGUM TREE. A relatively large American tree (*Liquidambar styraciflua*). The tree achieves its largest stature in the south Atlantic states. The mature tree rises straight up on a tall trunk that is usually free of branches for up to 70 feet (21.3 meters) above ground. In the immature tree, the head is narrow and pyramidal. As the tree matures, the head becomes irregular when not pruned. The leaves are star-shaped (5 lobes), glossy, dark, and rather thick. They turn into many brilliant colors in autumn. The flowers are quite inconspicuous. Brown and twisted seed capsules tend to hang by long stalks through most of winter. The wood has a tendency to warp and is difficult to season; otherwise it would rival black walnut for working. Sweetgum trees are known in Asia (*L. orientale*).

The record sweetgum tree growing in the United States is located in North Carolina. As compiled by the American Forestry Association, this specimen has a circumference (at 4.5 feet; 1.4 meters above ground level) of 178 inches (452 cm), a height of 136 feet (41.5 meters), and a spread of 66 feet (20.1 meters).

SWEET POTATO (*Ipomoea Batatas; Convolvulaceae*). The plant is a trailing perennial, the stems of which twine in counter-clockwise direction around supporting objects. These stems arise from much-thickened roots which are rich in starch. In cultivation many varieties have been developed, with many different leaf shapes. Dark green, heart-shaped leaves with shining surface occur in several varieties, while in others the leaves are variously lobed and dissected. The flowers, seldom produced in plants grown in northern latitudes, are about 2 inches (5 centimeters) across, purple, and borne either singly or in small axillary cymes. The fruit is a capsule. See Fig. 1.

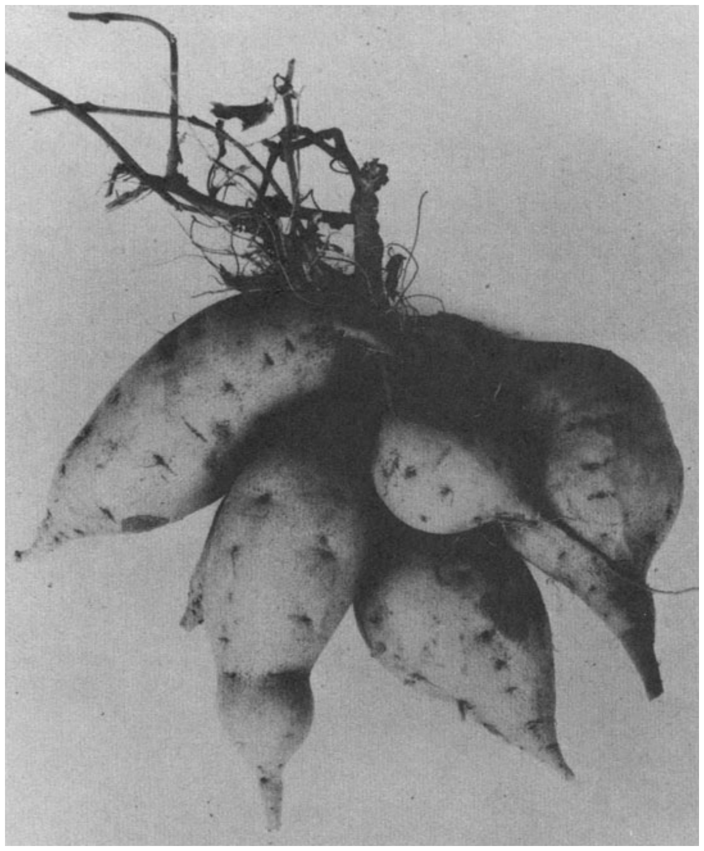


Fig. 1. A hill of *Big-Stem Jersey* sweet potatoes. (*USDA photo.*)

Various methods of propagation are employed. Small roots may be planted whole, adventitious buds soon forming and giving rise to shoots which appear above the ground in about four weeks. Root cuttings from growing plants may also be used, especially in regions where the growing season is long. On occasion stem cuttings may be used, but necessarily demand a long growing season and reach maturity very late in the season. Seed may be grown, but germination is slow and uneven, and the product not uniform.

The sweet potato is an American plant, a native of the West Indies and Central America. In cultivation it has gradually spread out of tropical lands, new varieties being developed to suit new localities. At present the crop is grown as far north as Cape Cod.

The principal use of the sweet potato is for human consumption, although some are fed to swine. The vines are frequently used as stock food. From the roots, starch, flour, glucose, and alcohol are extracted, to a limited extent.

Sweet potatoes are frequently called yams. This application of the name yam to the sweet potato is confusing, since the true yam is an entirely different plant, *Dioscorea alata* (*Dioscoreaceae*), widely grown in tropical lands for its edible tubers, which are rich in sugar, watery, and soft when cooked. The flowers are white. Propagation is mainly by cuttings of tubers, each containing one or more eyes, or small buds, such as are found in the white potato tuber. Yams are widely used as food.

Another species of *Ipomoea*, *I. purpurea*, is the Morning Glory, frequently cultivated for its showy flowers.

SWIFT (Lizard). (*Reptilia, Squamata*). A lizard. The name is applied without scientific accuracy to some of the iguanas, including small species of two different genera. Most of these lizards occur in the southwestern United States and Mexico but one species, the pine or fence lizard, *Sceloporus undulatus*, is found as far north as Michigan, New Jersey, and Oregon.

SWIFTS AND HUMMINGBIRDS (*Aves, Apodiformes*). The swift is a small bird with a short wide beak and long slender wings. The swift is superficially like the swallows, but is more closely related to the hummingbirds. The numerous species of swifts are widely distributed in both hemispheres, five species occurring in North America. The common chimney swift, *Chaetura pelagica*, which has abandoned its original habit of nesting in hollow trees to occupy chimneys, is both widely distributed and abundant. See Fig. 1. The remaining species are found only in the far west and southwest. The swift makes its nest of various materials cemented together and fastened to their support with saliva. One species of the Oriental region uses the secretion alone, without foreign materials. The nests of this species are attached to the walls of caves and are the famous edible bird nests of Chinese epicures.

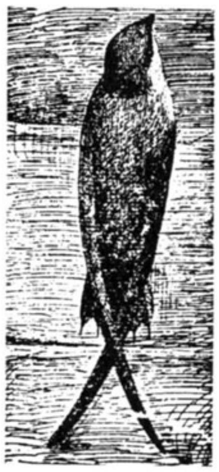


Fig. 1. Chimney swift. Sooty brown; tail feathers sharply pointed; long narrow wings.

The hummingbird is a small bird of the New World whose wings are moved so rapidly in flight that they produce a low-pitched sound. The hummingbird is capable of hovering in one spot in the air, and habitually poises before flowers which are visited for nectar and insects. The wings may beat up to 90 times per second. There are about 320 species of hummingbirds, many of these occurring in the northern Andes of South America. However, hummingbirds range from Alaska southward into much of the South American continent. They are found up to the snow line, 16,000 feet (4,800 meters), in Ecuador. The hummingbird has short legs, small feet, narrow wings, and long feathers. The bill may be from $\frac{1}{3}$ to 5 inches (0.8–13 centimeters) in length, depending upon the species. It usually is straight. The tongue is specialized and tubelike. The voice is comprised of a weak chippering sound. The nest is very small, cuplike, built of plant fibers and spiderwebs, and lined with down. The nest opening is only about 1 inch (2.5 centimeters) in diameter. There are usually two white eggs.

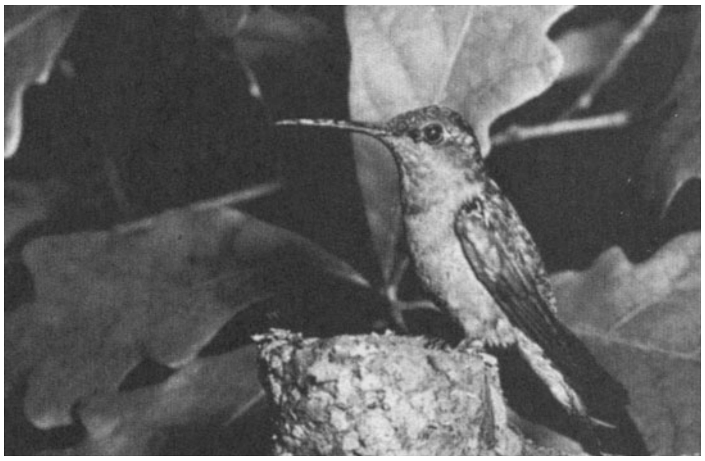


Fig. 2. Ruby-throated hummingbird. (*Harrison.*)

The hummingbird exhibits rather brave behavior around homes and people, wherever food is put out for them, or where flowers are in bloom. The bird can appear and disappear from the scene with great rapidity. The male hummingbird shows no concern for family affairs, leaving the nest building care and feeding of the young to the female. The female is often quite pugnacious when watching over her eggs and young, and can be quite dominating in terms of other hummingbirds when they encroach on a favorite feeding cup or flower bed. A few species of hummingbirds hibernate in winter. The ruby-throated hummingbird is one of the few species that inhabits the eastern United States. See Fig. 2. The coquette is one of a group of species found from southern Mexico to southern Brazil. These birds are distinguished by the crested head and conspicuous frills at the sides of the neck. See also **Apodiformes.**

SWORD-BEARER (*Insecta, Orthoptera*). A long-horned grasshopper of the group known as cone-headed grasshoppers from the conical prolongation of the head. The sword-bearer is named from the long ovipositor of the female. This organ is a slender slightly curved blade longer than the entire body. The species is found in the northern states of the United States, east of the Rockies. See also **Grasshopper.**

SYCAMORE AND PLANE TREES. Of the family *Platanaceae* (plane family), these trees are known as plane trees in Europe, but in America are more commonly referred to as sycamores or buttonwoods. To add to the confusion of nomenclature, the maples of Europe (family *Aceraceae*) of particular species, such as *A. pseudoplatanus*, are known as sycamore-maples. Plane trees are very common in Europe. At one time, it was estimated that over 60% of the trees in London were plane trees. The London plane is considered to be an early cross between the

RECORD SYCAMORE TREES IN THE UNITED STATES[1] (*Platanaceae* Family)

	Circumference[2]		Height		Spread		
Specimen	Inches	Centimeters	Feet	Meters	Feet	Meters	Location
Arizona sycamore (1981) (*Platanus wrightii*)	283	719	114	34.7	116	35.4	New Mexico
California sycamore (1945) (*Platanus racemosa*)	324	823	116	35.4	158	48.2	California
Western sycamore (1974) (*Platanus occidentalis*)	582	1478	129	39.3	105	32.0	Ohio

[1] From the "National Register of Big Trees," The American Forestry Association (by permission).
[2] At 4.5 feet (1.4 meters).

eastern plane (*Platanus orientalis*), which is native to Turkey, and the western plane (*P. occidentalis*), known as the sycamore or button wood in America. This tree is found extensively in the eastern United States, as well as in Turkey, Greece, southeastern Europe, and Asia Minor. The tree ranges from 75 to 100 feet (22.5 to 30 meters) in height, with a large trunk and broad spread. See accompanying table. The bark is dark gray, peels easily, with a lighter gray color underneath. The leaf is large, smooth, coarse, and permanently lobed. The flower is unisexual, about 1 inch across. The wood of the sycamore is light yellow, with a red-brown heartwood. It is tough, close-grained, and firm. Commercial sycamore wood, when cut and quartered, appears much like oak. The surface is lustrous and takes readily to a fine polish. The wood is used for tool handles, rollers, flooring, cooperage, furniture, and veneers. Other species include *P. silver*, native to the eastern United States; *P. racemosa*, found mainly in California; and *P. wrightii*, found in the southwest United States, notably Arizona.

SYENITE. A coarse-grained, granular, therefore intrusive, igneous rock of the general composition of granite except that quartz is either absent or present in a relatively small amount. The feldspars are alkaline in character and the dark mineral is usually hornblende. Soda-lime feldspars may be present in small quantities. The term syenite was originally applied to hornblende granite like that of Syene in Egypt from whence the name is derived. Syenite is not a common rock, some of the more important occurrences being, in the United States, in New England, Arkansas, Montana, and New York State (syenite gneisses), and elsewhere, in Switzerland, Germany, and Norway.

SYLVANITE. A mineral, a telluride of gold and silver approximating the formula $AgAuTe_4$. Sylvanite is monoclinic, occurring in bladed, columnar, and granular forms as well as arborescent and branching. It is a brittle mineral; hardness, 1.5–2; specific gravity, 8.16; luster, metallic; color and streak, steel gray to yellowish-gray. This mineral is found associated with gold and tellurides of gold and silver or with sulfides such as pyrite. It is found in Rumania, Australia, Colorado and California. It was named for Rumanian Transylvania where it was first found.

Krennerite is another telluride of gold and silver with a similar composition to sylvanite, but crystallizing in the orthorhombic system. Calaverite is a gold telluride with only a small silver content.

SYLVITE. A mineral, potassium chloride, KCl, occurring in cubes, or as cubes modified by octahedra. Sylvite is therefore isometric. It has a perfect cubic cleavage; uneven fracture; is brittle; hardness, 2; specific gravity, 1.9; luster, vitreous; colorless when pure but may be white, bluish, yellowish or reddish due to impurities. It is soluble in water. It is much rarer than halite and has been found as sublimates at Mt. Vesuvius and as bedded deposits at Stassfurt, Germany. Extensive deposits occur in sedimentary deposits in the Permian basin of southwestern New Mexico, near Carlsbad, in the United States.

It is used as a source of potash salts. Potassium chloride was called by the early chemists *sal digestivus Sylvii*, whence the name of the mineral.

SYMBIOSIS (Ecology). A close association between two organisms of different species in which at least one of the two benefits. The two organisms may be both plants, one may be a plant and the other an animal, or both may be animals. Three different kinds of symbiosis are recognized by ecologists:

1. *Mutualism*, the condition where both of the symbionts benefit from the association. A lichen, for instance, is a composite plant formed in actuality from two different plants. One, a fungus, gains nutrients from the alga, and the other, an alga, gains protection and an increased supply of water from the fungus. The result is a highly efficient plant structure which can grow in places where either type of plant alone could not exist. The nitrogen-fixing bacterium, *Rhizobus leguminosum*, lives in the roots of the legumes. The bacteria are protected in the nodules which form on the roots and the legumes benefit by an increased supply of nitrates fixed by the bacteria. The pollination of flowers by insects is an example of a more loosely knit mutualistic association, involving plants and animals. The rhinoceros and the tick bird provide a good example of mutualism among two higher animals (see illustration). The tick bird picks the ticks off the rhinoceros and gets food, and the rhinoceros gets rid of its ticks. Also, the tick bird has much keener eyesight than the nearsighted rhinoceros and gives warnings in times of danger by jumping up and down and uttering shrill cries.

An example of mutualism: a tick bird on the back of a rhinoceros. (*A. M. Winchester.*)

In some cases, the mutualism is *facultative*—either organism can exist without the other; but in other cases, the association is *obligatory*—neither organism can live without the other. An example of the latter is the association of termites and certain protozoa that live in their intestines. The termites eat wood, but have no enzymes for digesting wood. The protozoa have such enzymes, but cannot live on wood unless it is first chewed to a pulp by the termites. Termites die if the protozoa are killed by antibiotics and the protozoa die when they leave the body of the termites. The Smyrna fig and a tiny wasp are also entirely dependent upon one another. The fig can produce neither fruit nor seeds unless this species of wasp enters the young flower and accomplishes cross-fertilization. The wasps die without the fig because the eggs are laid within some of the figs.

2. *Commensalism*. When one symbiont benefits from the association and the other is neither harmed nor benefited. The shark sucker, *Remora*, which attaches itself to sharks and gets a free ride as well as some share of the food killed by the shark, benefits from the association, but the shark is not affected one way or another. Many of the bacteria living in the human mouth and intestine are also classed as commensals. The green hydra, *Chlorohydra viridisima*, is green in color because there are small one-celled algae in its gastrodermal cells. The hydra neither benefits nor is harmed by the presence of these plants; it gets along just as well without them. Hence, this is a case of commensalism.

3. *Parasitism*. When one symbiont benefits and the other is harmed. A tapeworm living within the intestine of a vertebrate animal is a good example of a parasite, since the host animal is definitely harmed. The tapeworm is an endoparasite, since it lives within the body of another animal; ectoparasites, such as fleas, leeches, and ticks, live outside the body and have a more temporary association.

SYMMETRIC. Arranged in accordance with a certain similarity with reference to a certain geometrical entity or position, which may be a point (center or point of symmetry), a line (axis of symmetry), or a plane (plane of symmetry), etc.

SYMMETRIC FUNCTION. A function whose value remains unchanged under any permutation of its independent variables. That is, any function $f(x_1, x_2, \ldots, x_n)$ not affected by an interchange of any x_i and x_j. If these are roots of an algebraic equation

$$x^n - c_1x^{n-1} + c_2x^{n-2} - \cdots \pm c_n = 0$$

then

$$c_1 = \sum x_i, \qquad c_2 = \sum x_ix_j, \qquad c_3 = \sum x_ix_jx_k \cdots$$

are the elementary symmetric functions, where the summations are extended over all distinct products of distinct factors, the x_i being themselves considered independently varying. Any rational symmetric function of the x_i is a rational function of the c_i.

See also **Permutation.**

SYMMETRY (Axis of). A line drawn within a body or within a set of points in such a location and direction that a rotation of the body through an angle $(2\pi/n)$ radians about the line as an axis, n being an integer, greater than unity, results in a configuration indistinguishable from the original configuration. A body or set having such an axis is said to have n-fold symmetry, and the line is said to be an n-fold axis. Thus a line through the center of a cube and parallel to a face is a four-fold axis of symmetry, while a body diagonal of the cube is a two-fold axis.

SYMMETRY (Center of). A symmetry element such that any line through it will intersect the crystal at equal distances on either side. Schoenflies symbol, subscript i.

SYMMETRY (Plane of). A plane passed through a body or through a set of points in such a location and direction that the reflection of all points in the plane results in a configuration indistinguishable from the original configuration. Thus, a cube has many planes of symmetry through its center, including those parallel to the faces and those passing through face diagonals.

SYMMETRY (Zoology). The arrangements of the parts of animal bodies in relation to centralized axes. The bodies of some one-celled animals are asymmetrical and of others, notably the Heliozoa, spherically symmetrical with the hard parts of the skeleton radiating in various directions from a common center. By far the most common forms of symmetry, however, are those known as radial and bilateral.

Radial symmetry is especially common among the sessile animals such as sea anemones and the related jellyfishes whose movements are weak. These animals have a principal axis passing through the mouth from which similar structures extend on several radii. The same form of symmetry appears in the echinoderms although these animals begin life as bilaterally symmetrical larvae. The radial symmetry of the adult accompanies sluggish movement and in some forms food-securing habits like those of sessile animals.

Bilaterally symmetrical animals have similar halves flanking a median plane in the principal axis of the body. Sense organs are concentrated near the end that goes first in locomotion, forming a head in which the mouth opens as a rule. This end of the body is the cephalic end, in contrast with the opposite caudal end where the tail is attached in the vertebrates. The originally upper and lower or dorsal and ventral surfaces are also differentiated, since the animal rests on the latter while the former is exposed to surrounding influences, and the sides of the body are known as right and left. This type of symmetry prevails in all actively moving animals.

The value of each type of symmetry is clearly correlated with the mode of life in which it is found. Sessile animals receive food and are subjected to dangers only when the responsible factors approach under their own powers of locomotion or on currents in the water. It is an advantage to the animal to be able to perceive such factors as easily in one direction as another. Bilaterally symmetrical animals move about in search of food, hence the end of the body that normally goes first has the chief need of powers of perception, while the upper and lower surfaces are exposed to different environmental conditions and the sides are similar in their contacts.

SYMPHYLA. Small and rare animals living in moist debris at the surface of the ground. They are related to primitive insects and centipedes and are usually regarded as a class of the phylum *Arthropoda*. They have a pair of antennae but no eyes. The segments of the body are well marked, bearing 11 or 12 pairs of legs. The animals breathe by means of tracheae.

SYNAPSE. The association between nerve cells of animals above the level of coelenterates. In the latter, the nerve network is composed of cells whose processes are structurally connected, but in higher nervous systems, the fine terminal branches of nerve processes merely come into close contact with those of adjacent cells.

Synapses can be defined as regions of structural specialization between two or more neurons. Since, as a rule, synapses conduct impulses only unidirectionally, some type of asymmetry might be expected in their structure—and there being regions of apposition between adjacent neurons, they may involve almost any parts of the two neuronal surfaces. The most common type of synaptic junction is that between an axon and a dendrite or a soma, the efferent fiber being expanded at its end to form a small bulb. When a nervous impulse reaches the terminal bulb of the axon, acetylcholine is liberated and depolarizes the secondary neuron, thus creating a nervous impulse throughout its length. Cholinesterase at the synapse, however, quickly inactivates the acetylcholine and prevents it from stimulating the dendrites of the secondary neuron more than once.

In the intact nervous system, the impulses normally travel only from the dendrites to cell body or soma, to the axon and finally to the end bulb.

How neurons form synapses in the development nervous system and distinguish appropriate from inappropriate synapses remains one of the central, unsolved problems in neurobiology. However, it has recently been found that cyclic AMP (adenosine $3':5'$-cyclic phosphate) effects synaptogenesis by regulating the expression of voltage sensitive Ca^{2+} channels suggesting that cyclic AMF affects post-translational modification of some glycoproteins and the cellular levels of certain proteins.

The nervous system mediates adaptive response of an organism to environmental changes or to changes in the organism itself. To this end, the nervous system is uniquely modifiable or plastic. Neuronal plasticity is largely the capability of synapses to modify their function, to be replaced, and to increase or decrease their number when required. Neuronal plasticity is maximal during development and is expressed after maturity in response to external or internal perturbations, such as changes in hormonal levels, environmental modification, or injury.

See also **Nervous System and The Brain.**

R. C. Vickery, M.D., D.Sc., Ph.D., Blanton/Dade City, Florida.

SYNCHRONOUS. A synchronous operation takes place in a fixed time relationship to another operation or event, such as a clock pulse. See also **Asynchronous.** When a set of contacts is sampled at a fixed time interval, the operation is termed synchronous. This situation is to be contrasted with that where the contacts may be sampled randomly under the control of an external signal. Generally, the read operation of a main storage unit is synchronous. The turning on of the X and Y selection drivers and the sampling of the storage output on the sense line are controlled by a fixed-frequency clock.

SYNCLASTIC SURFACE. A surface, or portion of a surface, on which the two principal radii of curvature at each point have the same sign. Also called *surface of positive total curvature*.

SYNCLINE. The syncline is a structure in which the strata are bent downward in an inverted arch, the sides of which are designated the limbs. The syncline may be a broad open fold or tightly compressed with steep dips, and pitch either upward or downward.

SYNCOPE. A fainting spell, in which the unconsciousness is due to a temporary cerebral anemia, i.e., insufficient circulation of blood in the brain. Differing only in degree from *true syncope* is dizziness (lightheadedness). Syncope also may occur from noncardiovascular causes, such as head injury, hypoglycemia, seizures, and hysteria.

SYNDROME. A group of symptoms characterizing or occurring in any abnormal state or disease.

SYNERESIS. The contraction of a gel with accompanying pressing out of the interstitial solution or serum. Observed in the clotting of blood, with silicic acid gels, etc. See also **Colloid System.**

SYNERGIC CURVE. A curve plotted for the ascent of a rocket vehicle calculated to give the vehicle an optimum economy in fuel with an optimum velocity. This curve, plotted to minimize air resistance, starts off vertically, but bends towards the horizontal between 20 and 60 miles (32 and 96 kilometers) altitude to minimize the thrust required for vertical ascent.

SYNODIC PERIOD. The synodic period of any member of the solar system is the time required for the object to go from some particular position relative to the sun as seen from the earth back to the same position. In the case of the moon, the synodic period is the time required to go from conjunction, or new moon, back to conjunction. This period of approximately 29.5 days is the original month as used by ancient astronomers in the construction of the calendar.

Since a planet is best observed at opposition, the synodic period of the planet gives the interval of time between successive positions of favorable observation. The synodic period is related to the sidereal period, i.e., the actual period of revolution of an object about the sun, by a simple relationship:

let P be the sidereal period of the object;
S the synodic period of the same object;
E the sidereal period of the earth (approximately 365.25 days);

then, for planets with orbits inside that of the earth, $1/S = 1/P - 1/E$; and for planets with orbits outside that of the earth, $1/S = 1/E - 1/P$.

SYNTHESIS (Chemical). The process of building chemical compounds through a planned series of steps (reactions, separations, etc.). Synthesis usually is the method of choice (1) when the desired compound is not present in natural materials from which it can be isolated, (2) when the compound cannot be easily obtained from reacting readily available materials in a few simple steps; and (3) although a compound may be available within a natural complex, the economic separation and purification are prohibitive, or often in the case of biochemicals, too little natural raw material is available to meet the demand.

Even more important, synthesis plays a key role in developing new, untried chemical structures which, on paper, appear to have properties that may be of great value, e.g., a new synthetic material, a new drug, or a new fuel. Chemicals by design from prior knowledge of related materials generally are created via the route of synthesis. Further, synthesis is fundamental to broadening the base of chemical knowledge. Sometimes unexpected results occur, i.e., compounds with unusual, unexpected, and often desirable practical chemical and/or physical properties.

Because of the hundreds of thousands of organic substances already established, but many yet remaining to be "built," organic synthesis predominates. Most of the synthetis (elastomers, fibers, and other polymers, coatings, films, adhesives, and numerous other products) that have appeared during the last 30 to 40 years resulted from research involving organic synthesis. Some of the early work in organic synthesis dealt with the creation of certain fatty acids and ketones. A few examples are given to provide an insight into the workings of synthesis.

In the following examples, only the main starting ingredients and products are shown. No attempt is made to indicate byproducts or the conditions of the reactions involved:

(a) Target compound: Ethylpropylacetic acid, $(C_2H_5)(C_3H_7)CH{:}COOH$
  (1) Acetic anhydride $\rightarrow$ ethyl acetate
    (+ alcohol)
  (2) Ethyl acetate $\rightarrow$ ethyl acetoacetate
    (sodium + dilute acids)
  (3) Ethyl acetoacetate $\rightarrow$ sodium derivative of ethyl acetoacetate
    (+ sodium ethoxide)
  (4) Sodium derivative of ethyl acetoacetate $\rightarrow$ ethyl ethylpropyl acetoacetate
    (+ propyl iodide)
  (5) Ethyl ethylpropyl acetoacetate $\rightarrow$ ethylpropylacetic acid
    (concentrated alcohol and potash)
(b) Target compound: Butyl acetone, $CH_3{\cdot}CO{\cdot}CH_2{\cdot}C_2H_4$
  (1) through (3), same as given in example (a)
  (4) Sodium derivative of ethyl acetoacetate $\rightarrow$ ethylbutylpropyl acetoacetate
    (+ butyl iodide)
  (5) Ethylbutylpropyl acetoacetate $\rightarrow$ butyl acetone
    (+ dilute alcohol and potash)
(c) Target compound: *n*-valeric acid, $CH_3{\cdot}CH_2{\cdot}CH_2CH_2COOH$
  (1) Potassium chloroacetate $\rightarrow$ potassium cyanoacetate
    (+ potassium cyanide)
  (2) Potassium cyanoacetate $\rightarrow$ ethyl malonate
    (+ alcohol and hydrogen chloride)
  (3) Ethyl malonate $\rightarrow$ sodium derivative of ethyl malonate
    (+ sodium ethoxide)
  (4) Sodium derivative of ethyl malonate $\rightarrow$ ethylpropyl malonate
    (+ propyl iodide)

The compounds on the right-hand side of intermediate reactions are often called *intermediates*. See also **Intermediate (Chemical).**

Some of the notable syntheses from the early history of the technique include:

Inorganic Syntheses
1746 Sulfuric acid (chamber process)
1800 Soda ash (Le Blanc process)
1861 Soda ash (Solvay process)
1890 Sulfuric acid (contact process)
1912 Ammonia (Haber-Bosch process)
Organic Syntheses
1828 Urea (Wohler)
1857 Mauveine (Perkin)
1869 Celluloid (Hyatt)
1877 Ethylbenzene (Friedel-Crafts)
1884 Rayon (Chardonnet)
1910 Phenolic resins (Baekeland)
1910 Neoarsphenamine (Ehrlich)
1920 Aldehydes, alcohols (Oxo synthesis)
1925 Insulin (Banting)
1927 Methanol
1930 Neoprene (Nieuwland)
1935 Nylon (Carothers)
1940 Styrene-butadiene rubber
1950 Polyisoprene

SYNTHESIS GAS. For a number of industrial organic syntheses that proceed in the gaseous phase, it is advantageous to prepare a chargestock to specification. When a mixture of gases is so prepared, the term *synthesis gas* is often used. Thus, there are several mixtures which qualify under this definition: (1) a mixture of H_2 and N_2 used for NH_3 synthesis; (2) a mixture of CO and H_2 for methyl alcohol synthesis; and (3) a mixture of CO, H_2, and olefins for the synthesis of oxo-alcohols. Ammonia synthesis gas is described briefly here.

The hydrogen required for NH_3 synthesis gas may be obtained in commercial quantities from coke oven water gas; from steam reforming of hydrocarbons; from the partial oxidation of hydrocarbon chargestocks; or from the electrolysis of H_2O. The nitrogen required may come from the introduction of air to the process, or where specifically required, pure nitrogen may be obtained from an air separation plant. Since NH_3 synthesis occurs under high pressure, it is advantageous to generate the synthesis gas at high pressure and thus avoid additional high compression costs. For this and other economic situations, coke oven gas and hydrogen from electrolysis are eliminated. This leaves hydrocarbons as the logical choice.

In the steam-hydrocarbon reforming process, steam at temperatures up to 850°C and pressures up to 30 atmospheres reacts with the desulfurized hydrocarbon feed, in the presence of a nickel catalyst, to produce H_2, CO, CO_2, CH_4, and some undecomposed steam. In a second process stage, these product gases are further reformed. Air also is added at this stage to introduce nitrogen into the gas mixture. The exit gases from this stage are further purified to provide the desired 3 parts H_3 to 1 part N_2 which is the correct empirical ratio for NH_3 synthesis. See also **Ammonia.**

SYNTHETIC DIVISION. An abbreviated process, using detached coefficients, for finding the quotient of a polynomial in one variable x by a divisor of the form $x - r$, where r is a constant. The procedure may be illustrated by the polynomial $a_0x^4 + a_1x^3 + a_3x + a_4$. The results can be obtained in the following form, which should be self-explanatory:

$$\begin{array}{ccc}
a_0 & a_1 & a_2 \\
 & a_0r & A_1r \\
a_0, & A_1 = a_0r + a_1, & A_2 = A_1r + a_2
\end{array}$$

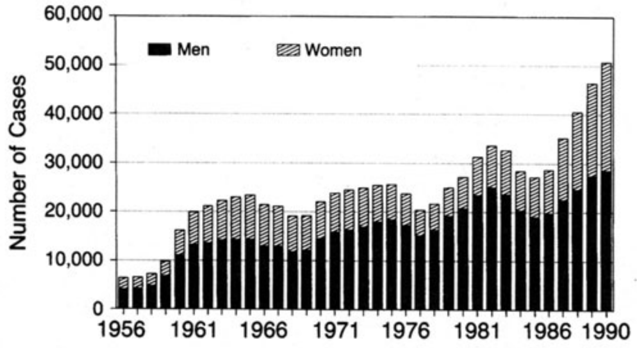
$$\begin{array}{ccc}
a_3 & a_4 & r \\
A_2r & A_3r & \\
A_3 = A_2r + a_3, & R = A_3r + a_4 &
\end{array}$$

The quotient is $a_0x^3 + A_1x^2 + A_2x + A_3$ and the remainder is R. A similar process applies to a polynomial of any other degree. In the general case, if the coefficient of any power of x is missing, a zero should be supplied in the appropriate place in the first line of the scheme. If the remainder in the synthetic division is zero, then the divisor is a factor of the given polynomial.

See also **Polynomial.**

SYPHILIS. An infectious disease, congenital or acquired, caused by the spirochete, *Treponema pallidum*. Syphilis is characterized by primary and secondary stages during which it is highly contagious, and a noncontagious, late or tertiary stage, marked by involvement of many organs and tissues. In the congenital form there is no primary lesion and late manifestations predominate.

Epidemiology. The number of cases of syphilis reported and the reporting requirements range widely from one country to the next. In most industrialized nations, the disease is reasonably well documented, but even then most health officials believe that the disease is underreported, mainly in the underdeveloped countries, as exemplified by parts of Africa. The accompanying chart clearly depicts the alarming trends of case numbers in the United States since the mid-1950s. As observed by E. W. Hook (University of Alabama, Birmingham) and C. M. Marra (American Social Health Association) in April 1992, "In the early 20th century, syphilis was a major public health problem. In the 1920s more than 20 percent of the patients in U.S. mental institutions had tertiary syphilis (general paresis). Even before the therapeutic efficacy of penicillin was described, public health measures such as serologic-screening programs and treatment through government-funded rapid-treatment centers had begun to reduce the rate of syphilis in the United States. By 1941, only 10 percent of patients in mental institutions had tertiary syphilis. These changes accelerated after World War II as penicillin became widely available. In 1979 only 162 syphilis-related admissions to mental institutions were reported, 98 percent fewer than 60 years earlier. Similarly, the number of syphilis-related deaths declined more than 50-fold between the early 1940s and the 1980s."



Number of cases of primary and secondary syphilis in the United States between 1956 and 1990. (SOURCE: *Centers for Disease Control, Atlanta, Georgia.*)

By the mid-1950s only about 6500 cases of primary and secondary syphilis were reported per year in the United States. In the late 1950s and early 1960s, the number of cases increased, fluctuating between 19,000 and 26,000 cases per year until 1978. The rates increased slightly in the late 1970s and early 1980s, and a disproportionate number of cases occurred in homosexual men. In the mid-1980s, primarily because of behavioral changes adopted in response to the AIDS epidemic, syphilis declined among homosexual men, as did the male-to-female ratio of cases.

In 1985, the incidence of syphilis among heterosexual men and women began to increase rapidly, a trend that also led to dramatic increases in congenital syphilis. Between 1983 and 1990, when 50,225 cases of primary and secondary syphilis were reported, there was a 75% increase in the incidence of the disease. Although widespread, the changes were not geographically or racially uniform. The rates among black men increased 126% (from 69 to 156 per 100,000) and among black women 231% (from 35 to 116 per 100,000), whereas the rates for other subgroups of the population remained stable or even declined. Rates for non-Hispanic white men declined 50%, from 6 to 3 per 100,000 during the same period. Although rates of syphilis remain highest in urban areas of the southeastern United States and Texas, changes have been reported nationwide in both rural and urban areas. From 1989 to 1990 some of the greatest increases occurred in midwestern cities where syphilis had previously been uncommon. For example, in three Ohio cities the rates increased between 235% and 293%, and similar increases occurred in St. Louis (172%) and Milwaukee (153%).

Several factors have contributed to the changing epidemiology of syphilis: (1) limited access to health care has delayed the diagnosis of syphilis; (2) a linkage with concomitant epidemics of illegal drug use, particularly of "crack" cocaine because cocaine is associated with prolific engagement in sexual practices as well as the "exchange" of sex in payment of drugs; and (3) the difficulty in tracing sexual partners among drug users. To address these problems, public health officials in some areas have modified traditional intervention activities, performing screening at crack houses and considering mass treatment of well-defined, identifiable risk groups.

Social Vectors of the Disease. A. M. Brandt (Harvard Medical School) has reviewed how past experience in connection with the sociological aspects of syphilis may be helpful in relation to the current crisis concerning AIDS (acquired immunodeficiency syndrome). Brandt observes that AIDS, like syphilis in the past, engenders powerful social conflicts about the meaning, nature, and risks of sexuality; the nature and role of the state in protecting and promoting public health; the significance of individual rights in regard to communal good; and the nature of the doctor-patient relationship and social responsibility. The analogs that AIDS poses to syphilis are striking: the pervasive fear of contagion, concerns about casual transmission, the stigmatization of victims, and the conflicts between public health and civil liberties. The importance of the history of syphilis (reviewed by Brandt in *Science*, **239**, 375–369, January 22, 1988) is that it reminds the public of that range of forces that influence disease, health, and social policy.

Chronology of Syphilis. The early history of syphilis is not entirely clear. It is thought by some historians that the disease was first introduced into Europe by Columbus' returning sailors, and subsequently spread through Italy where it became a great scourge, by the soldiers of Charles V. Another view is that the disease has existed in civilized man since antiquity. Early Egyptian and Assyrian inscriptions as well as bony changes found in mummies are interpreted as supporting this theory. In the Middle Ages, syphilis occurred in severe, widespread epidemics with enormous mortality rates. Later it became a milder disease, and its venereal nature was recognized. It was not until 1905, when Schaudinn discovered the *Treponema pallidum* to be the causative organism, that syphilis and gonorrhea were recognized as two distinct diseases. Development of the darkfield microscope made it possible for the first time to detect the causative spirochete.

Acquired Syphilis

The disease is almost always transmitted during sexual intercourse. The causative agent gains entry through an abraded surface, usually on the genital organs. In the male, the primary sore is commonly on the penis, where it is easily seen. In the female, it is found either on the mucous surface of the vulva or within the vagina where it may remain unnoticed.

The primary stage is marked by the appearance of the chancre, which develops at the site of invasion, usually 2–4 weeks after exposure. Before the chancre develops, the organisms often have invaded the body tissues by way of the lymph channels and blood stream. A fully developed chancre appears as a clean, slightly raised, hard circumscribed ulcer, which exudes a thin, highly infectious secretion. By darkfield microscopic examination, this secretion can be seen to teem with spirochetes. The lesion is painless, but tender lymph nodes may develop in the groin.

The secondary stage begins 6 weeks to 6 months after the primary. By the time of onset of this period, the organisms have invaded the tissues, and the body's defense reactions are active. Antibodies are being produced, and the Wassermann reaction is positive. The second stage may be so mild as to pass unnoticed, but it is usually ushered in with a rash over the skin and mucuous membranes. The eruption is extremely variable and may imitate any skin disease. The commonest type is the macular, in which the lesions consist of flat, or slightly raised, rose-colored spots, most prominent over the abdomen and chest. Characteristic features of the secondary eruption are its symmetrical distribution, painless, non-itching nature, and its tendency to appear on the palms and soles. Its duration is variable from weeks to months, and it may fade and leave a faint pigmentation for a time. With treatment the rash disappears promptly. Spirochetes are present in skin and mucous-membrane lesions, and are particularly easy to demonstrate in the soft moist sores on the genitalia, the condylomata.

Constitutional symptoms in the secondary stage are usually mild. They consist of sore throat, slight fever, headache, and some enlargement of the superficial lymph nodes.

The latent period is the interval between the secondary and tertiary phases of the disease. During this time, the patient may enjoy excellent health, and be unaware that he has syphilis. The duration is variable—from a few weeks to many years—even 25 to 30. During this time, the various organs and tissues are harboring spirochetes, but these are inactive, hibernating as it were.

The tertiary stage is the late picture, which is usually seen only in untreated syphilis. It is marked by the development of gumma, tumor-like masses in the skin and visceral organs, and inflammatory changes in the cardiovascular and central nervous systems. Since any organ or tissue may be attacked in tertiary syphilis, the signs and symptoms are extremely varied.

The most common types of tertiary syphilis are those involving the cardiovascular and central nervous systems. Cardiovascular involvement usually appears earlier—10 to 15 years after infection. It occurs as aortitis, aortic aneurysm and sometimes coronary occlusion. In syphilitic heart disease, aortitis is commonly followed by insufficiency of the aortic valve. This mechanical defect puts a tremendous strain on the heart, which compensates by hypertrophy of the muscle. Eventually the strain is too great for the reserve, and heart failure and finally death are the result.

Central nervous system syphilis may not develop for 30 years after the primary infection. It occurs as a meningitis, as an arteritis or inflammation of the small cerebral vessels with secondary changes in the brain tissue, and as involvement of the cells of the brain and cord in tabes dorsalis and paresis.

Congenital Syphilis

It is important that no pregnant woman have syphilis, for the causative organism of this disease is one of the few that can pass through the placental barrier to the unborn child. Once the disease establishes itself in the baby, abortion or stillbirth may follow. If syphilis is diagnosed early in pregnancy, treatment can be started which may curb the disease and allow the baby to develop normally. In the most severe forms of infection, the child may show at birth an extensive skin eruption, fissures about the angles of the mouth and nose, a characteristic nasal discharge, bone lesions, and enlargement of the liver and spleen. Underdevelopment and poor nutrition are conspicuous. Such a child rarely lives longer than a few days. In other instances, the infected infant may appear normal at birth, yet after a few months develop the typical signs and symptoms of the disease.

Children who survive the active congenital disease, or those in whom the infection remains latent, often show certain permanent stigmata which make the diagnosis apparent at a glance. "Saddle nose" and deformed (Hutchinson's) teeth—peg-shaped notched incisors which are widely spaced—are characteristic. Bony lesions, inflammation of the cornea of the eye (interstitial keratitis), deafness, and central nervous system syphilis, such as occur in the acquired form, are also seen in children with congenital syphilis.

A test for syphilis should be performed on all pregnant women at the initial visit to the obstetrician and repeated during the third trimester in all cases where there may be a risk of acquiring syphilis. Where diagnosis is positive, the patient should be treated. Early syphilis in pregnancy is treated with the same dosage of antibiotics as for any other person in whom the disease has been diagnosed. Tetracycline is not recommended for syphilitic infections in pregnancy because of potential toxicity for mother and fetus. Monthly tests for syphilis should be made during the full period of pregnancy. Where effective treatment is given, the risk of congenital syphilis in the newborn is minimized.

Diagnosis of Syphilis

The first blood test for the disease was developed by August von Wassermann, a German bacteriologist, in 1907. The Wassermann test was used for many years, but during the interim a number of other serologic tests have been developed. These tests fall into two main categories: (1) treponemal tests which are designed to detect the treponemal antibody produced in response to syphilitic infection; and (2) nontreponemal or reagin tests, which detect an antibody-like substance (reagin), the latter found in the serum of an infected patient. Reagin is assumed to result from the interaction of *T. pallidum* with body tissue. Compared with the nontreponemal tests, the treponemal tests are relatively time-consuming and costly.

To guide therapeutic decisions and disease-intervention activities, syphilis is divided into a series of clinical stages. Despite its usefulness, clinical staging is imprecise. Patients with late stages of disease may have no recollection of signs of earlier stages, possibly because most syphilitic lesions are painless or because some patients may not have clinically apparent primary or secondary lesions. Also, there is considerable overlap between stages. For greater detail pertaining to diagnostic procedures, refer to the Hook-Marra paper.

Therapy for Syphilis. The primary goals of therapy are to prevent transmission and avoid the late complications in affected patients. *Treponema pallidum* cannot readily be propagated in vitro. Consequently, much research has been conducted using rabbit models.

Penicillin remains the best-studied and the preferred therapy for syphilis. A single dose of 2.4 million units of penicillin G benzathine results in a serum penicillin concentration of more than 0.018 micrograms per milliliter, which remains effective for about 3 weeks. Studies have shown that this dose for early syphilis is not complete and that retreatment is required in up to 10% of patients. There have been reports that penicillin G benzathine does not relieve neurologic and ocular syphilis. However, prior to the era of the HIV epidemic, penicillin G benzathine therapy for early syphilis had a success rate of 95% or higher. This high rate was attributed to the fact that the therapeutic regimen and an intact immune system act together to clear peripheral organisms as well as those sequestered in the central nervous system and the eye. When symptoms of neurosyphilis are evidenced, the preferred treatment may be the use of high-dose intravenous penicillin (sterile) or a combination of intramuscular penicillin G procaine with oral probenecid.

Treatment of persons who are allergic to penicillin, repeated dosage of tetracyclines (although not for pregnant women), and possibly erythromycin may be considered. Information is still being collected on the efficacy of other antibiotics. Persons with the HIV infection present a considerably more complex situation. Both syphilis and HIV are protean diseases, and they interact on a number of levels. Syphilitic genital ulcers may enhance the acquisition and transmission of HIV. The natural history of syphilis may be modified in patients coinfected with HIV (human immunodeficiency virus). The results of laboratory tests for syphilis may be different in HIV-infected persons, thus misleading the proper therapy to be used. Numerous studies have shown that syphilis, as well as other genital-ulcer diseases, are disproportionately common in HIV-infected patients and vice versa, suggesting that syphilis increases the risk of the acquisition and transmission of HIV. Much more detail on this subject is given in the Hook-Marra reference listed.

Untreated Syphilis

As a cause of death, untreated syphilis has received extensive attention of medical statisticians and long and complex studies of this topic continue. Questions for which specific answers have not been fully provided include: To what degree does syphilis masquerade as another disease? If left untreated, what is the ultimate pathway of destruction caused by syphilis? To what degree is syphilis indirectly associated with death from other causes?

In an extensive study (Gjestland) of figures gathered by Boeck and Brusgaard pertaining to 2,000 syphilitic patients who lived in Oslo, Norway during the period 1891–1910 and who had received no specific therapy for the disease, the following findings were revealed: (1) Untreated syphilitic patients exceeded their expected mortality rates by 53% (male) and 63% (female); (2) the male patients developed cardiovascular syphilis in 13.0% of the cases—7.6% (female); (3) neurosyphilis developed in 9.4% (male) and 5.0% (female); (4) the mortality rate from syphilis among males was double that of females; (5) clinical or autopsy evidence of syphilitic pathology of a serious nature was noted in 23% of all patients.

A similar study was made of males 25 years or older in the United States. Known as the Tuskegee Study, these findings have been summarized as follows: (1) Only one-fourth of the untreated syphilitics were normal after an infection of 15 years' duration. Most of the abnormal findings were associated with the cardiovascular system; (2) During the first 12 years of observation, 25% of the syphilitics and 14% of the control group (no syphilis) of comparable ages had died. At age 25, untreated male syphilitics would have a reduction of life expectancy of about 20%; (3) After 25 years of follow-up to this study, the life expectancy of individuals with syphilis (ages 25 to 50) was found to be 17% of the control group. A male with untreated syphilis of more than 10 years' duration and a sustained reactive serology would have approximately a 1:1 chance of having demonstrable cardiovascular involvement (autopsy findings).

Another interesting study concerned with the outcome of untreated syphilitic infection was done by Rosahn (Yale) by analyzing autopsy records from 1917 to 1941, consisting of 3,907 cases over age 20. Major observations included: (1) About 9.7% of the population studied had clinical laboratory or anatomic evidence of syphilis. Earlier, in 1938, Vonderlehr and Usilton estimated that one person in ten in the United States would have syphilis before dying; (2) Males with lesions at autopsy comprised 4.7% of the male population; 2.7% of the females, indicating a sexual bearing upon the resistance to tissue changes of late syphilis; (3) Syphilis significantly reduced longevity, whether or not there were tissue lesions; (4) No greater frequency of anatomic lesions was apparent among blacks than whites, but whites with such lesions died at a greater rate than blacks, indicating that as previously believed, syphilis does not run a more fatal course among blacks than among whites; (5) About 39% of untreated syphilitics indicated anatomic evidence of syphilis at autopsy; (6) About 23% of the untreated syphilitics died primarily as the result of the disease. There were several parallels between the Yale study and the Brusgaard study previously described.

Additional Reading

Brandt, A. M.: "No Magic Bullet: A Social History of Venereal Disease in the United States Since 1880," Oxford University Press, New York, 1987.
Brandt, A. M.: "The Syphilis Epidemic and Its Relation to AIDS," *Science*, 375 (January 22, 1988).
Handsfield, H. H., and J. Schwebke: "Trends in Sexually Transmitted Diseases in Homosexually Active Men in King County, Washington," *Sexually Transmitted Diseases*, **17**, 211 (1990).
Hook, E. W., and C. M. Marra: "Acquired Syphilis in Adults," *N. Eng. J. Med.*, 1060 (April 16, 1992).
Hutchinson, C. M., and E. W. Hook: "Syphilis in Adults," *Medical Clinician North America*, **74**, 1389 (1990).
Sparling, P. F.: "Natural History of Syphilis," in *Sexually Transmitted Diseases* (K. K. Holmes and P. A. Mandh, Editors), McGraw-Hill, New York, 1990.
Staff: "Prevention: Sexually Transmitted Disease Surveillance," Centers for Disease Control, Atlanta, Georgia, 1991.

SYSTEMIC LUPUS ERYTHEMATOSUS. Generally grouped with the rheumatic diseases, *systemic lupus erythematosus* (frequently shortened to *lupus* and abbreviated, SLE) was, until a few years ago, considered to be a rather rare condition. With an improved understanding of the disease and the use of more definitive diagnosis and precise nomenclature, the disease has become better identified. It is now variously estimated to affect 5 cases per 100,000 population—to as high as 50,000 new cases per year in the United States alone. Records to date indicate that SLE occurs in women at a rate nearly ten times that of men and that it usually appears during the second and third decade of life. Thus, the generalization can be made that SLE is predominantly a disease of young women.

Systemic lupus erythematosus is a chronic multisystem inflammatory disease with currently unknown origin. Clinical features are quite variable. SLE ranges from a moderate, benign condition to one with serious consequences that can lead to death. As with rheumatoid arthritis, SLE is an autoimmune disease, in which the body's natural defense mechanisms behave erratically. See **Immune System and Immunology.**

The diagnosis of SLE is frequently complex, but the physician will suspect the condition when a young female is seen with a combination of fever, skin rash, and arthritis. The fever is usually low-grade, but in a so-called "lupus crisis" may reach 104.9°F (40.5°C). The typical skin lesion appears in the facial region, particularly involving the nose and cheeks, producing what is sometimes called a "butterfly rash." Alopecia (spotty loss of hair) also may occur. There also may be ulcerations of the mouth and lips. It is believed that the skin conditions arise from a superabundance of keratin which plugs the hair follicles and sweat and sebacious glands. The arthritis most often involves the hands, wrists, elbows, knees, and ankles. Cartilage or bone degradation is not usually found. Tendonitis is not uncommon.

Further examination may show involvement of SLE in other organs. causing, for example, acute hemolytic anemia, thrombocytopenic purpura, or pericarditis. Most frequently, after a flare-up, there will be characteristic remissions which may extend from months to years. The risk of an acute attack depends to a large degree upon the number of organs that have been affected. Prognosis of recovery and remission diminishes in cases where the kidneys and central nervous system are involved. Situations which appear to precipitate the occurrence of SLE and which aggravate symptoms appear to include infections, stress (emotional and physical), surgery, pregnancy, some drugs, and undue exposure to sunlight.

Moderate involvement of the kidneys may occur in nearly all cases. Less commonly, renal malfunction may be serious and pose a life-threatening situation. Nervous system involvements occur in about half the cases, these varying considerably in magnitude. These may be manifested in psychoses and, in particular, organic psychoses which include deterioration of intellectual capacity, memory, and disorientation. These manifestations are usually of a transient nature, but they may occur any time during the course of the disease. Cardiac involvement is usually confined to pericarditis, but endocarditis occurs in some cases. Involvement of the respiratory system may cause pleuritic chest pain. Pleural fluid, in such cases, will almost always show high concentrations of protein. In some cases, bacterial pneumonia, pulmonary infarcts, uremic pneumonitis, and congestive heart failure may develop.

The treatment of SLE, as of the early 1990's remains somewhat controversial and consequently nonuniform. This will change with time as both fundamental scientific information, better clinical statistical data, and empirical observations are collected. Treatment is complicated by the variety of symptoms presented by SLE in various patients. Generally, therapy is directed to greater periods of rest to reduce the effects of emotional and physical stress. Aspirin is commonly suggested as an anti-inflammatory drug. When skin rashes are difficult to control, antimalarials, such as chloroquine and hydroxychloroquine, have been used.

Glucocorticoids will demonstrate marked suppression of SLE manifestations in many patients, but these drugs are usually reserved for patients with serious life-threatening conditions.

Currently, there is much research directed toward a better understanding of the origin and course of SLE. This research is motivated not only in the interest of developing improved therapy for SLE patients, but also as a probable rewarding avenue to a better understanding of many autoimmune phenomena. Histocompatibility antigens located on cell surfaces are antigens that elicit rejection of transplanted organs by the immune system. Researchers are interested in these antigens as regards SLE to determine if they are directly involved in provoking autoimmunity or that they may be indirectly involved because of their genetic association with immune response genes. Faulty control of immune responses has been strongly implicated in the etiology of lupus.

SYSTEMS ENGINEERING. A term widely used in engineering and industrial planning, defined somewhat differently in various organizations. Probably there is general agreement that it is the application of the scientific method to a communications system, a data processing system, or an aircraft, a ship or an entire business. Moreover, most of its practitioners would extend the definition to groups of businesses, that is, to industries or even entire societies.

SYZYGY. The points in the moon's orbit about the earth at which the moon is new or full.