

Automobile Technology

Hans Joachim Förster

Daimler Chrysler Company, Technical University of Karlsruhe

Hermann Gaus

DaimlerChrysler Company

I. Introduction
II. Man as Driver
III. The Vehicle
IV. The Power Train
V. Driving
VI. Passive Safety
VII. Outlook

GLOSSARY

Chassis Assembly of frame, axles, suspension, and wheels.

Diesel engine Engine in which the fuel is injected into the compressed air and ignited by its heat.

Drive train Power path from the engine via transmission, axle gear, shafts, and joints to the wheels.

Efficiency Quotient of output by input power of a drive train, part of it, or any other power-using or power-transmitting machine.

Energy density Energy content of fuel divided by the mass or the bulk volume of fuel alone or of fuel plus container.

Kinesthesia The perception or consciousness of one's own muscular movements.

Servo power Reduces the effort of actuating.

Spark-ignition (SI) engine Engine inwhich the compressed air–fuel mixture is ignited by an electric spark.

Speed converter Equipment that permits different speeds at input and output without altering the transmitted torque.

Torque converter Equipment that permits one to vary the power factors torque and angular velocity between input and output without altering the transmitted power. Synonyms include transmission, gearbox, and continuously variable ratio transmission (CVT).

AUTOMOBILE TECHNOLOGY comprises the technical means and processes by which the automobile is developed, built, and improved. At the center of automobile technology stand humans—in a double role. On the one hand, they are the customers for whose transportation needs the automobile was created. On the other hand,

as drivers, they are an integral component of the technical system of the motor vehicle. Even if an automobile has been built as a safe, economic, and nonpolluting unit, these qualities are of value in practice only if the driver drives safely, economically, and with consideration for the environment. In most countries, nearly everyone over the ages of 16–18 years can obtain a driver's license after a short period of training and testing. As a result, the species "vehicle driver" incorporates the whole spectrum of human characteristics: sex, body size, age, predisposition, character, strength, and intelligence. Automobile engineers therefore must reckon with "ordinary people" with their amazing capabilities to learn and react, but also with limited long-term performance and, in the technical sense, basic unreliability. Since man cannot be selected or molded according to the requirements of the automobile, the automobile engineer must adapt the vehicle to humans. Automobile technology is therefore a high-level technology whose results are measured by how well and sensibly humans handle the product in traffic.

I. INTRODUCTION

The motor vehicle is more than 100 years old. In 1885, Gottlieb Daimler rode a motorcycle for the first time, and on July 3, 1886, the *Neue Badische Landeszeitung*, Mannheim (Germany), reported the first excursion of Carl Benz in a car, his velocipede (Fig. 1). Since that event, the development of the automobile and its technology has continued unceasingly in many places throughout the world. Today there are ~700 million motor vehicles, of which ~510 million are passenger cars. The remainder are commercial vehicles, buses, motorcycles, and tractors. This worldwide distribution and the continuing demand for motor vehicles are due to the fact that the motor vehicle fulfills the basic human need for individual mobility, the ancient dream of the seven-league boots. It provides free choice of when, where, and with whom one will travel, wide range and high speed, and nearly effortless travel. In addition, it offers protection from the weather and from danger; it serves as both draft horse and pack mule for trailers and luggage; it creates a kind of privacy; and it provides a means of self-expression.

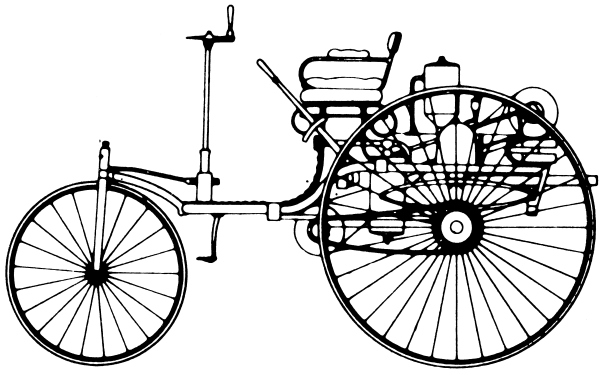


FIGURE 1 Benz velocipede, 1886. One cylinder; bore, 91.4 mm; stroke, 150 mm; displacement, 0.99 liter; power, 0.89 hp at 400 rpm; maximum speed, 15 km/hr.

So far, only the positive aspects of the automobile have been mentioned. Unavoidably, however, the large number of cars also creates serious problems, including traffic accidents, noise and air pollution, energy consumption, and changes in the environment brought about by the construction of streets and highways. Solving these problems, which affect everyone, is the most important task facing all those working in the field of road traffic, where many components—the motor vehicle, driver, roadway and surroundings, traffic information, and organization—interact.

II. MAN AS DRIVER

Of the many characteristics and capabilities of man, some must be taken into consideration with which man as driver can manage the task of vehicle driving. There are three main areas:

1. Taking in information that is relevant to the traffic, both from the outer field and from the vehicle
2. Processing information—that is, comparing it with the desired values to determine the necessary corrections when there are deviations
3. Operating the vehicle

The *taking in of information* occurs via five channels: visually (through the eye), vestibularly (through the balance organs in the ear), kinesthetically (through the perception of one's own muscular movements), haptically (through the sense of touch), and aurally (through the ear).

The visual channel is the most important, because more than 90% of all information passes through it, and if this channel fails steering of the vehicle is no longer possible. The vestibular sense registers accelerations. The messages regarding lateral movements are especially important, because they are already sensed before a sideways deviation can be seen by the eyes. Feedback from the motor movements to the legs and arms affect mainly the kinesthetic and sometimes the haptic senses also. The auditory channel is used very little nowadays for driving feedback, since engine noise has been reduced considerably. However, because mobile phones are widespread, the auditory channel is used more and more for information transfer (speech)

to and from the outside world. Traffic information is often supplied by radio or communication systems. The role of the auditory channel will increase in the future due to increased oral instead of visual information (e.g., navigation systems with spoken directions).

Information processing can take different paths (Fig. 2). In the conscious, cortical function loop, the sequence of recognition, decision, and action operates relatively slowly, 0.6–2 sec (16 bit/sec) per operation. The capability to grasp information appears to be limited to these 16 bits/sec and, with close attention, tiring occurs rapidly. However, during habitual, subcortical, subconsciously occurring processes, considerably better figures are obtained: 0.14–0.45 sec (10^6 bit/sec) per process. From large quantities of information, relevant partial quantities are selected sensibly. Tiring occurs only after longer periods of activity.

This human capability to habitualize (or establish) certain reactions and sequences of actions so that they are carried out quickly and reliably subconsciously is the crucial precondition for mastering the task of vehicle steering. Automobile and traffic engineers must ensure that this type of vehicle steering is possible (see Section V) and that the demands made on drivers by both the traffic and the vehicle remain within the range of their capabilities.

The *operations for handling* the vehicle require the use of all the available extremities: legs, feet, arms, and hands. However, because there are more tasks and more levers and buttons than there are feet and hands, many operations must be carried out sequentially. The working capacity of the weakest person determines the limits of reasonable operating forces, therefore, it is often necessary to reduce the required work of the driver by servo power.

The size of the vehicle must be determined by human dimensions or those of the goods to be transported, where the driver's working space has priority. Table I lists body dimensions and weights for men and women in North America. Table II shows those for selected regions. These

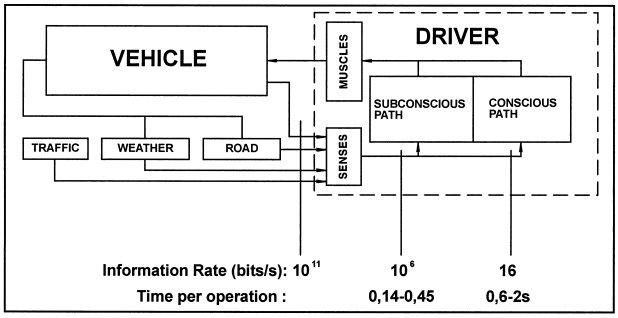


FIGURE 2 Ways of processing information when driving an automobile, according to W. Durth in "*Der Mensch als massgebliche Groesse im Strassenentwurf*," T. H. Darmstadt.

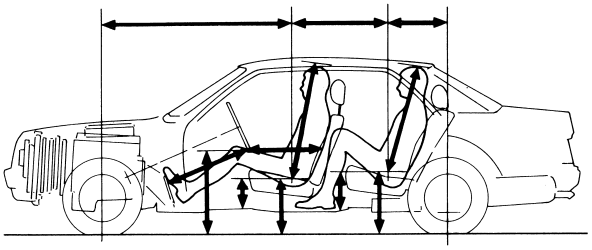


FIGURE 3 A human as a dimensional entity for the concept of a car. The seating position of 95% of humans is used for the arrangement of the seat, pedals, and steering wheel and for the position of the front and rear axles.

data determine the size of seats, the passenger space, the relative position of seat, steering wheel, pedals, and so on. Figure 3 shows a side view of the most important dimensions for determining the areas for passengers and driver in a passenger car. The quality of vehicle driving improves with the increasing physical and psychological well-being of the driver; therefore, other limitations to be taken into account are the endurance of vibrations, comfort in the ambient temperature and humidity, and the level of noise.

III. THE VEHICLE

Technically speaking, the vehicle is a container on wheels that carries all the necessary components for operation, such as axles, wheels, suspension, fuel, engine, transmission, and drive shafts, as well as the necessary operating and control systems; it is designed to meet the requirements of the load to be transported (i.e., people or goods) and other objectives of transportation (efficiency, safety, comfort, style, and price). Figure 4 shows examples of the three families of motor vehicles (namely, passenger cars, commercial vehicles, and buses) and gives an impression of the large number of variations. Other motor vehicles, such as motorcycles and work vehicles (e.g., tractors), are not dealt with in this article. For all the differences among vehicles with regard to use and functioning in traffic, many of the tasks of automobile technology are the same for all types of vehicles, especially in the fields of handling dynamics, vehicle construction, equipment, transmission, vehicle control, and passive safety (crashworthiness).

A. Handling Dynamics

Handling dynamics describes the generation and interaction of all the forces that occur through acceleration (deceleration) of the total and partial masses of the moving vehicle. The control of these forces is the prerequisite for safe handling characteristics.

For reasons of stability, automobiles, with very few exceptions, stand on at least four wheels arranged at the

TABLE I Body Measurements (mm) for North America

Measurement	Men			Women		
	Percentile			Percentile		
	5	50	95	5	50	95
Stature	1670	1790	1900	1540	1650	1760
Sitting height	890	930	990	820	880	930
Eye height, sitting	760	810	860	700	760	810
Forward reach (fingertips)	790	850	910	730	800	870
Shoulder breadth (bideltoid)	420	460	500	370	400	430
Shoulder breadth (biacromial)	365	395	430	330	360	385
Hip breadth (standing)	305	340	365	310	350	405
Knee height	505	550	600	460	500	550
Lower leg length (popliteal height)	405	445	500	360	410	450
Elbow-grip length	325	360	385	295	330	355
Buttock-knee length	550	600	660	520	570	630
Buttock-heel length	990	1080	1180	940	1030	1130
Hip breadth (sitting)	310	350	400	310	370	430
Hand length	175	190	205	160	170	190
Hand breadth	80	85	90	70	75	85
Foot length	250	265	285	220	245	260
Head circumference	550	575	600	520	545	575
Head length	185	195	205	165	180	195
Head breadth	145	155	165	135	145	155

From International Data on Anthropometry, International Labour Office, Geneva, Switzerland.

corners of a rectangle. Passenger vehicles therefore generally have two axles, and buses and commercial vehicles, depending on their weight, have two or more axles. On passenger cars all the wheels and on buses and commercial vehicles the steered front wheels, have single tires. On commercial vehicles and on buses, owing to the higher load, dual tires are often used on the rear axles. The wheels are fitted with inflated tires whose spring-damper qualities have a great influence on the whole vibration system and whose friction characteristics against the road surface co-determine the limits of the guiding forces of the vehicle. All the foces for carrying, propelling, braking, and steering are transmitted to the ground via the wheels (Fig. 5). The connection between wheel and ground is positive only for load-support forces and nonpositive for longitudinal and lateral forces. From this is derived great maneuverability, but it is also the basis for problems caused by the limits of frictional contact.

TABLE II Stature of Human Beings (mm) from Selected Regions

	North America	Northern Europe	France	Iberian Peninsula	Japan
Women (5%)	1540	1580	1530	1510	1510
Men (95%)	1900	1910	1890	1830	1820

From International Data on Anthropometry, International Labour Office, Geneva, Switzerland.

The vehicle has six degrees of freedom—three translatory movements in the longitudinal, lateral, and vertical directions and three rotational movements about the longitudinal (roll), the lateral (pitch), and the vertical (yaw) axles. Of these six possible movements, only two are controlled by the driver (as long as he remains within the physical limits of the tire grip): movement in the longitudinal direction by pressing the accelerator or by braking and turning around the vertical axis by steering, which leads to the lateral movement of the vehicle. The movements in the other degrees of freedom are determined by the road profile and the vehicle parameters, because the body and wheels are coupled through the kinematics of the suspension-damper system. Three dynamic requirements must be satisfied:

1. Permanent positive wheel load, as the guiding forces F_H are limited by the product of the vertical wheel load (normal force) F_N and the traction coefficient μ_w between tire and road surface:

$$F_H = F_N \mu_w \text{ (N)}$$

Both factors are variable over wide limits.

A

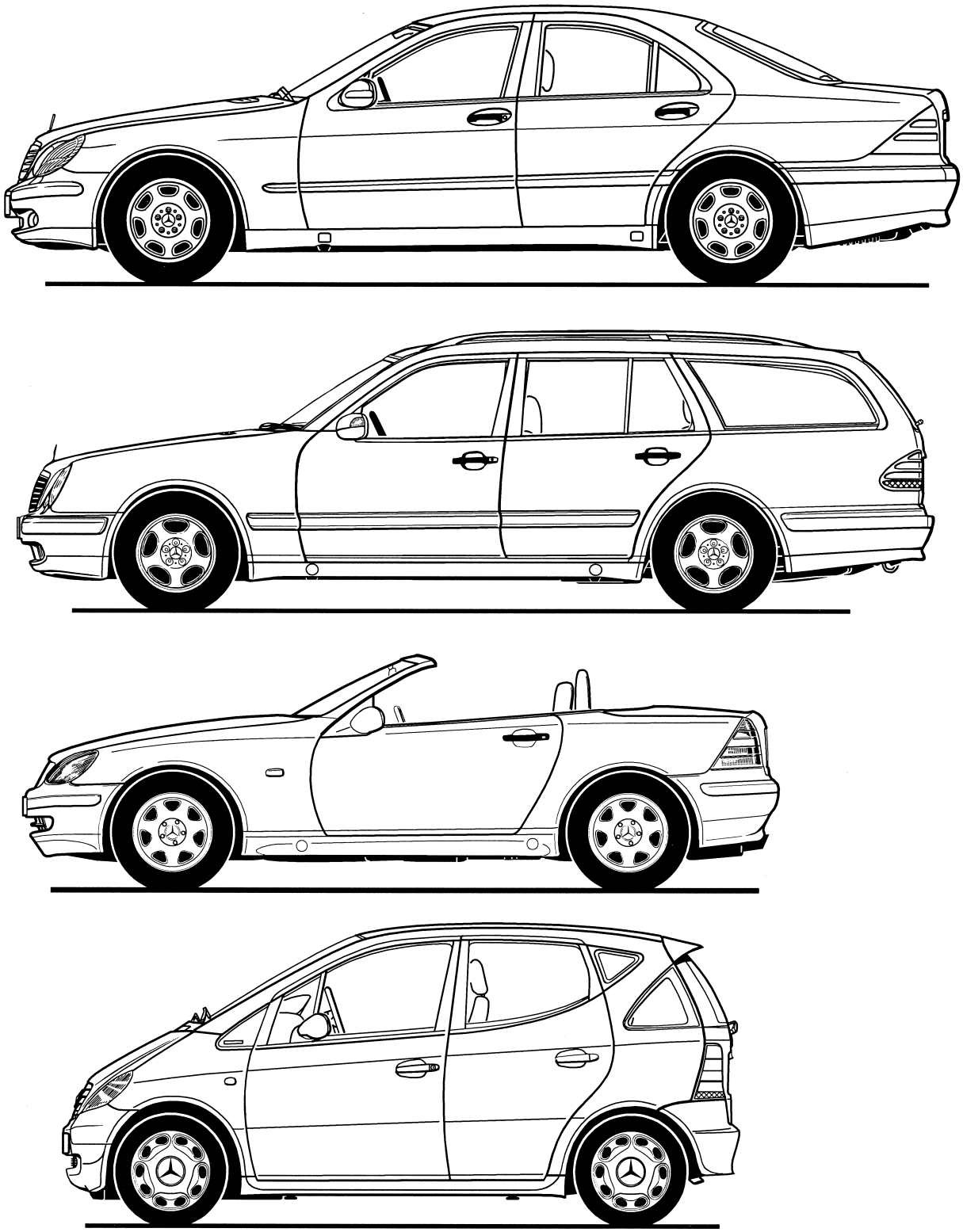


FIGURE 4 (A) Types of passenger cars (Mercedes-Benz); top to bottom: limousine, station wagon, roadster, compact car. (B) Types of commercial vehicles (Mercedes-Benz); top to bottom; delivery van, truck, semitrailer truck. (C) Types of buses (Mercedes-Benz); top to bottom, articulated bus, city bus, coach, small bus, minibus.

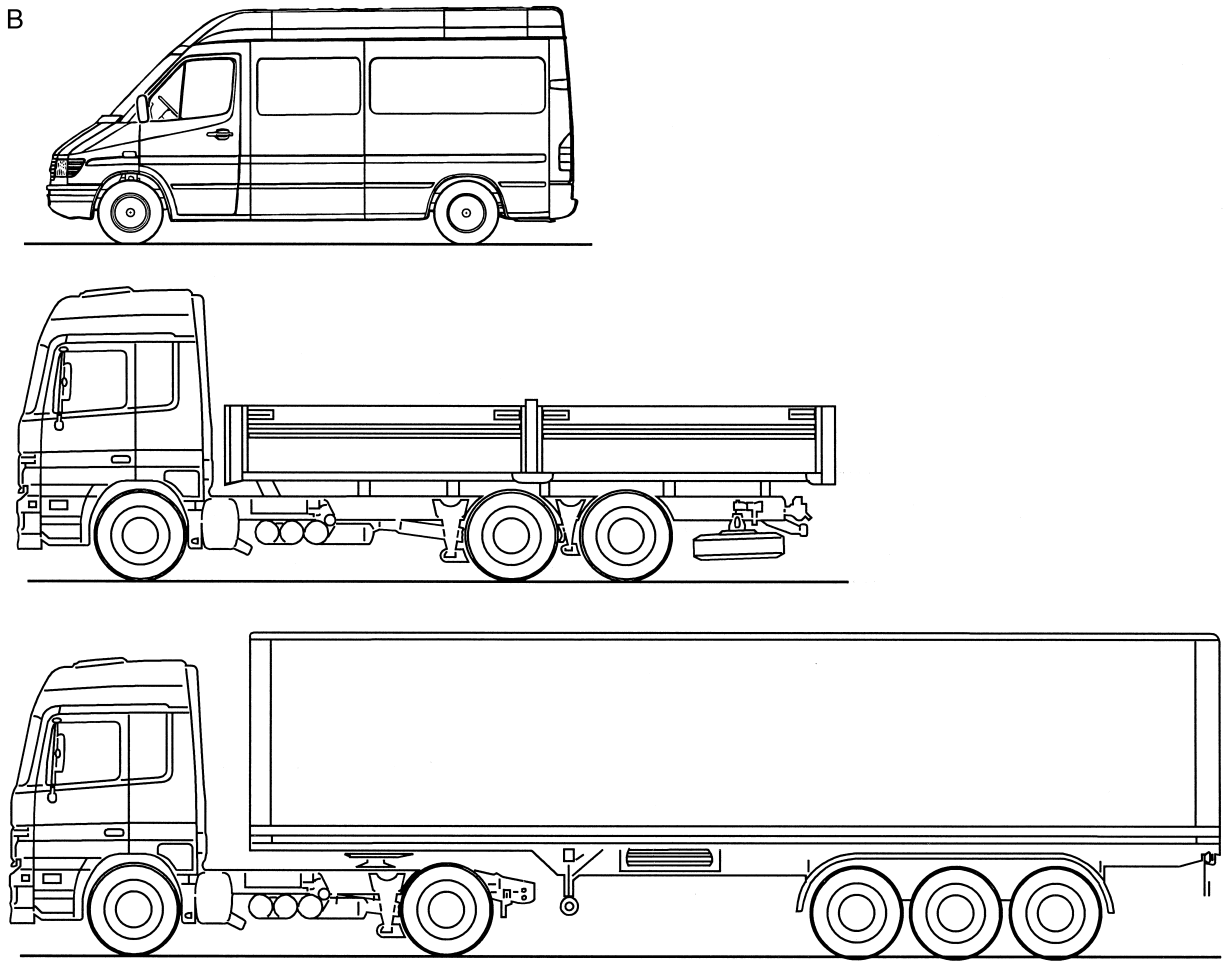


FIGURE 4 (*continued*)

2. "Floating" of body independent of the ground profile, or at least in forms of vibrations whose amplitude, frequency, and damping the passengers (freight) can stand.
3. Self-stabilizing steering characteristics.

1. Road Grip

The vertical load F_N on each wheel is determined statically by the total weight of the vehicle (passenger car in the range 1–3 tons, commercial vehicles up to 50 tons), the number of wheels (car, 4; commercial vehicle, 4–10), and the position of the center of gravity of the vehicle. Superimposed on this static load distribution are three types of additional dynamic forces, which occur during acceleration or braking, during cornering, and through oscillations of the body. During acceleration, the rear axle is loaded further and the front axle is correspondingly relieved, while during braking the front wheels take up more load and the rear wheels are relieved. During cornering, the centrifugal forces acting at the center of gravity in addition to the weight are supported on the ground by an additional load on the wheels at the outside of the curve and a corresponding reduction on the wheels at the inside of the curve. Because in principle only one axle is necessary to support the centrifugal forces, the way the division between the two axles occurs provides additional freedom for influencing the self-steering performance during cornering. Both types of additional forces depend on the height of the center of gravity (Fig. 6), so that it is kept as low as possible, though this is limited by requirements for sufficient passenger or cargo space.

Finally, vibrations of the body and unevenness of the road can increase or decrease the load on the wheels. Since the maximum ground adhesion is achieved with an even distribution of the total weight among the load-carrying

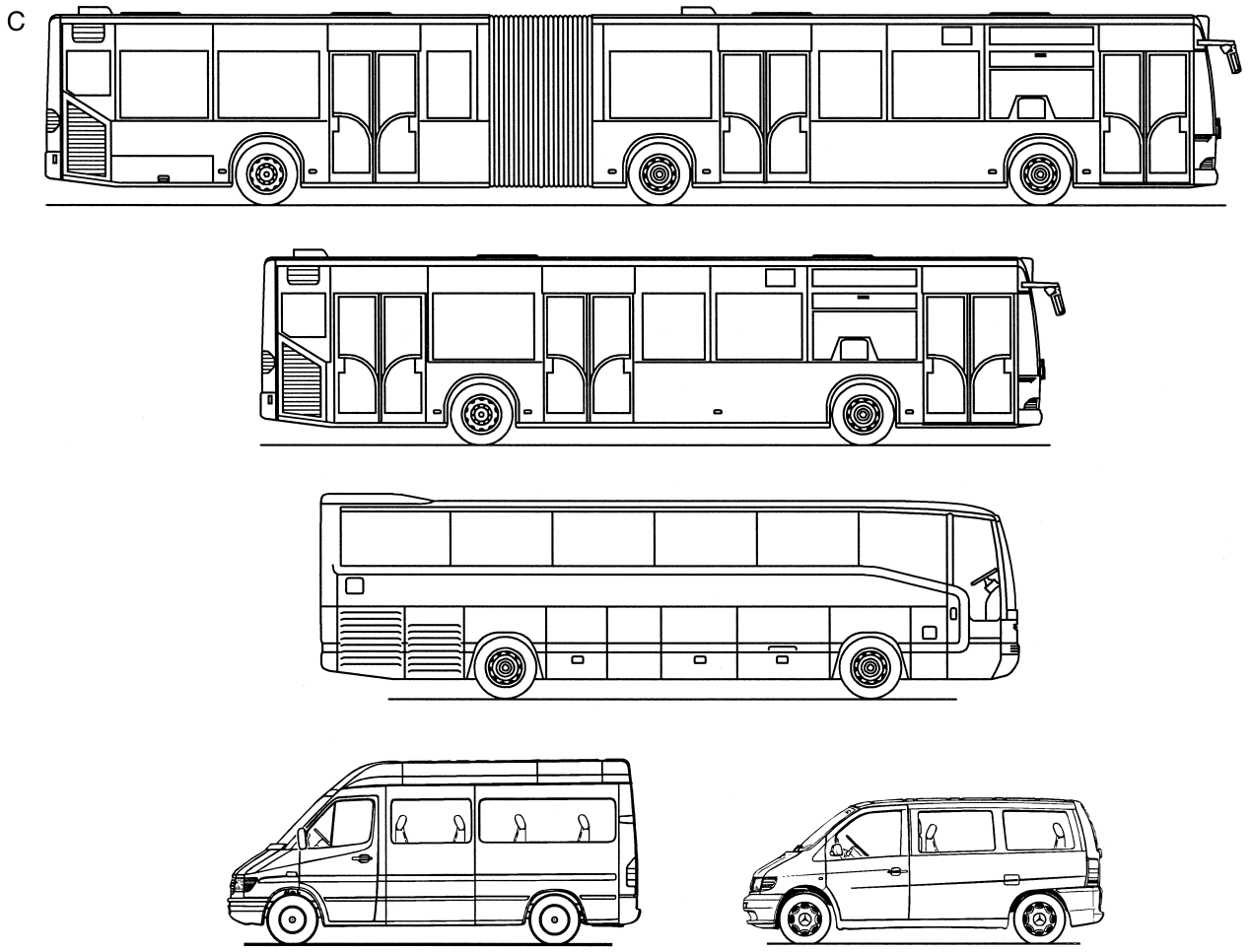


FIGURE 4 (*continued*)

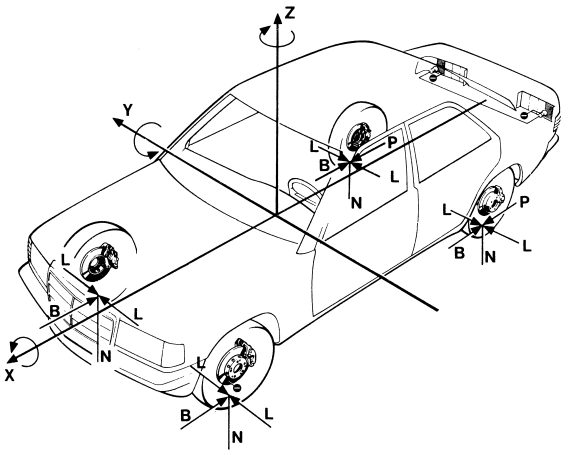


FIGURE 5 Degrees of freedom for movement and forces transmitted by wheels (tires). *X*, *Y*, *Z* are coordinates with the center of gravity as the zero point; *N*, normal (vertical) force; *B*, braking force; *P*, propulsion force; *L*, lateral force. *B*, *P*, and *L* are working in the horizontal plane ($\Sigma N = F_N$; $\Sigma B,\ P,\ L = F_H$).

wheels, the dynamic forces, which create additional load fluctuations, must be kept as low as possible.

The maintenance of contact between the road and the wheels involves the whole vibratory system, including the pneumatic tire, wheel or axle against the frame, and the frame against the body, with their frequencies and attenuations. The following are technical devices for achieving good road grip: low-pressure tires, small unsprung masses of the wheel-axle system (e.g., independent wheel suspension), long spring displacements even under load, low-friction springs, and adequate damping.

The traction coefficient μ_w, the second factor that determines road adhesion, depends on many parameters (Fig. 7). The following apply to the rolling tire:

1. The total force that a tire transmits is divided into circumferential and lateral forces roughly according to the Pythagorean theorem.
2. Circumferential force is transmitted only when there is slip between tire and road surface.

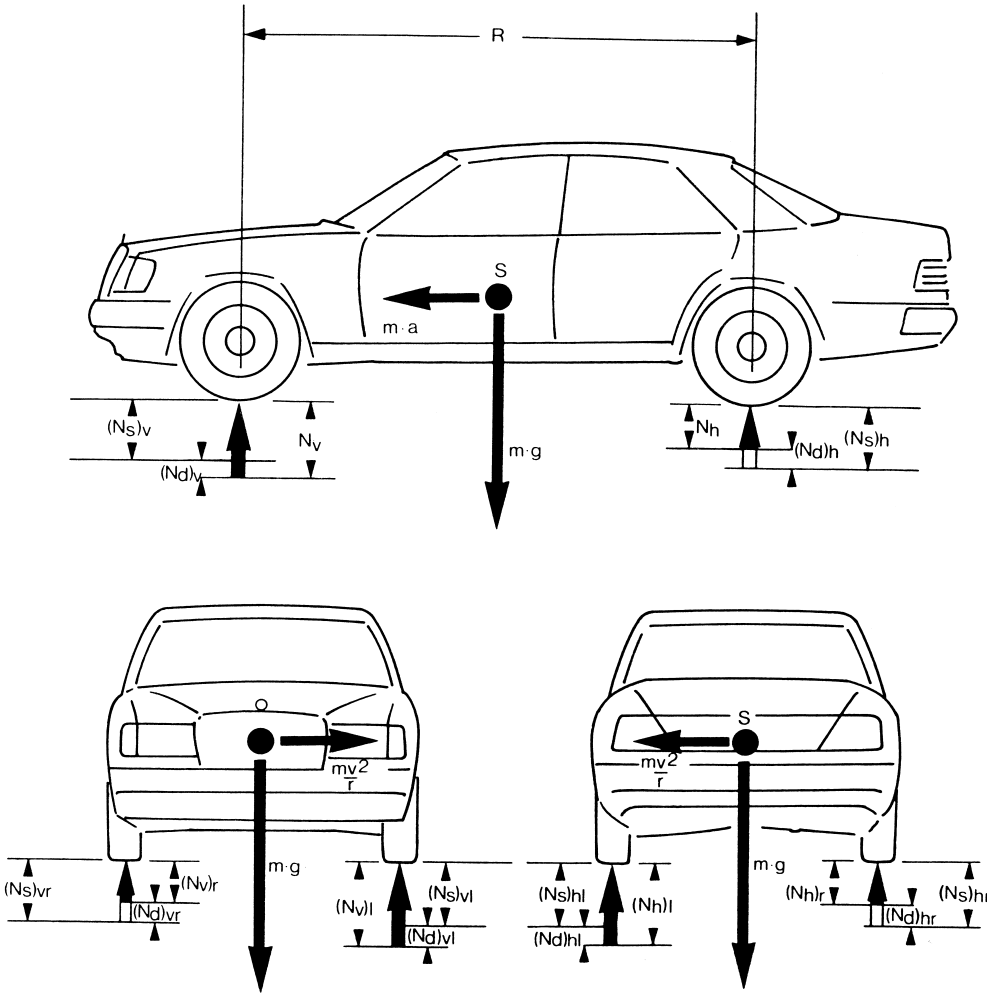


FIGURE 6 Effect of dynamic forces on the normal load of the wheels. *S*, center of gravity; *R*, Wheel base. (Top) Braking (deceleration): *mg*, weight; *ma*, forward force due to deceleration (both are effective in the center of gravity), $(N_S)_v$, static load, front wheel; $(N_S)_h$, static load, rear wheel; $(N_d)_v$, dynamic load, front wheel; $(N_d)_h$, dynamic load, rear wheel; $N_v = (N_s + N_d)_v$, total load, front wheel; $N_h = (N_s - N_d)_h$, total load, rear wheel. (Bottom) Cornering: mv^2/r, centrifugal force when cornering with speed v and bend radius r; $(N_s)_{vr}$, static load, front wheel right; $(N_s)_{vl}$, static load, front wheel left; $(N_s)_{hr}$, static load, rear wheel right $(N_s)_{hl}$, static load, rear wheel left; $(N_d)_{vr}$, dynamic load, front wheel right; $(N_d)_{vl}$, dynamic load, front wheel left; $(N_d)_{hr}$, dynamic load, rear wheel right; $(N_d)_{hl}$, dynamic load, rear wheel left; $(N_v)_r = (N_s - N_d)_{vr}$, total load, front wheel right; $(N_v)_1 = (N_s + N_d)_{vl}$, total load, front wheel left; $(N_h)_r = (N_s - N_d)_{hr}$, total load, rear wheel right; $(N_h)_1 = (N_s + N_d)_{hl}$, total load, rear wheel left.

3. Lateral force is generated only if the wheel (tire) is running at an angle to its plane (slip angle), and it tends toward a limiting value with an increase in slip angle.
4. The circumferential force increases rapidly with increasing slip and reaches a maximum at ~15% slip, after which it decreases.
5. The lateral guidance force is at a maximum at zero circumferential slip and reduces roughly linearly to almost zero at 100% slip. A locked wheel has almost no lateral guidance.

Superimposed on these rules are the effect of the tires, such as their construction (height/width ratio and whether they are radial- or cross-ply), rubber mixture, type and depth of tread, and tire pressure, as well as the larger effects of the road surface, such as type and roughness, wetness and cleanliness, and degree of water removal with rain, snow, and ice. The grip-force limits can thus vary over a very wide range of 1–10, although in practice they are mainly utilized only over less than 20%. The surprise (and incorrect behavior) is then all the greater under special driving conditions with small values

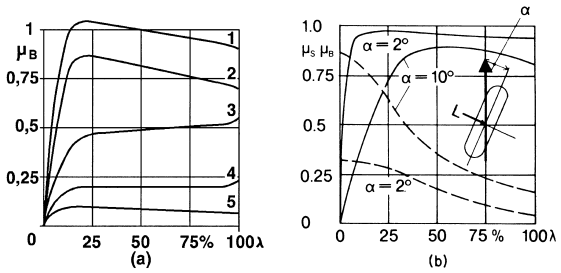


FIGURE 7 (a) Circumferential force coefficient μ_B as a function of slip λ between tire and road (slip angle $\alpha = 0$). 1, radial-ply tire on dry asphalt; 2, radial-ply tire on wet asphalt; 3, radial-ply tire on gravel; 4, radial-ply tire on powder snow; 5, radial-ply tire on glaze. (b) Lateral force coefficient μ_S (– –) and circumferential force coefficient μ_B (—) as function of slip λ and slip angle α between tire and road. L = lateral force; traction coefficient, $\sqrt{\mu_w} \approx \mu_B^2 + \mu_S^2$.

of the traction coefficient, such as icy roads and aquaplaning.

2. Floating State

The "floating state" of the passenger cell relative to the wheels, which must follow every undulation of the ground, is striven for by the following:

1. Large movements of the springs and frictionless springing (helical, torsion bar springs or air springing)
2. Frictionless damping (hydraulic shock absorbers, gas-pressure spring struts with large stroke)
3. Stiffening of suspension with alternate wheel motion (torsion stabilizer, composite springs)
4. Constant spring displacements and constant vibration frequency with greatly differing vehicle loads (level adjustment with additional springs, hydropneumatic or air suspension)

3. Self-Steering

The self-steering characteristics of a vehicle describe its directional stability when disturbing forces occur and its reaction to steering motions. Two types of disturbances can cause deviations from the desired course in a driving vehicle: (1) external effects due to the air, unevenness of the ground, and road gradients in both the longitudinal and lateral directions; and (2) internal forces caused by movements of the wheels and the body when these alter the position of the wheel planes relative to the vehicle.

Of the external forces, the effect of the aerodynamic forces can be reduced primarily by suitable form. To counter the second type of disturbance, it is necessary to eliminate the effect of all dynamic forces and the movements of the axles and the body on the straight-line motion or the track holding during cornering by modifying the axle kinematics and the types of suspension and damping.

B. Vehicle Design

Vehicle design includes the chassis and the cell as well as the large number of units and components fitted to the vehicle that make it possible for the vehicle to operate. In this section, only the most important of these are mentioned (the power train is discussed in Section IV), but each component, however small and unimportant it may seem to be, requires the same intensive development as the large units; in an automobile, nothing is unimportant. Longevity, reliability, and easy serviceability are major targets of design.

1. Frame and Cell

There are two basically different methods of motor vehicle construction—namely, the separate assembly of frame and body and the integral method, in which the body, as a load-bearing shell, also takes on the function of the frame. Mixed forms of both methods also exist.

With chassis construction, the chassis is the backbone of the vehicle from which the axles are suspended and on which the body is placed; it carries the engine, gearing, and drive shafts. This is still the predominant method for commercial vehicles (Fig. 8). Here, too, the cab and load compartment are often separate so that the design, such as platform, box type, or interchangeable body, can be adapted to the required purpose (Fig. 4B). Depending on the application, the frame can be torsion resistant or flexible.

However, the integral method of construction, in which the chassis and body are combined into a single load-carrying shell, is well established for automobiles (Fig. 9) and for many buses. These integral vehicle shells are welded together from many (often complicated) sheet-metal parts formed in large presses. The advantage is increased stiffness combined with lower weight, although insulation of the passenger cell from road shocks and noise is more difficult. For this reason, parts of axles are sometimes suspended from subframes. External forces from the road are applied to the frame or body only at the mounting points of the suspension. The construction of the doors with hinges and locks, often with central locking (electric or pneumatic) which also fulfills a load-carrying function, and the sealing of fixed and movable windows (which require lifting and closing

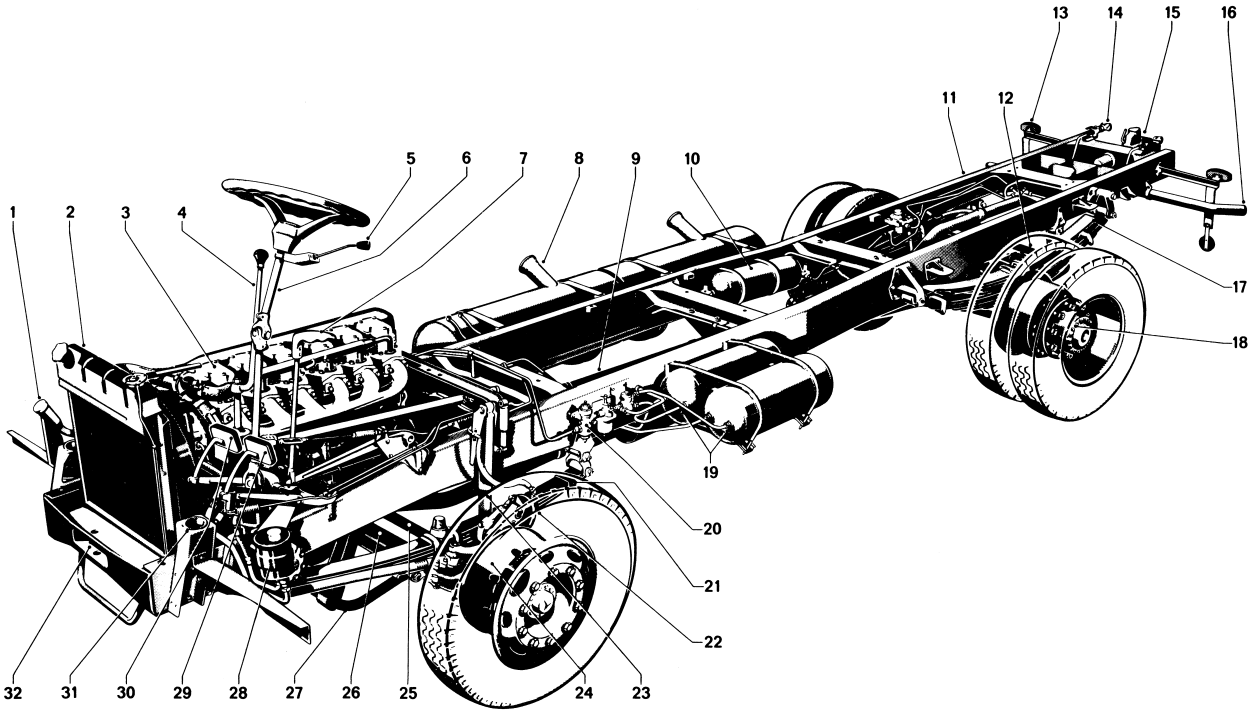


FIGURE 8 Commercial vehicle frame with assemblies (Mercedes-Benz). (1) oil filling tube; (2) radiator; (3) engine; (4) gearshift lever; (5) direction indicator; (6) steering; (7) hand brake; (8) fuel tank; (9) universal shaft; (10) accumulator, compressed air; (11) frame; (12) stabilizer rear axle; (13) tail lamp; (14) air connection, trailer; (15) coupling trailer; (16) exhaust pipe; (17) rear suspension; (18) rear axle; (19) accumulator, compressed air; (20) air pressure control valve; (21) front suspension; (22) break servo; (23) shock absorber; (24) wheel brake drum; (25) tie rod; (26) front axle; (27) stabilizer front axle; (28) oil tank, power steering; (29) clutch pedal; (30) brake pedal; (31) mounting pads for driver cab; (32) coupling front.

mechanisms, often electrically operated) require extraordinary solutions owing to the high demand made on them. The optimum design of the frame, and especially the correct forming of the integral shell, can be done successfully only with the aid of the finite-element method calculation, in which a component is assumed to consist of a large number of individual elements whose effect on one another at the intersections can be calculated (Fig. 10). This method is also used for the design of many other components, such as housings, gears, shafts, and frames. Severe and unusual demands are made on the chassis and body construction by the conditions of passive safety (Section VI).

2. Axle

Axle is the generic term for all elements that connect the wheel with the frame or the body. There are two groups: the rigid axle and independent wheel suspension.

The rigid axle (Fig. 11), which is still predominant in commercial vehicles, buses, transporters, and off-road vehicles, consists of a rigid support with connections at either end for attaching wheels with brakes. On driven axles, the axle transmission is mounted on the axle or formed as part of the axle, from which support tubes lead to the wheel-bearing housings. The body of the axle with steered wheels carries at either end the suspension for the axle steering pivot (king pin). The rigid axle is guided either by leaf springs directly or by individual steering arms, especially with helical springs or air suspension (Watts' linkage, Panhard rod, links, etc.) and connected with the frame. The advantage of the rigid axle design is its simple construction, especially for driven axles, and the tidy parallel guidance of the wheel; however, the vertical travel of one wheel influences camber and toe-in of the other one. The disadvantages are the large, unsprung masses (which make the maintenance of contact between the road and the wheels more difficult) and the large space requirement of the floating axle.

A great number of different designs for independent suspension of the wheels include wishbones, longitudinal control arms, torque struts, tension rods, and spring struts

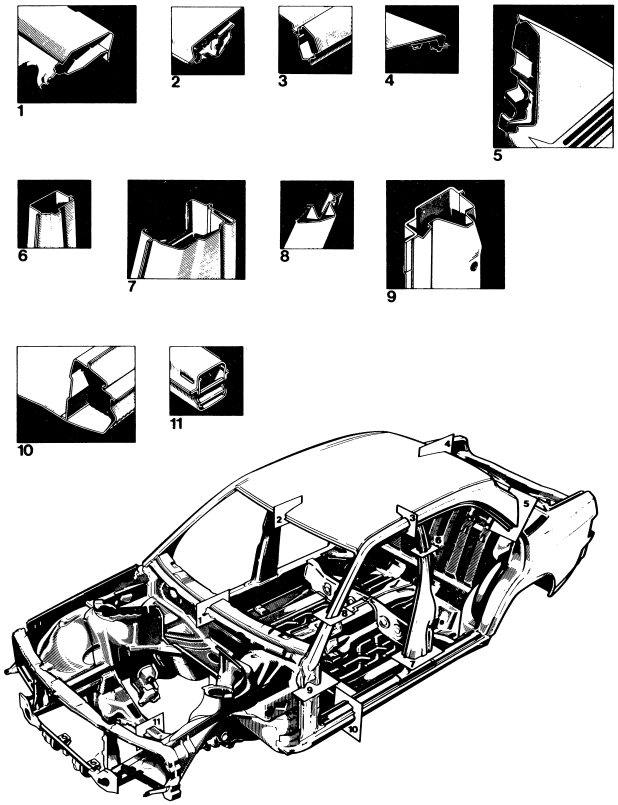


FIGURE 9 Integral body for passenger car, shell.

(Figs. 12 and 13). As with independent wheel suspension, the axle transmission and the steering gear are suspended from the chassis and do not follow the movement of the wheels; these gears have to be connected to the wheels through shafts and joints. There are also combinations or rigid axle and independent suspension (De Dion axle). With this multiplicity of variations it is possible to give vertical wheel movements any desired characteristic, although the effect of the desired flexible guides and the cell must be taken into account. Usually a compromise must be sought between simplicity of design and ideal suspension. The advantage of the small unsprung masses of the independent suspension must be compared with

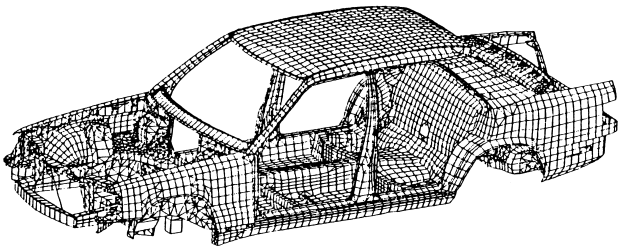


FIGURE 10 Body replacement structure for stiffness and strength calculations by the finite element method; example passenger car body.

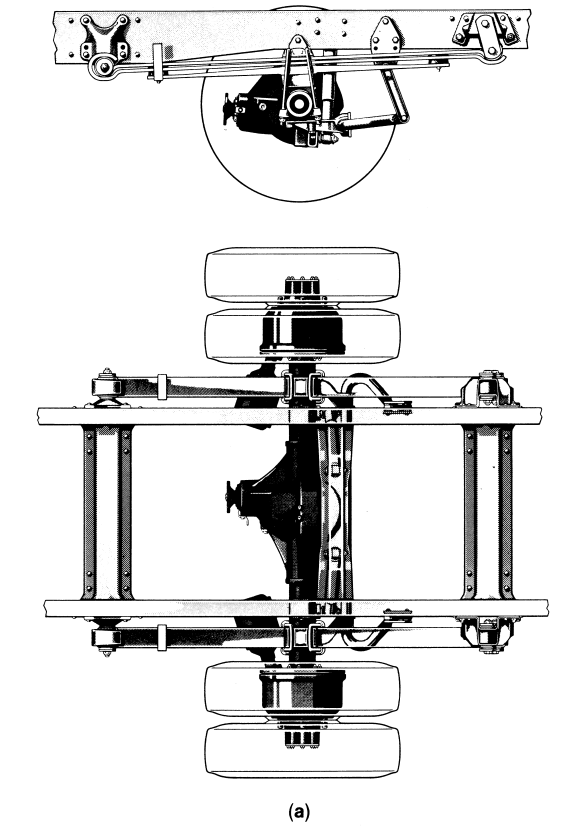


FIGURE 11 (a) Rigid rear axle for commercial vehicle (Mercedes-Benz) with leaf spring suspension, torsion bar, and telescopic shock absorber. (b) Rigid front axle for commercial vehicle (Mercedes-Benz) and cross-section of the wheel. Front axle, designed as a stub axle, has needle bearing king pins, balanced brake drums, and an air-actuated brake with wedge operation.

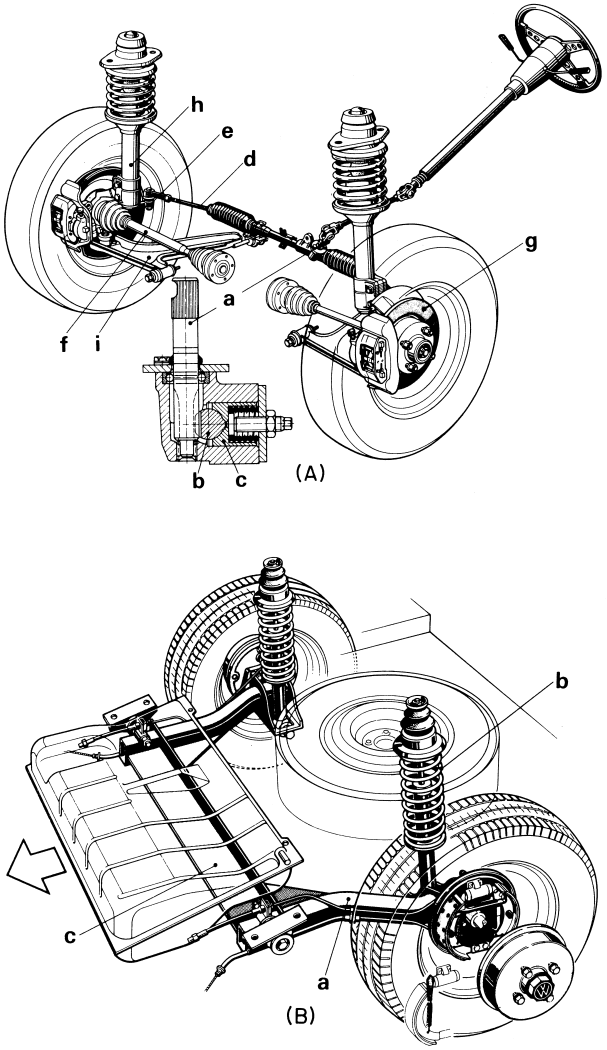


FIGURE 12 (A) Independent wheel suspension for front axle of passenger car with front-wheel drive (VW). *a*, Steering gear, pinion; *b*, steering gear, rack; *c*, pressure sleeve; *d*, tie rod; *e*, steering lever; *f*, drive shaft; *g*, disk brake; *h*, McPherson spring strut; *i*, wishbone. (B) Independent wheel suspension for rear axle of passenger car with front-wheel drive (VW). *a*, Trailing arm; *b*, spring strut; *c*, traverse beam as quasi-torsion bar.

the increased cost of wheel guidance, steering, and drive elements.

3. Steering

The vehicle is steered by turning of the front wheels only. Cornering stability can be increased by countersteering of the rear wheels only a few degrees. The steering kinematics are designed to ensure that all the wheels are cornering around the same center but have to take account of deviations from the geometric movements for large slip angles (at large side force). The types of steering gear include rack and pinion (Fig. 14A and B), worm and nut steering, or a gear unit with recirculating balls between worm and nut. To reduce the steering, effort in medium and large cars, power steering with hydraulic power assistance is fitted.

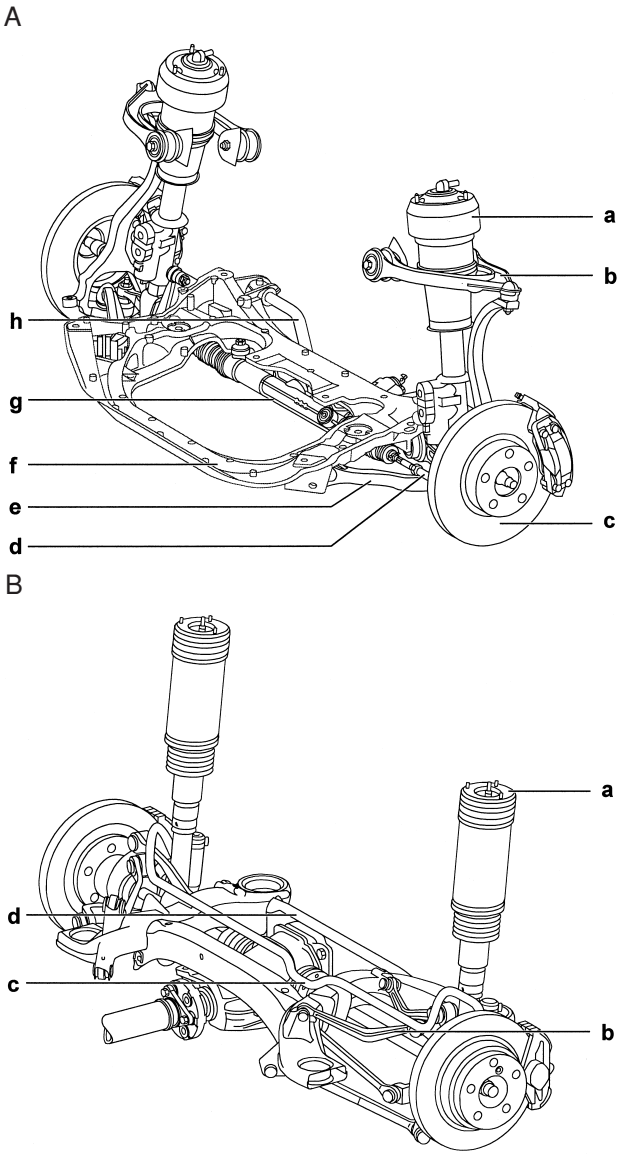


FIGURE 13 (A) Independent wheel suspension for front axle of motorcars with rear-wheel drive (Mercedes-Benz). *a*, Air suspension strud; *b*, upper wishbone; *c*, disc brake; *d*, tie rod; *e*, lower wishbone; *f*, chassis subframe; *g*, rack-and-pinion steering; *h*, stabilizer (torsion bar). (B) Independent wheel suspension for rear axle of cars with rear-wheel drive (Mercedes-Benz); multilink suspension, where the movement of the wheel carrier is guided by five links flexibly mounted on the axle support. *a*, Air suspension strud; *b*, one of five links; *c*, stabilizer (torsion bar); *d*, chassis subframe.

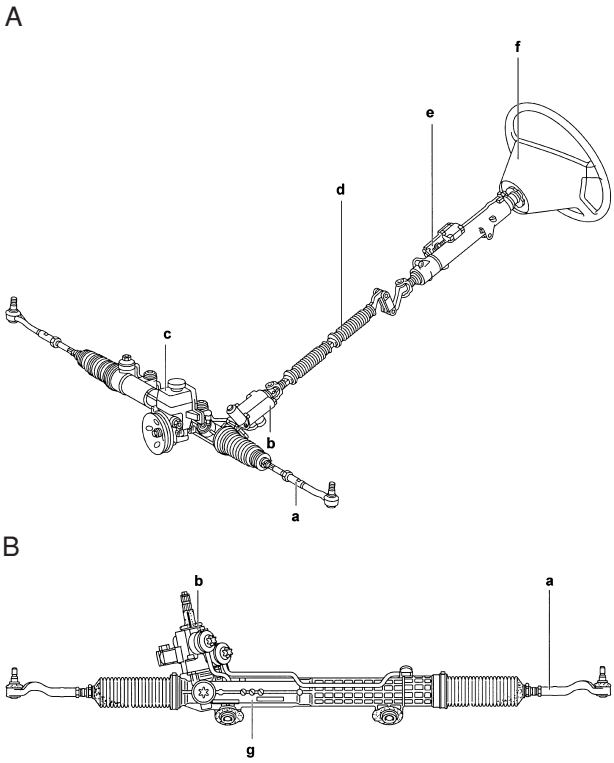


FIGURE 14 (A) Power-assisted steering (Mercedes-Benz). (B) Rack-and-pinion steering. *a*, Tie rod; *b*, steering gear with internal steering shock absorber; *c*, power-steering pump and reservoir for power-steering fluid; *d*, collapsible steering column; *e*, casing tube; *f*, steering wheel; *g*, steering rack (not visible).

4. Braking System

The braking system consists of friction brakes in all wheels and, for safety, of several brake circuits (Fig. 15). On most commercial vehicles and on some smaller passenger vehicles, only the rear-axle drum brakes are fitted: however, in passenger vehicles disk brakes are mainly being used. These can stand a greater overload and exhibit less fading but need much higher pressure. The wheel–brake cylinders, which are mounted in or near the wheels, are operated by the master cylinder of the brake pedal through a hydrostatic linkage (special brake fluid). For power brakes, which have been introduced in almost all vehicles, the foot power is usually assisted by a vacuum intensifier (on medium and heavy commercial vehicles, mainly compressed-air brakes). The proportioning of the brake force between the wheels must take into account the different static loads as well as the dynamic increase and decrease of the wheel loads during the braking process. The rear wheels must never lock first, because the vehicle then becomes unstable. Today, antilock brake systems, which prevent locking by an electronically controlled intervention in the brake circuit, are being used in passenger and in commercial vehicles (Fig. 16). The speeds of the individual wheels are measured and processed electronically. As soon as one of the wheels threatens to exceed a certain slip, the brake pressure for this wheel is correspondingly reduced. In this way not only is it possible to reduce the stopping distance in many cases, but also the lateral guidance force, so important for stability and swerving maneuvers, is fully maintained. In commercial vehicles and buses, often a third brake for continuous operation is fitted in the form of an engine exhaust gas throttle or a hydraulic (hydrodynamic brake) or an electric (eddy current brake) retarder. In addition, all vehicles have a parking brake (food, hand, or spring brake).

5. Electrical System

The electrical installation, which initially included only the ignition, is growing very quickly. This is due to new possibilities of electronic data acquisition and processing. The electrical installation consists of the following basic parts:

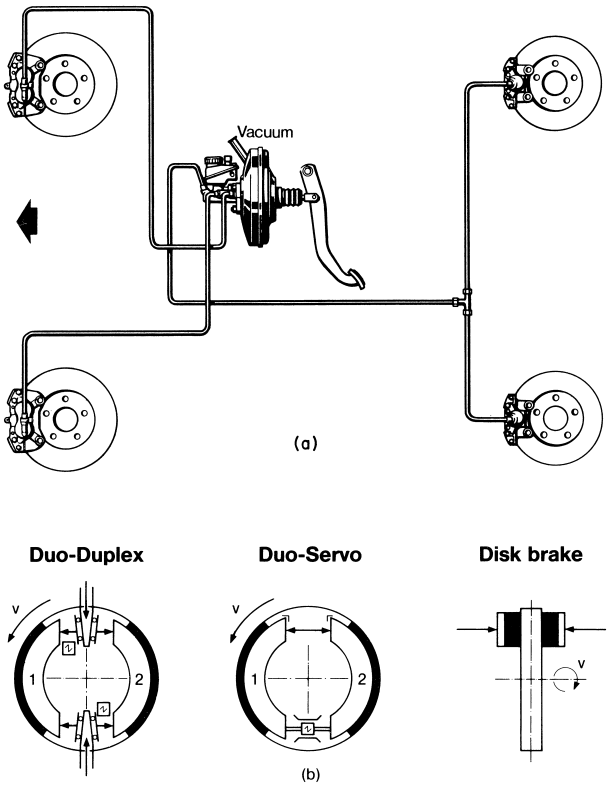


FIGURE 15 (a) Dual-circuit brake system for automobile, vacuum assisted. (b) Diagrammatic illustration of brake design. *v*, Circumferential speed; 1, 2, both leading shoes. $C^* =$ braking force/applied force, with a friction coefficient of the brake lining of 0.4. Duo-duplex, $C^* = 3$; duo-servo, $C^* = 5$; disk brake, $C^* = 0.8$.

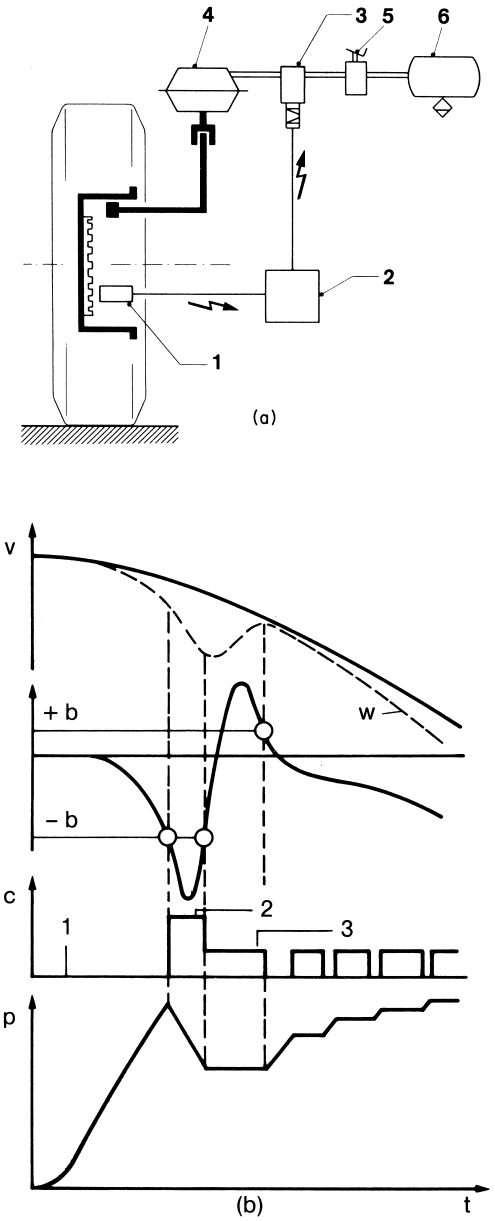


FIGURE 16 (a) Antilock system for commercial vehicle, schematic. 1, Wheel-speed sensor; 2, electronic control unit; 3, solenoid valve; 4, servo unit; 5, brake pedal; 6, compressed air accumulator. (b) Antilock working diagram. *t*, Time; *v*, vehicle speed; *w*, wheel speed (dashed line); $+b$, acceleration of wheel; $-b$, deceleration of wheel; *c*, control impulses; 1, increase pressure; 2, drop pressure; 3, maintain pressure; *p*, air pressure in servo unit.

1. The generator, driven by the engine, which supplies rectified three-phase current direct ot the user or to storage (e.g., battery).
2. The battery (mainly lead–acid battery), which stores surplus energy, supplies electrical energy for strating the engine, and feeds current into the electrical system when the engine/generator speed is low.
3. The extensive cable harness, which makes the connections between all the elements of the electrical installation (i.e., control units, switches, sensors, actuators, etc.). In order to reduce the cable harness in cars with a large number of electric components, data networking is increasingly used (Fig. 17A). In data networks, the bulk of the cable harness is divided into wires for voltage supply and wires for data transfer (information which is digitally coded mostly by means of different voltage levels). The data are distributed via a tree structure with microprocessors at the handover point to the end-user. The cable harness normally consists of copper wires; however, due to data networking, copper wires have been replaced in some fields (e.g., communications systems) by optical fibers in which the information is transmitted by light impulses instead of electicity.
4. The operating devices (e.g., switches).
5. The end users (e.g., the starter motor, ignition, internal and external lighting, headlamps, direction indicators, brake lights, windshield wiper and washer, ventilation fan, control motors for mirrors, seats, window lifters, central locking, window, seat and auxiliary heating, sensors and indicators, horn, and communication system).
6. The electronic systems, which have assumed very strong significance in recent years (see below). Elecrtonic systems include engine and transmission and suspension control, antilock brake system, antislip control, electronic stability program, air bag, belt tensioner, etc.

Progress in electronic and information technology as well as numerous complex functions that have to be carried out while operating a vehicle together with the high output of automobile production will lead to large markets for electronic and information technology in automobiles and traffic. The continous preferential treatment of electronic systems in control operations will at the same time lead to a preferential treatment of electric actuators in order to avoid a conversion of electric signals in a different technique (e.g., hydraulic).

6. Electronic Systems

There are four fields in which the new techniques of electronic data acquisition and processing are used:

1. Control functionalities required for the operation of the vehicle which were partly carried out mechanically, hydraulically, or pneumatically in the past. Examples are overall engine manangement systems to cut back on fuel consumption or pollutant emissions as well as transmission control in order to

choose the optimal ratio of transmission or to improve the shifting operation in the automatic transmission.

2. Control functionalities concerning the movement of the vehicle on the road which also improve the active and passive safety by recognizing when the driver reacts too late or in an inappropriate fashion and correcting his steering and braking maneuvers in good time. Examples of the combination of axles, steering, and brakes with electronic data processing follow. The *antilock brake system* (ABS; Section 4) is sometimes backed up by *brake assist*, which shortens the stopping distance of a vehicle in an emergency braking situation. The brake assist recognizes a fear-induced reaction from the speed at which the driver depresses the brake pedal. It then proceeds to develop maximum brake power boost from the start of braking. The *acceleration skid control system* (ASR) prevents the drive wheels from slipping. ASR is an electronically controlled intervention of the engine power output and the brakes which is being used increasingly in both passenger cars and commercial vehicles. The *electronic stability program* (ESP) (Fig. 17B) drastically reduces skidding and sliding tendencies in the moving vehicle. The system detects potential dangerous situations on the basis of different sensor signals (e.g., rotational speed of the wheel, yaw angle, and steering angle) faster than even the most experienced driver can do. ESP than intervenes with high precision and stabilizes the vehicle with exactly measured activation of the brakes and/or reduction of the engine power output. Electronic control of the brakes and engine power helps in utilizing all the possible friction forces between the wheels and the road, but only within their physical limits. The *electronic distance control* regulates the distance to the vehicle in front via radar or infrared sensors, intervenes in the engine management, and, if necessary, activates the brakes. Activation of the airbag and belt tensioner is controlled by a *deceleration sensor*, which is rigidly connected to the car body in a central location (Section VI).

3. Information and communication transfer between the vehicle, the driver, and the outside world. For

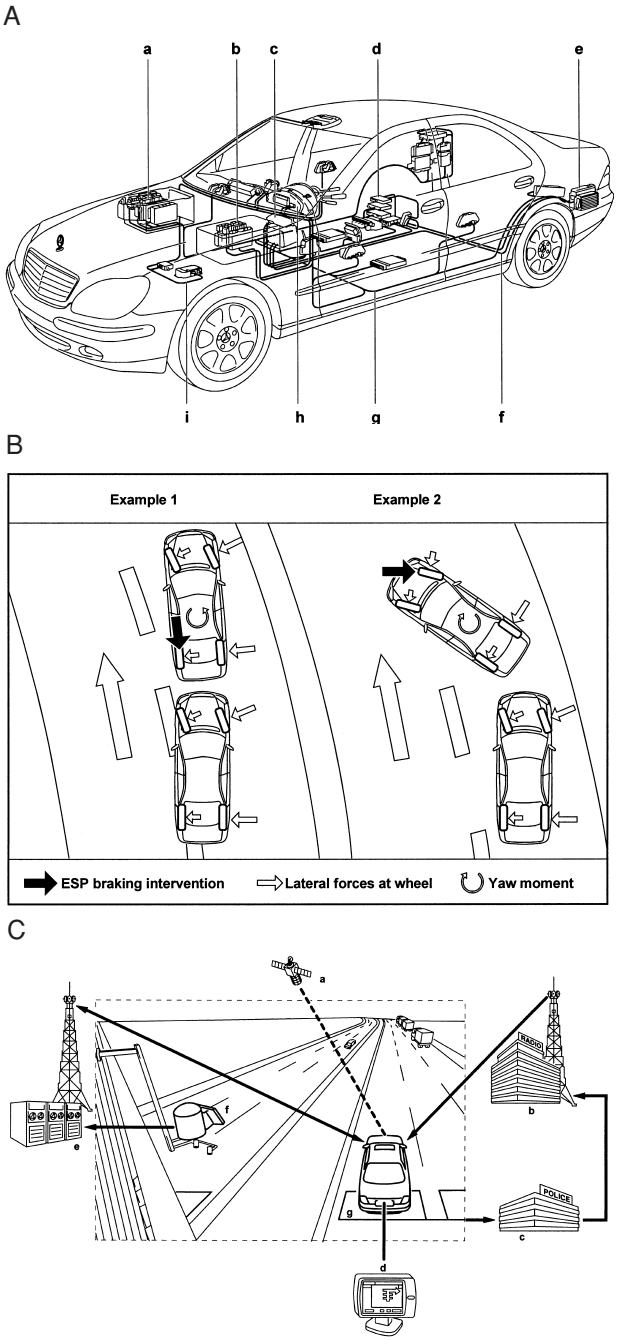


FIGURE 17 (A) Data communication in electronic networks (Mercedes-Benz). *a*, Engine, transmission, and electronic stability program (ESP) control units; *b*, air-suspension control unit; *c*, navigation system and automatic temperature control; *d*, control unit for emergency call; *e*, compact disc (CD) changer and sound system; *f*, fiberoptic data-bus; *g*, class B controller-area network (CAN) data-bus; *h*, instrument cluster; *i*, class C CAN data-bus. (B) Electronic stability program (ESP) (Mercedes-Benz). Example 1: The car understeers; vehicle is stabilized through automatic braking intervention at rear left wheel. Example 2: The rear end breaks away due to oversteering; vehicle is stabilized through automatic braking intervention at front right wheel. (C) Dynamic autopilot system using radio data system (RDS) or cellular telephone (Mercedes-Benz). *a*, GPS satellites; *b*, radio station broadcasting traffic messages on traffic message channel (TMC); *c*, regional traffic computer centers; *d*, navigation system display; *e*, computer center for traffic messages via cellular phones; *f*, infrared sensor; *g*, induction control loops.

decades, the car radio provided the only—one-way—link with the outside world, and it was one way. Recently, the mobile phone ushered in an era of two-way communication. Satellite-assisted navigation systems to keep track of the exact position of a vehicle and an operating and display monitor, the basic components of an in-vehicle communication or telematics module, are now part of a modern car. This permits a wide range of new applications, including navigation systems to guide drivers to their destination with the aid of a CD-ROM road map. If current traffic data are also piped into the vehicle via the data interface (Fig. 17C), the route recommendation can take into account the traffic situation and, wherever possible, help avoid delays.

4. Monitoring of all important vehicle functions such as ABS, airbags, lighting, etc. These functions are constantly checked. In case there is a fault, the driver receives a warning indication and/or buzzer.

Despite these advances, however, there is no chance of electronics taking over entirely for the driver in the foreseeable future. Drivers will retain responsibility for monitoring the traffic situation themselves and reacting in the event of danger.

7. Air Conditioning

The air-conditioning system of the passenger compartment consists at least of a controllable air heater (heated by the engine-cooling water through a heat exchanger), fans, air pipes, and adjustable openings and nozzles. In ever-growing numbers, refrigeration systems are being installed. These consist of a compressor, evaporator, condenser, and water separator (Fig. 18). The fans for the air circulation are driven electrically, while the compressor is driven directly by the engine through an electromagnetic clutch or, in newer systems, are clutchless. The compressor output is then controled externally through an adjustment of the stroke.

Since comfort is a concept that varies widely among individuals, an extensive system of pipes, nozzles, and dampers makes possible manual or automatic individual regulation of air flow and direction, heating and cooling (often separately for the two sides in front and for the front and rear of the passenger compartment), fresh and circulating air, defrosting of the windows, and so on.

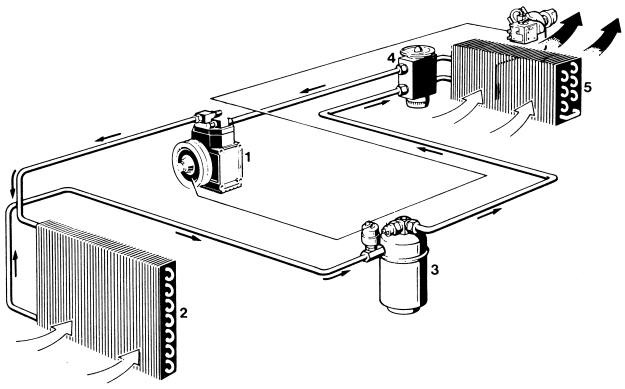


FIGURE 18 Air-conditioning system. (1) Compressor, cooling fluid; (2) condenser (heat exchanger); (3) container for cooling fluid; (4) expansion throttle; (5) evaporator (heat exchanger).

IV. THE POWER TRAIN

The power train consists of fuel supply, engine, gearbox, axle drive, shafts, and drive wheels in many different arrangements. (Fig. 19). The fuel tank is mounted in a position protected against impact (in passenger vehicles often behind or below the rear seats). It is fitted with a filler neck, vent (with evaporation control by means of a carbon filter), and a sensor for the fuel indicator. The fuel is supplied to the engine through feed and return pipes with filters, since the electrically driven fuel pumps (located in or near the fuel tank) pump more fuel than the engine uses in order to avoid vapor locks.

In addition, there are also fuel supply systems without a return pipe. Here, the engine receives only the required amount of fuel from the continously feeding fuel pump. Excess fuel is returned via a pressure regulator located close to the fuel tank. The advantage of this system is the reduction of the warming of the fuel in the tank which leads to minimized fuel vaporization.

In passenger vehicles, the engine is generally at the front (front engine), in front of, above, or very rarely behind the

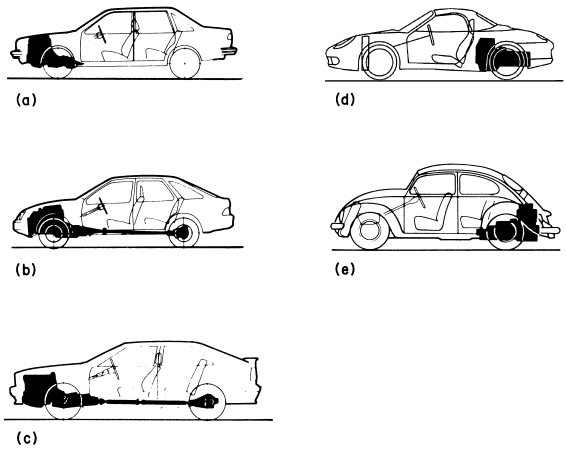


FIGURE 19 Power trains in passenger car. (a) Front-wheel drive; (b) standard drive; (c) all-wheel drive (here, four-wheel drive); (d) middle-engine drive; (e) rear-engine drive.

front axle; longitudinally or across the car; or (rarely) in front of the rear axle (middle engine) or behind the rear axle (rear engine). In commercial vehicles, the preferred engine position is in front of or above the front axle longitudinally. On buses, special advantages occur with the engine positioned behind the rear axle, either longitudinally or transversely.

For all vehicle categories, propulsion through a single axle predominates. This is sufficient for all normal road conditions, because a driving force exceeding 20–30% of the vehicle weight is rarely required for overcoming gradients or for acceleration. For passenger cars, front-wheel drive is preferred. Only for large passenger vehicles with high engine power and for commercial vehicles and buses is the preference for rear-wheel drive. The following combinations are recognized:

1. Front-wheel drive: Front engine (longitudinal or transverse), front-axle driven.
2. Standard drive: Front engine longitudinal, rear-axle driven.
3. Middle-engine drive: Engine in front of rear axle behind the front seats (longitudinal or transverse), rear-axle driven.
4. Rear-engine drive: Engine behind rear axle (longitudinal or transverse), rear-axle driven.
5. All-wheel drive: Front or rear engine, all axles driven.

Front-wheel drive makes very compact construction of the power train possible. There are only a few short wheel shafts, and the drive unit takes up little usable space. It has the advantage of retaining driving stability even with slipping drive (front) wheels. Although the basic load on the driven axle is rather high, a load reduction on the drive wheels occurs during acceleration, climbing, and towing. The drive through the jointed front axle requires homokinetic joints (mostly ball joints) with large angular displacement which transmits the driving torque uniformly even during cornering.

Standard drive is used for larger passenger cars and for most commercial vehicles. Since the engine and the drive axle are generally mounted separately, the optional fitting of different sizes of engine is facilitated. The additional loading of the driven rear axle during acceleration, on gradients, and through trailers also permits the use of large engine power. On a loaded truck, the rear axle generally carries two thirds of the total load, and with a truck train it is also stressed by the drawbar force. However, if the drive wheels of the standard drive slip, the road behavior becomes unstable and this can lead to skidding. In order to prevent this, ASR (Section 6) is used.

Middle-engine drive is only used in sports cars, because the high static load on the drive axle (the center of gravity is displaced toward the rear) and the low moment of inertia about the yaw axis (the engine mass close to the center of gravity) make cornering faster but more difficult. The space required by the engine between the axles reduces the usable space.

On buses, rear-engine drive is preferred because the absence of transmission shafts favors a low floor on city buses and a large baggage space on coaches. However, buses that are built on truck chassis keep their arrangement of engine and transmission.

All-wheel drive, which makes possible full utilization of the propulsion force given by the weight and traction coefficients, is used by off-road vehicles and trucks as well as by construction vehicles. The large degree of traction safety on a slippery road is attracting an increasing number of adherents for all-wheel drives, even for passenger vehicles. The cost of all-wheel drives is higher because they require additional driven axles, auxiliary gearboxes, differentials, and differential lock systems.

All the variations of the drive train arrangement have advantages and disadvantages. In each case, selection requires a careful analysis of the conditions for the application of the vehicle.

A. Driving Dynamics

Driving dynamics is the description of the interaction of vehicle and drive and thus gives the limits of performance (climbing ability, acceleration, and speed).

In order to move a vehicle, tractive resistance F_R, which consists of rolling, climbing, acceleration, and aerodynamic and possibly braking resistances, has to be overcome. Their reduction contributes greatly to fuel saving and this is therefore a constant aim in automobile development. The tractive resistance is given by:

$$F_R = mg(f_R \cos\alpha + \sin\alpha) + m_t a + F_B + (\rho/2)c_W A \cdot v^2(\mathrm{N})$$

where mg is weight, f_R the rolling resistance coefficient, α the angle of gradient, m_t vehicle and rotary mass, a acceleration, c_w the air resistance coefficient, A frontal area, ρ air density, v speed of travel (correct relative to air), F_B braking force (if applied) and g acceleration due to gravity.

The energy for overcoming gradients is first transformed into potential energy and that for acceleration into kinetic energy, which is degraded only to overcome the other resistances or during braking (increase of entropy). Because the rolling, climbing, and acceleration resistances are proportional to the mass of the vehicle, the reduction of weight is of great importance (conflicting with other requirements such as safety, noise reduction, and

TABLE III Energy Density of Transportable Storage Systems

		Fuel		Fuel plus container		
Type of fuel	Engine	(MJ/kg)	(MJ/liter)	(MJ/kg)	(MJ/liter)	Remarks
Diesel	Diesel	42.5	35.3	36.0	30.0	
Gasoline	Spark ignition	43.1	32.3	37.0	28.0	Super pressure tank
LPG[a]	Spark ignition	46.1	24.9	22.0	17.7	
Ethanol	Spark ignition	26.8	21.2	23.0	18.0	
Methanol	Spark ignition	19.7	15.6	17.0	13.0	
Hydrogen (gaseous)	Spark ignition	120.0	10.8[b]	6.6	1.6	Pressure tank (250 bar)
Hydrogen (liquid)	Spark ignition	120.0	8.5	20.0	4.9	−235°C alcryogenic tank
Hydrogen (hydride)	Spark ignition	120.0	—	2.3	3.5	Titaniron + Mg
Electricity	Electric motor	—	—	0.09	0.25	Lead-acid battery (25 Wh/kg)
Hydrogen (gaseous)	Fuel cell and electric motor	120.0	10.8[b]	6.6	1.6	Pressure tank (250 bar)
Hydrogen (liquid)	Fuel cell and electric motor	120.0	8.5	20.0	4.9	−253°C alcryogenic tank
Methanol	Fuel cell and electric motor	19.7	15.6	17.0	13.0	Requires fuel processor

[a] LPG-liquefied petroleum gas.
[b] MJ/m^3.

cargo space). That means reducing to a minimum any vehicle features that do not contribute to the conveying of persons or cargo. Technical means of reduction include integral body construction, shell instead of solid construction for load-bearing components, and the use of specific light materials such as high-tensile steel, light metal, plastics, and liminated material. Part of the weight-reduction effort, however, has to compensate for the weight of additional components that contribute to safety, environmental protection, and comfort. Braking always creates losses and should be avoided whenever possible. Driving in a manner that reduces the necessity of braking is the best way to save fuel. A reduction in the loss coefficients for rolling and aerodynamic resistance always has advantages, so work on their improvement continues. The rolling resistance has been improved by the introduction of the radial-ply tire and depends strongly on tire pressure and type of road surface. Since aerodynamic resistance varies with the square of the speed, in order to achieve low fuel consumption it is important to maintain not only a low speed for a given traveling time, but also a uniform one.

B. Fuel

Mechanical energy (force times distance) is required for overcoming road resistance. Automobiles achieve their independent mobility through the store of energy that they carry, usually enough for journeys of 400–1000 km. Because their weight and volume reduce the usable load and space, fuels of high energy density (and storage containers) are being sought (Table III). Liquid hydrocarbons in the form of gasoline and diesel oil are distinguished from other types of energy by their high energy yield per weight and volume (tank included). These fuels are also particularly favorable because, (1) they can be used over a large range of ambient temperatures, (2) the fuel tank can be refilled quickly and easily, (3) they have low toxicity and are safe, and (4) they are inexpensive (in principle), since hydrocarbons are refined products of primary energy (mineral oil). For this reason, liquid hydrocarbons have always been the preferred fuel for the automobile. Owing to the large battery weight, electric drive has had very little chance of finding wide application in the automobile so far. However, new improvements in the field of fuel cells could make the use of electrically driven vehicles likely in the future.

C. Engine

To ensure the propelling power of the vehicle, an engine is needed that converts the latent chemical energy in the fuel into mechanical energy. Unfortunately, doing so is always connected with losses (waste heat) owing to the laws of thermodynamics. Therefore, all combustion engines require a large cooling system. For water-cooled vehicle

engines, the system comprises a radiator, fan (driven by the engine, often via viscous coupling or electrically), water pump (driven by the engine), temperature controller, and water-cooling circuit (requiring antifreeze), which is usually pressurized in order to increase the cooling capacity.

When considering fuel economy, it must be remembered that the engine in the vehicle represents a power station that, in addition to the power for movement, supplies the power for operating the auxiliary equipment, such as servo pumps (for steering, automatic transmission, and suspension), refrigeration and air compressors, and electric generator. It also creates a vacuum for vacuum servos and, in addition, provides the necessary heat for heating the vehicle from its waste heat.

Within the power range required to drive a modern automobile (about 20–300 kW for passenger vehicles and 20–350 kW for commercial vehicles), spark-ignition and diesel engines, mostly of the four-stroke and reciprocating type (Fig. 20A and B), are superior to other types of thermal engines as a result of the following:

1. High power density relative to mass and volume
2. Favorable specific fuel consumption
3. Reliability and ease of manipulation
4. Few pollutants in the exaust gas and good possibilities of aftertreatment
5. Little noise or good potential for sound proofing or noise design

Neither the automobile gas turbine nor the Stirling engine has reached the same state of developement as the spark-ignition or diesel engines.

The torque M is the product of the mean effective pressure p_e and the swept volume V_h:

$$M = \frac{V_h \cdot p_e}{(2)2\pi} \text{ (Nm)}$$

where (2) is the four-stroke factor. The mean effective pressure is a measure of cylinder charge, efficient utilization of the air, and the quality of energy conversion. It can be increased by precompressing the induction air by charging.

In practice, there are two means of achieving this:

1. Turbocharging (exhaust gas turbocharger): The residual energy in the exhaust gas drives a small gas turbine (speed greater than 100,000 rpm), which in turn drives the compressor (Fig. 21).
2. Supercharging (mechanical charger or compressor): A spiral blower (Roots) is driven mechanically by the engine (mainly via toothed belts and a clutch; Fig. 21A).

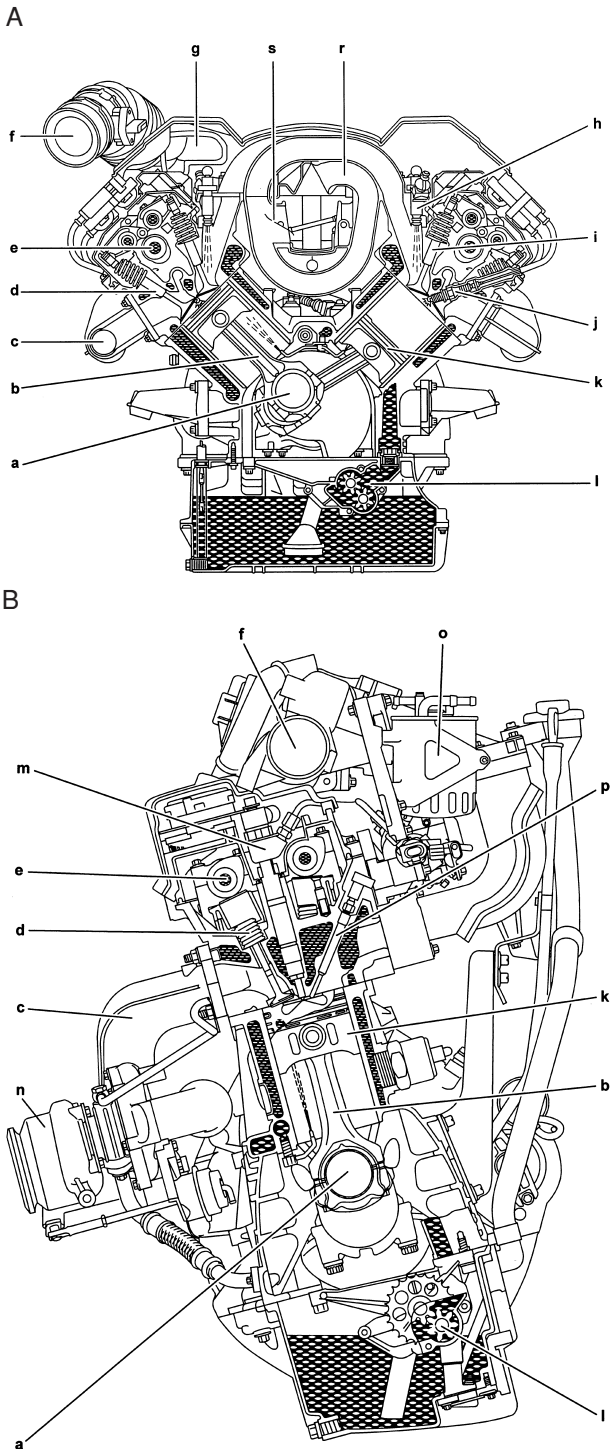


FIGURE 20 (A) Spark-ignition engine (SI, or otto-engine) (Mercedes-Benz). *a*, Crankshaft; *b*, connecting rod; *c*, exhaust pipe; *d*, outlet valve; *e*, camshaft; *f*, air intake; *g*, air filter; *h*, fuel-injection nozzle; *i*, inlet valve; *j*, spark plug; *k*, piston with rings; *l*, oil pump; *r*, switchover induction system; *s*, switchover induction system door. (B) Direct-injection diesel engine for passenger car (Mercedes-Benz). See legend to part (A); *m*, common-rail injection system; *n*, turbocharger; *o*, fuel filter; *p*, glow plug.

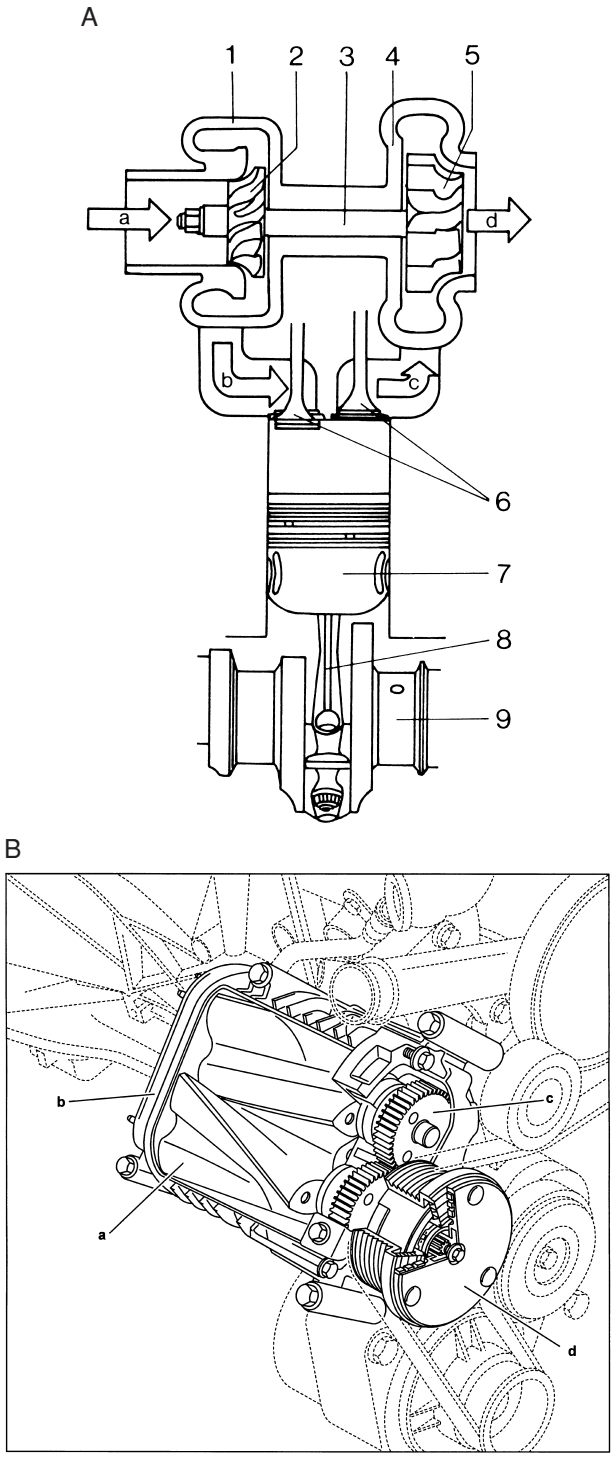


FIGURE 21 (A) Schematic of turbocharging of internal combustion engine. *a*, Air intake to turbocompressor; *b*, air intake to cylinder; *c*, gas exhaust from cylinder; *d*, gas exhaust from gas turbine; 1, compressor housing; 2, compressor wheel; 3, compressor turbine shaft; 4, turbine housing; 5, turbine wheel; 6, engine valves; 7, piston; 8, connecting rod; 9, crankshaft. (B) Roots compressor (Mercedes-Benz). *a*, Rotary piston; *b*, housing; *c*, gear wheels for synchronization of rotary pistons; *d*, electormagnetic clutch.

The essential parameter for the production of a required torque is the swept volume V_h, which in motor vehicles can be found over a wide range between 0.5 and 20 liters. This swept volume is divided between 2, 4, 6, or 8 cylinders or, rarely, between 3, 5, 10, or 12 cylinders. A small number of cylinders is less expensive and has lower friction losses but suffers from a very uneven torsional force diagramm (heavy flywheel) and free inertia forces. That is why 4, 6, or 8 cylinders in passenger cars, 12 cylinders in luxury cars, and 4, 6, 8, and more cylinders in commercial vehicles are preferred, but 5-cylinder engines are also to be found.

The engine power P, which determines the vehicle performance, is the product of the torque M and the angular velocity ω for which the speed $n = \omega/2\pi$ ($n/60$ if n is in revolutions per minute) can be substituted:

$$P = M\omega(W)$$

The highest speed is reached by the spark-ignition engine followed by the diesel engine for passenger cars. The lowest speeds are those of commercial vehicle diesels, which require considerably higher torques for the same output. This low speed is less process dependent but results from the requirement for extremely long running periods between overhauls (~150,000 km) and for the highest efficiency. Table IV illustrates typical data for automotive engines.

The classical aims of all engine developement are high power yield, high mean effective pressure, low fuel consumption, high reliability, and low cost. Additionally, there are strict requirements for low pollutant content in the exhaust gas. The products of complete combustion of hydrocarbons (HC) are water (H_2O) and carbon dioxide (CO_2). In real combustion, however, remains of carbon monoxide (CO) and unburned or partially burned hydrocarbons (CH) as well particulate matter (e.g., soot) can be found in the exhaust gas. In addition, especially at points of efficient engine operation, different types of nitrogen oxide (called NO_x) are formed from the constituents of the intake air oxygen and nitrogen. Because of the large number of vehicles, the small individual quantities of pollutants in the exhaust gas of individual cars when taken together have detrimental effects on health and the environment. Therefore, the amount of pollutant is strictly limited by law (emission standards), though differently in different countries. Since the quantity of pollutant in the exhaust gas depends both on the type of engine and on its means of operation, no uniform recipes for reduction can be given. In addition to possibilities of influencing the thermodynamic process (operation with excess air, recycling of exhaust gas, etc.), processes for exhaust purification (e.g., with catalytic converters) are important.

TABLE IV Typical Data for Automotive Engines

Engine type[a]	Angular velocity (rad/sec)	Compression ratio (ε)	Mean effective pressure (MPa)	Output per		Engine efficiency maximum (η_e)
				Swept volume (kW/liter)	Mass (kW/kg)	
Spark ignition						
Passenger car						
NA	450–750	8–12	0.8–1.5	35–70	0.33–1.0	0.24–0.36
SC	500–700	7–10	1.1–2.0	50–105	0.33–1.0	0.22–0.34
Commercial vehicle						
NA	250–500	7–9	0.8–1.0	20–30	0.17–0.33	0.22–0.31
Diesel						
Passenger car						
NA	350–500	19–24	0.6–0.8	20–30	0.20–0.33	0.26–0.35
SC	350–450	17–24	0.9–1.2	35–50	0.25–0.50	0.29–0.35
Commercial vehicle						
NA	200–400	16–18	0.7–1.0	10–15	0.11–0.25	0.35–0.40
SC	200–320	15–17	1.0–1.3	15–20	0.13–0.33	0.37–0.40
Intermediate cooling						
SC	180–260	14–16	1.3–1.8	20–25	0.20–0.33	0.37–0.43

[a] NA = naturally aspirated; SC = supercharged.
From SAE. "Bosch Automotive Handbook," Society of Automotive Engineers, Warrendale, PA. With permission.

In the following sections, two types of vehicle engine will be discussed: spark-ignition and diesel.

1. Spark-Ignition Engine

The spark-ignition engine (Fig. 20A) is the universal engine for all fuels of low inflammability, especially gasoline (a mixture of several hydrocarbons), but also liquefied petroleum gas (LPG), liquefied natural gas (LNG), methanol and ethanol (alcohols), and hydrogen. The thermal efficiency depends mainly on the compression ratio ε, which is limited by the knock tendency (i.e., tendency toward spontaneous ignition of the fuel, which is characterized by the octane number). High octane numbers approaching 100 are sought. The knock resistance of a fuel depends on the composition of the hydrocarbon mixture and additives, some of which act as knock inhibiters (lead, tetraethyl, and tetramethyl). The spark-ignition engine works with a mixture of air and atomized fuel produced by fuel injection into the induction pipe (widely used), by fuel injection into the combustion chamber (increasing in application), or in the carburetor (classical method, rarely used today).

The ignition of the fuel mixture by an electric spark is possible only when conditions are close to being stoichiometrically correct, that is, only when the quantity of oxygen in the air is correct for complete oxidation of the hydrocarbons. For petrol (and for diesel fuel as well), this is ~14.5 kg of air per kilogram fuel ($\lambda = 1$). Because a mixture ratio close to stoichiometry must be provided within the cylinder under all air conditions (temperature, pressure), output control can be achieved only by altering air density, which is done by throttling the intake air.

The stoichiometric ratio of the gas–air mixture ($\lambda = 1$) makes possible very effective purification of the exhaust gas in a three-way catalyst (Fig. 22). With the aid of an O_2 probe measuring oxygen in the exhaust, the mixture before it reaches the engine is regulated to the stoichiometric ratio as accurately as possible so that in the first part of the catalyst, owing to a shortage of oxygen, any traces of carbon monoxide reduce nitrogen oxides to nitrogen by oxidizing to carbon dioxide. Eventually, any remaining carbon monoxide and unburned hydrocarbons are fully oxidized in the other part of the catalyst. Operation with an exhaust-gas catalyst requires lead-free gasoline; otherwise, the effect of the catalyst materials (platinum and rhodium) would be rapidly reduced.

Spark-ignition engines can run very smoothly and at low noise level and are therefore expecially suitable for passenger cars. Their only disadvantage is their low efficiency at part loads due to losses by throttling the air. The development of these engines is therefore aimed at an improvement of part-load efficiency through improved knock-resisting combustion chambers; variable valve timing and compression ratios; three, four, and five valves per cylinder (central location of the spark plugs); more accurate

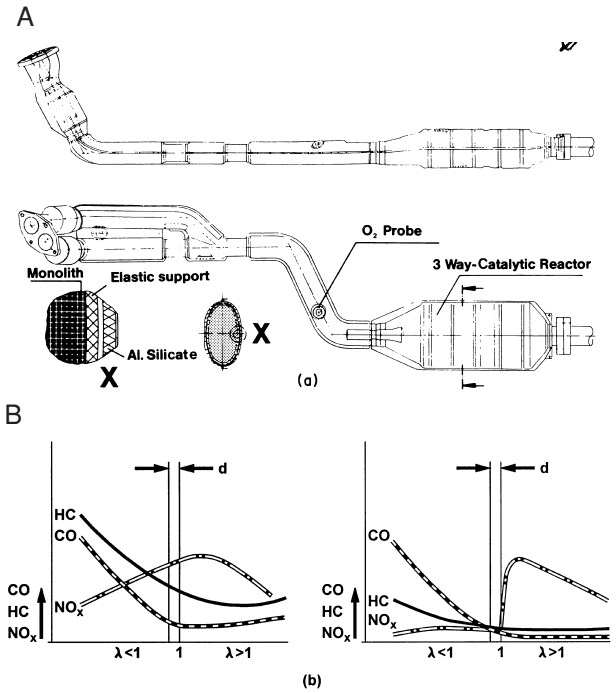


FIGURE 22 (A) Emission control by three-way catalyst (catalytic reactor) and λ control (O_2 probe); exhaust pipe. (B) Effect of closed-loop emission control. Quantity of pollutants CO, HC, and NO in exhaust gas as function of ratio λ:

$$\lambda = \left[\frac{\text{kg fuel}}{\text{kg air}}\right]_{\text{stoichiometric}} \Big/ \left[\frac{\text{kg fuel}}{\text{kg air}}\right]$$

The left side is measured before the catalyst; the right side is measured after the catalyst. $\lambda < 1$, excess of fuel; $\lambda = 1$, stoichiometric mixture; $\lambda > 1$, excess of air. d = Acceptable range to be maintained by closed-loop control of air–fuel mixture via λ control.

electronic control of mixture formation; timing along the knocking limit; warming-up period; and idling.

Another potential measure to improve efficiency is to de-throttle the spark-ignition engine. Modern direct-injection control produces an ignitable mixture around the spark plug, whereas the overall cylinder is filled with a lean mixture ($\lambda > 1$). By doing so, the fuel-consumption can be lowered at part load by 8–15%, depending on the characteristics of the lean-burn operation in the engine map.

The residual oxygen in the exhaust gas has a detrimental effect, however, as it inhibits the use of the three-way catalyst (oxidation occurs before reduction). This means that the catalysts for lean-burn engines must be used, so-called NO_x-storage catalysts.

In the lean exhaust gas ($\lambda > 1$), NO_2 is adsorbed as nitrate on the storage component (e.g., barium) of the catalyst. Desorption occurs in the rich ($\lambda < 1$) exhaust gas. The nitrates then disintegrate and the generated NO_x is simultaneously reduced because there is enough reducing agent in the rich ($\lambda < 1$) exhaust gas. This requires a discontinuous (lean/rich) operation of the engine with typical intervals of about 60 sec (lean) and 3 sec (rich).

The use of lean-burn engine catalysts places extra demands on the fuel: all fuels must meet set parameters and be as much as possible free from sulfur, which has a toxic effect on the described catalytic coatings.

2. Diesel Engine

In the diesel engine (Fig. 20B), fuel is ignited purely by the heat of compressed air. This requires, on the one hand, fuel with a high ignition quality (characteristic high cetane number approaching 50) and, on the other hand, a high compression ratio ε to heat the air sufficiently, which increases peak pressure and makes the running less smooth. The induction air is unthrottled, which provides higher efficiency at part load. The output is controlled by the fuel quantity alone; the fuel is injected under high pressure up to 250 MPa into the already compressed and therefore heated air in the cylinder and thus ignited. The diesel engine works (with smoke-free exhaust gases) only with excess air (equivalence ratio $\lambda > 1$), so that the mean effective pressure, the torque, and the power output are lower but the efficiencies, especially at part loads, are higher than for a spark-ignition engine.

Due to the good fuel-economy it offers, the direct-injection diesel engine has been the preferred engine in commercial vehicles for many years. Today's diesel-powered cars take advantage of novel direct injection techniques such as common rail or unit-injector systems. These techniques improve the field noise, handling, and emissions, which were previously pre- and swirl-chamber diesel engine domains. The use of charging techniques (i.e., exhaust turbochargers) yields optimal torque and power output.

The diesel engine produces less CO, CO_2, and HC than the equivalent spark-ignition engine, although the permanent presence of residual oxygen in the diesel fuel causes considerable problems for exhaust gas treatment. Particulate emissions are reduced either in the engine itself (e.g., by increased pressure during injection) or by using filter systems.

The total of nitrogen oxides produced is cut down to one third with the help of exhaust gas recirculation. Complex technology (e.g., so-called $DENO_x$ catalysts) is required to further reduce this value which unfortunately take their toll on the traditional advantages of the diesel-engine such as fuel economy and reliability.

The $DENO_x$ catalyst shows a certain potential to reduce NO_x already in the lean ($\lambda > 1$) exhaust gas with sufficient reducing agent such as CO and HC. The reducing agent is supplied either in the raw exhaust gas or

actively via post-injection (only with common-rail systems). Unlike the lean-burn, spark-ignition engine, the diesel engine operation is therefore continuous.

D. Drive Train

The combustion engine in a vehicle cannot be connected directly to the wheels and cannot move the vehicle directly because (Fig. 23):

1. The engine operates only above a minimum speed (the idling speed) and cannot produce any power below that.
2. Every engine has a maximum speed, determined by the limit of the process or by strength factors.
3. The engine torque is, in principle, constant at full load and independent of speed. (In practice, the torque first rises somewhat with increasing speed before dropping again.)
4. With absent or insufficient fuel supply, an external torque must be applied to drive the engine, which amounts at full speed to only about one third of engine torque.

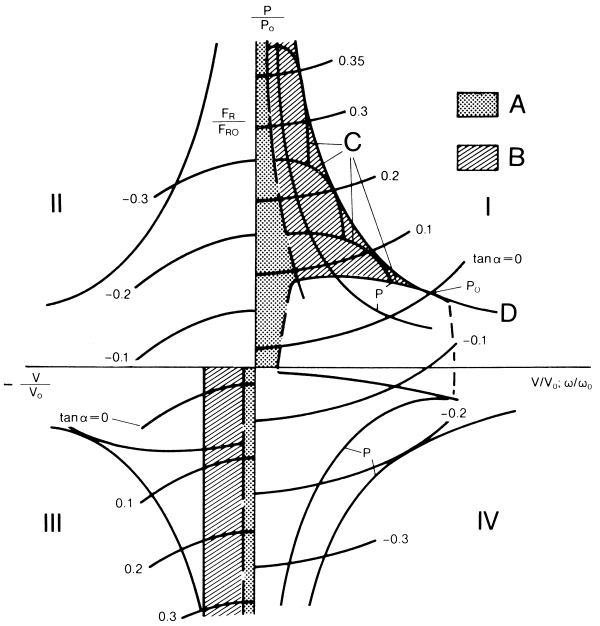


FIGURE 23 Vehicle-engine operation diagram standardized to the point of maximum engine power $P_o = M_o\omega_o = F_{Ro}v_o$, where P is power; F_R, traction force, v, vehicle speed; ω engine angular velocity; $\tan\alpha$ gradient $\tan\alpha = 0$ tractive resistance at plane. (A) Speed conversion (torque transmission under slip); (B) torque conversion (torque increase); (C) tractive force (torque) at different gear ratios of the transmission; (D) hyperbola of maximum engine power P_o. *I*, Forward drive, traction; *II*, reverse drive, coasting; *III*, reverse drive, traction; *IV*, forward drive, coasting.

5. The maximum output P_O is obtained only at one speed, that is, at only one point of the operating field.
6. Automotive engines rotate in only one direction.
7. The angular velocity range of the engine is high.
8. The crankshaft of the engine has a rotary movement.

In the engine operating area, which is bounded by the minimum and maximum speeds and the full-load and towing torques, every point can be set by altering the engine control element (throttle valve, control rod, or electronic control). Compared with this, the vehicle requires the following (same items as for the engine):

1. The motive force is required at zero speed for starting and at low speed for traveling.
2. The upper limit of speed is theoretically determined only by the transportation task.
3. A high traction force may already be required at zero and at low speeds to overcome large gradients and for high accelerations.
4. For braking on a gradient, a brake force equaling the traction force may be required.
5. The vehicle requires the maximum driving power over a wide speed range.
6. The vehicle operates in all four quadrants of the performance map—namely, forward, reverse, uphill, and downhill.
7. The speed range of the vehicle (measured as the angular velocity of the wheels) is low.
8. The movement of the vehicle is longitudinal.

It must be possible to operate the vehicle at every point of the complete performance map between the chosen limits of gradients, up and down, and forward and reverse.

The matching of engine and vehicle is done through the drive train, including transmission, axle gear, and wheels. The transmission must accordingly fulfill the following requirements:

1. Transfer of the engine torque at slipping speed in order to bridge the speed gap between engine and vehicle during the starting process (speed converter)
2. Adjustable conversion of engine torque M at the same output P to overcome the driving resistance F_R according to:

$$P = M\omega\eta_d = F_R v \text{ (W)}$$

where η_d is the efficiency between engine and wheels
3. Reversal of rotation for driving in reverse

In order to comply with these requirements, all transmissions consist of a speed converter, a torque converter, and

a reversing mechanism. The design of speed converters is described below, together with the gear boxes.

The real torque converter is the gearbox, which has to have a variable ratio range I, which is determined by the driving requirements and the vehicle characteristics. Calculating with the net engine output to weight ratio $(P/G)_{\text{eff}}$, the overall ratio I can be approximately expressed as:

$$I = \frac{(f_R + \sin\alpha)v_0}{(P/G)_{\text{eff}}\varphi} \approx \frac{\tan\alpha\, v_0}{(P/G)_{\text{eff}}\varphi}$$

Here, the index 0 refers to the point of maximum engine power, $\tan\alpha$ is the gradient that must be overcome, and v_0 is the highest possible speed the vehicle could reach. The effective power/weight ratio $(P/G)_{\text{eff}}$ must be calculated with the power actually available at the wheels (after subtracting all the power for auxiliary drives and losses) and the actual weight. Special transport requirements must be taken into account, such as drop of air density with increasing altitude (power reduction) and additional weight from towing a trailer.

The speed factor φ characterizes the driving resistance curve in the engine map at top gear (smallest gear ratio) and influences the highest attainable speed v_m, the power reserve ΔP (acceleration and climbing capacity without shifting), and fuel consumption (Fig. 24).

$$\varphi = \left[\frac{v_0}{v}\right]_{\omega_0} = \left[\frac{\omega}{\omega_0}\right]_{v_0}$$

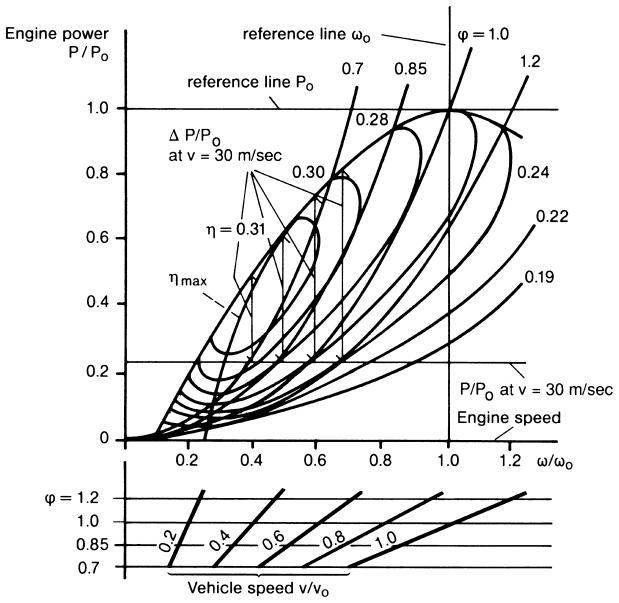


FIGURE 24 Engine map with tractive resistance lines for various speed factors φ standardized to the point of maximum engine power $P_0 = M_0\omega_0$. $\Delta P/P_0$ is the power reserve for various factors φ when driving at 30 m/sec (108 km/h). $\eta = \eta_e$ engine efficiency.

The following applies:

$\varphi > 1 : v_m < v_0$;	ΔP increasing;	η_e decreasing
$\varphi = 1 : v_m < v_0$;	ΔP mean;	η_e mean
$\varphi < 1 : v_m < v_0$;	ΔP decreasing;	η_e increasing

The greater requirements for climbing, the higher the highest speed v_0, and the lower the power/weight ratio the greater is the required overall ratio I of the transmission. Commercial vehicles, whose power/weight ratio usually amounts to only one tenth of that of passenger cars (0.5–1 versus 5–15 W/N, respectively), require a particularly large overall torque ratio I of the transmission.

1. Types of Transmission

For torque conversion, a continuously variable ratio transmission (CVT) would be the ideal solution. It would make it possible to operate the engine always at its optimum efficiency. Unfortunately, the efficiency of CVTs and their reliability, cost, and weight are not yet satisfactory. Owing to their inherent advantages, however, more work is being done on their development. Of the different types of CVTs (mechanical, hydrostatic, electric) for road vehicles, the belt-type CVT with a steel belt reaching maximum efficiencies of 95% has gone into production (Fig. 25). By varying the axial distance of the cone disks, the working radius of the belt and hence the torque ratio of the shafts can be altered over a range of about 5–6. A speed converter and reverse gear are also required.

Today, gearboxes with a discrete number of fixed ratios (speeds) that can deliberately be put in operation are the quasi-standard transmission in all motor vehicles because of their outstanding qualities, including simplicity,

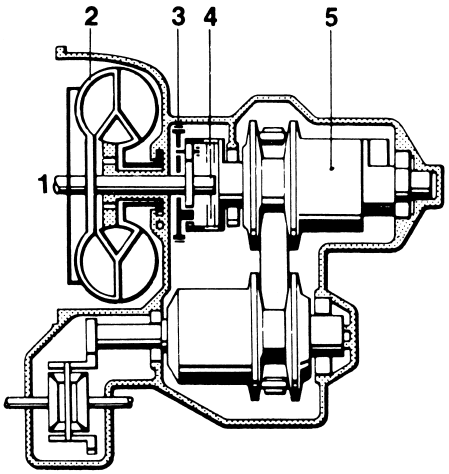


FIGURE 25 Continuously variable transmission (CVT) of the steel-belt type (PIV, Reimers). (1) Input shaft; (2) torque converter; (3) reverse gear; (4) forward clutch; (5) CVT.

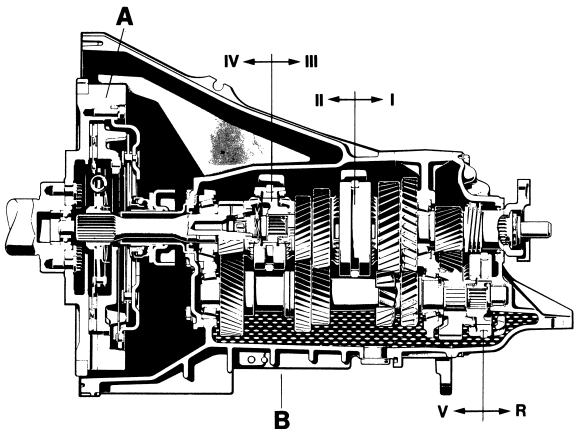


FIGURE 26 Manually shifted five-speed gearbox for passenger car (Mercedes-Benz). (A) Main clutch. (B) Gearbox. *I*–*V*, Five forward speeds; *R*, reverse gear. The arrows show the direction in which the shift fork has to be moved to engage the indicated speeds.

high efficiency, low cost, and favorable power/weight ratio. The required operation point of the vehicle is then a combination of one of the fixed gear ratios and the variable set of the engine operation point. The number of fixed gear ratios depends on the size of the overall gear ratio I, the flexibility of the engine, and the degree of adaptation to the hyperbola of maximum engine power P_0 required.

For a passenger car, the overall ratio I is 3–6, and the number of speeds 4–6. For commercial vehicles, I is 6–17, the number of speeds 5–16. Figures 26 and 27 illustrate two examples of manually operated gearboxes. The speed converter used for manually operated gears is the friction clutch (Fig. 26), which acts at the same time as the separating clutch for gear shifting. In its normal position, it is closed by spring pressure and can be opened by the driver through the clutch pedal (connected through mechanical or hydrostatic linkage). In order to shift one of the speeds, the main clutch must be released and the power flow interrupted. The loose gear of that speed is then connected with its shaft through a jaw clutch. To facilitate changing gear, the jaw clutch is preceded by a small cone clutch, which permits the gear change only when the speeds of the parts to be engaged are synchronized (have the same speed).

Advances in control technology and actuators have allowed the production of electronically controlled clutch systems in large-scale series. In the electronically controlled clutch system, the clutch is activated by a computer-assisted hydraulic servo unit, depending on driving conditions and the gear selected by the driver. These clutch systems can be coupled to an automated (often sequential) transmission, which, in the automatic mode, allows the advantages of a manual gearbox (no losses due to slip) without the driver having to shift the gears or activate the clutch pedal (improved comfort). With the automated transmission, the gears are shifted not by the driver but by computer-controlled actuators. However, one drawback is the interruption of tractive force during gear change. These fully automated manual transmissions are not to be confused with the classic automatic transmissions as described below.

For automatic transmissions (Fig. 28), the best "speed converter" has proved to be the fluid drive. (Although it

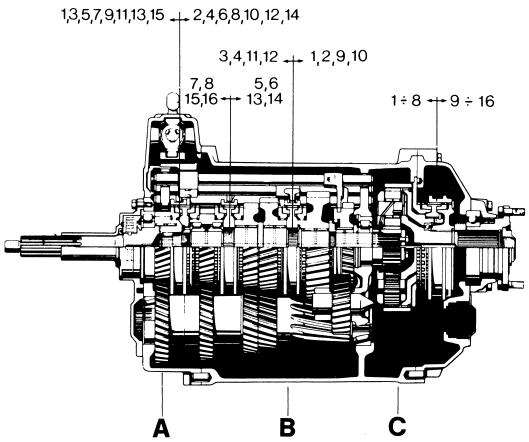


FIGURE 27 Manually shifted gearbox for trucks, 16 speeds (ZF Zahnradfabrik Friedrichshafen). (A) Splitter group (two speeds); (B) main gearbox (four forward, one reverse speed); (C) auxiliary group (two speeds). The ratio of each engaged speed i_s is the product of the ratios of the gear groups $i_g = i_A i_B i_C$. The arrows and figures indicate the speeds in which the respective position of the shift fork is used.

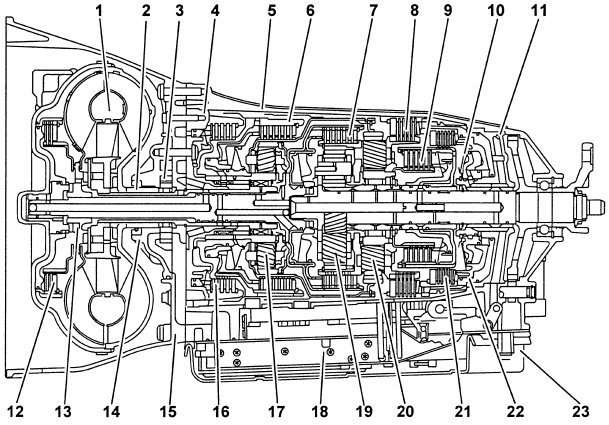


FIGURE 28 Five-speed automatic transmission for passenger car (Mercedes-Benz). (1) Torque converter; (2) drive flange oil pump; (3) oil pump; (4) front cover; (5) transmission housing; (6) disk carrier; (7), (9) multidisk clutch; (8), (16), (21) multidisk brake; (10) back-pressure spring; (11) cylinder; (12) lockup clutch; (13) flange output rotor; (14) oil pump housing; (15) converter housing; (17), (19), (20) planetary gear set; (18) control unit; (22) piston; (23) oil pan.

is usually a hydraulic torque converter, in the automobile drive consists of a centrifugal pump driven by the engine which transforms the mechanical energy into the hydraulic form of the mass flow of oil; a turbine, connected to the input of the gearbox, which reconverts the hydraulic energy to mechanical energy; and a reactor wheel, which is connected to the stationary housing and supports the differential torque between pump and turbine wheels. The three vane wheels are assembled in a rotating housing closely following one another. When the turbine is stalled, the input torque of the pump, depending on the design, is measured by a factor of two at the output of the turbine. With increasing turbine speed, the turbine torque drops almost linearly and reaches the input torque at the coupling point. The maximum efficiency during torque conversion hardly exceeds 90%. Since the reactor is usually supported by a free-wheeling mechanism, a negative reactor torque releases it and it starts to rotate. The fluid drive has now become a fluid coupling (that is, a true speed converter) which can reach efficiencies of greater than 98% (two-phase converter).

An automatic transmission (Fig. 28) saves the driver the burden of operating the clutch and shifting the gears. Because gear change has to be done automatically and without interrupting the tractive force (power shift), all the servo units for doing this must be of the friction type. Automatic transmissions for motor vehicles are usually a combination of a torque converter for starting and planetary gear sets with several speeds. With planetary gears, the arrangement of the servo elements, such as disk clutches, disk brakes, and band brakes, is more favorable. The program of automatic gear changing is usually a function of engine torque, driver's intention (gas pedal and selector lever position), and vehicle speed. The control system was in the beginning hydraulic and is now a combination of electronic and hydraulic components. The former gear-shift lever has become a selector lever with which the driver can select the positions: Park (output shaft is locked; the engine can be started), reverse, neutral (the gear is in idle; the engine can be started), and several positions for forward (in order to adapt to special traffic conditions, such as braking with the engine when going downhill). The development of the automatic transmission has resulted in an increased number of speeds as a consequence of speed factors $\varphi < 1$ and an increasing use of electronics for transmission and engine control (simultaneously).

2. Differential Gears

In the axle gear (Fig. 29), the different speed levels of engine and vehicle are matched by a constant gear ratio. On trucks, additional fixed gears are also to be found as wheel hub gears (planetary axle). All-wheel drive requires an additional auxiliary gearbox to divide the power between the front and rear axles. Because the speeds of all the driven wheels are never the same, and during cornering even differ considerably, they cannot be connected rigidly. In all axle transmissions and many auxiliary gearboxes, differentials are fitted. In some auxiliary gearboxes, torque splitting between front and rear axles is done by viscous couplings.

The differential gear (Figs. 29 and 30) is a planetary set of gears with bevel gears in which the planet carrier is driven and the central gears are connected to the wheel

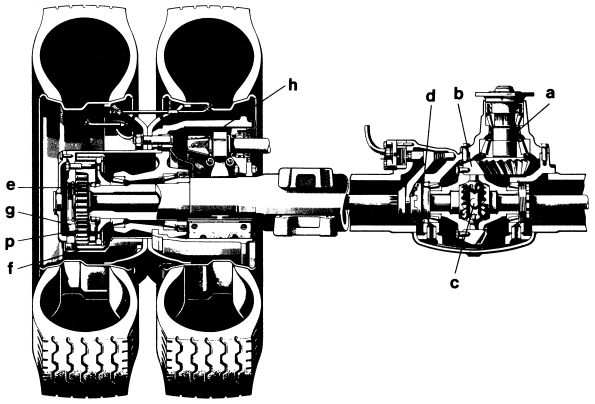


FIGURE 29 Rear axle for truck; speed reduction in bevel gear and a planetary gear within the wheel hub, cross-section (Mercedes-Benz). Bevel gear: (a) bevel pinion gear; (b) bevel ring gear; (c) differential gear; (d) differential lock; (e) sun gear; (f) ring gear; (g) planet carrier; (h) brake drum; (p) planet gear.

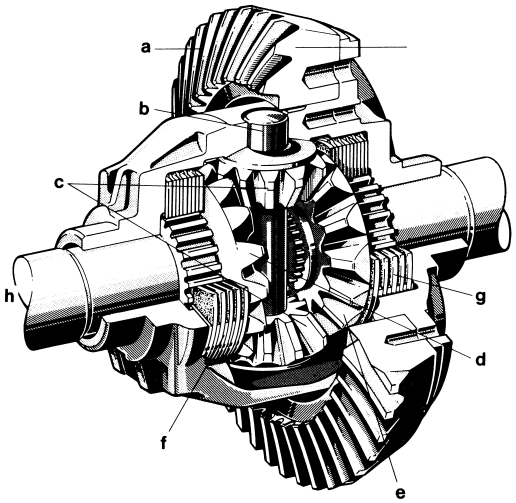


FIGURE 30 Differential gear for passenger car with limited slip (Mercedes-Benz). (a) Bevel ring gear; (b) carrier with planet axle; (c) planet gears; (d) central gear, left side; (e) central gear, right side; (f) disk clutch, left side; (g) disk clutch, right side; (h) wheel shaft, left side.

shafts. In a planetary gear set, the input torque is divided between the output shafts in proportion to the ratio of the number of teeth of the central gears, which may run at different speeds. This also means, however, that the traction of all wheels is strongly reduced when it is slippery and one wheel slips. In order to reduce or eliminate this disadvantage, differential gears with limited slip are fitted to passenger cars (Fig. 30), while in some commercial vehicles manually operated differential locks are mounted with which a rigid connection between the wheels can be made (Fig. 29). (If engaged, any speed difference between the wheels must then be compensated for through slip between the tires and the road.)

Another possibility for a differential lock is to specifically intervene by braking a slipping single wheel, thereby transferring the propelling force to the other half of the axle (two-wheel drive) or to the other axle (four-wheel drive). During this procedure, brake regulation is controlled by computer.

The power train ends at the driven wheels, through which the traction and braking forces are transmitted to the road surface by friction. Currently, only wheels with steel or light metal rims and tubeless tires are used (Figs. 11B and 29). Radial tires have almost completely replaced the former cross-ply tires because they have a longer life and a lower rolling resistance. Their harder rolling characteristic can be accepted owing to progress in road construction and suspension design.

V. DRIVING

The automobile gains its capacity to function only with a human as sensor, controller, and operator in the control loop of roadway, vehicle, and driver. The most important task of automobile technology is to use all the technical means available to make this difficult control function as simple and as safe as possible, so that the driver can carry it out, to a large extent, subconsciously. The concrete demands made on the technology are derrived from an analysis of the driving function, as discussed in this section.

The desired path of motion is determined macroscopically by the destination but microscopically, within narrow limits, by the road width, other road users, and traffic regulations. Since these limits are fluid, safe driving requires both uninterrupted observation of the external conditions, in order for the driver to be instantly aware of often rapid changes, and readiness to react in order to compensate immediately for changes and disturbances. The recognition of this essential (though not sufficient) condition for safe driving arises from the long distances covered in 1 sec (e.g., ~25 m at 55 mph) compared with the often much narrower space for maneuvering that is available for collision avoidance maneuvers or braking. A collision could have been avoided in a great majority of cases if one of the two participants in the accident had reacted only 0.5 sec earlier (accident compensation time according to Enke, 1979).

The technical tasks listed below, which arise from an analysis of the control process of driving, have not all been fulfilled by any means but remain the aim of all development projects. They concern both the vehicle and the power train. It is possible to formulate them comprehensively only when certain knowledge about the vehicle and the transmission is available. This underlines the unity of the complete vehicle with all its assemblies and components and all its partial functions; no change can be made anywhere without the overall function and safe driving of the vehicle being affected. The control task of the driver includes: (1) intake of information, (2) information processing, and, when necessary, (3) correction of the vehicle movement in the longitudinal and/or the lateral directions. For the driver to carry out this control function, the conditions discussed in the following subsections must be fulfilled.

A. Keeping all Information Channels Open

1. Visual

The driver must have an all-around view of the front and sides through glass panes; a view to the rear through mirrors that is as far as possible without blind spots and distortion and is of undiminished transparency (transmission grade of glass type, inclination, and tint); windshield cleaning by wiper and washer with sufficient wiping area; and rapid defrosting of all windows by hot air or electric heating. The entire field of road lighting by headlamps is also part of this requirement.

2. Vestibular

Because the driver is not rigidly connected to the vehicle, a sufficient positive and nonpositive connection between vehicle and driver must be created through the hardness, shape, and springiness of the seats and backrests so that information regarding acceleration (force) can be transmitted and uncontrolled movements of the driver can be avoided. Distortion of information through vibrational resonances must be eliminated.

3. Kinesthetic

A "feel" for road contact and for the correct proportioning of the gas and break pedals requires proportional action

feedback of the forces acting during steering and braking (especially important with servo power). In all the steering, accelerating, and braking systems, backlash, friction, and other distortions must be avoided. Intended nonlinearities (variable steering ratio or, with power steering, increasing steering resistance with growing speed, etc.) must be checked very carefully for their subconscious effects.

4. Auditive

The noise level in the vehicle must be reduced and all irritating individual noises must be eliminated so that "useful noises" from the outside or from a radio or communication system can be taken in immediately without distractions.

5. Information Output

The vehicle must convey information about itself to other road users. This includes its contours, color, lights, direction indicators, and brake lights.

B. Subconscious Operation of Control Functions

The operation of the vehicle must not require any conscious attention of the driver. This means the following:

1. Operations must be limited to a few easily established procedures, as far as possible only steering and accelerating or braking.
2. The number of necessary additional operations should be reduced to those that are indispensable (direction indication, headlight dimming, etc.).
3. It must be possible to perform the unavoidable remaining operations "blind"—that is, without searching and looking—which predetermines the number, location, and form of the actuator. Actuators for operations whose timing is prescribed by the traffic (direction indication, headlight dimming, speed regulation, etc.) must be located within reach of the steering wheel so that the wheel does not have to be released if at all possible. Actuators for operations for which the driver can decide the timing himself can be located more remotely (radio, air conditioning, lighting, etc.).
4. Anything that can be automated should be (e.g., air conditioning system, lighting, automatic transmission, skid and stability control, automatic four-wheel drive).
5. All operations that require too great an effort or must be performed too frequently must be reduced with the aid of servo power, such as power breaks, power steering, and power sun roof, window, and seat adjustment, etc.
6. All operations that require the driver's attention for searching, reading, or handling must be avoided wherever possible.

C. Linear Control Characteristic of the Vehicle

The directional behavior of the vehicle must follow the commands of the driver in a repetitive and predictable manner. Any vehicle-specific training must be rejected because drivers handle unfamiliar cars very frequently. Included in this are the following:

1. There should be no rebound of the front-wheel suspension (anti-lift) and no pitching of the body during braking (anti-dive). This can be achieved through the axle kinematics when the lines of action of the braking and starting supports pass through the instantaneous poles. Usually it is possible to obtain only an approximation.
2. The steering should be between neutral and understeer, which means constant or increasing steering motion for the same curve radius and at increasing speed (Fig. 31). This is achieved by displacing the major part of the centrifugal force support to the steered front wheels. It has been proved that a normal driver masters a vehicle with neutral steering or with understeer more safely in critical situations (no requirement for sudden countersteering, of which sports car drivers are so proud of).
3. There should be compensation for varying driving or braking forces on the steered front wheels through the negative steering offset (Fig. 32). With a positive steering offset, unequal forces at the two front wheels

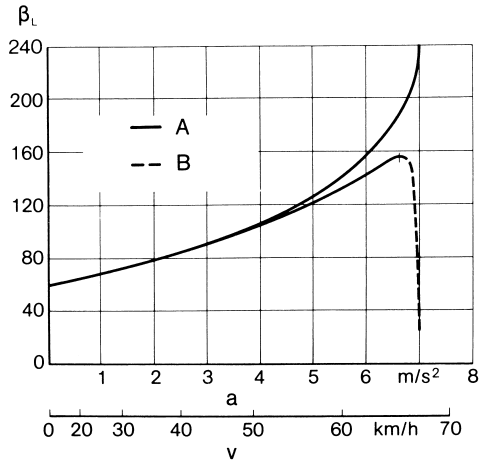


FIGURE 31 Directional behavior of a vehicle traveling in a circle of radius 50 m. *A*, understeering; *B*, oversteering; β_L, angle of steering wheel; *v*, vehicle speed; *a*, lateral acceleration.

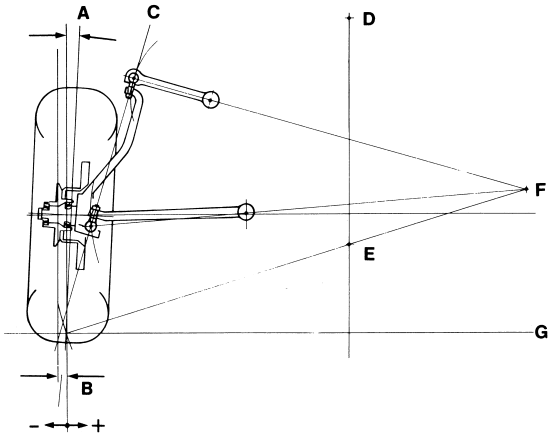


FIGURE 32 Suspension geometry, front wheel of passenger car, front view. (A) Wheel rake (camber); (B) steering offset; (C) steering axis; (D) center line of vehicle; (E) instantaneous center; (F) instantaneous pole; (G) ground.

cause a steering motion because the moments about the steering axles are no longer balanced. On the other hand, a negative steering offset stabilizes straight-line driving.

4. There should be automatic return of the steered front wheels on release of the steering wheel and self-stabilizing straight-line running through caster action or the inclination of the front wheels (Fig. 33). With both these measures, a realignment torque of the turned front wheels is obtained which is proportional to the speed and promotes both a steering "feel" and saves the necessity of straightening the steering wheel actively from the curve.
5. All disturbances should be eliminated from straight-line running through vertical or horizontal movements of the wheels and the body—for example, through the three-dimensional kinematic suspension of all the wheels (space control arm shaft, spring strut, and wishbone). This requirement is extremely difficult to satisfy because all intentional and unintentional resiliences of the axle and wheel suspensions can lead too easily to movements of the wheels through jouncing or blows to the wheels; this causes a change in the direction of the vehicle, thus requiring a steering correction by the driver (Fig. 34).
6. The steering capability must be retained during all driving maneuvers and traction conditions up to the speed limit for the curve, by the use of antilock, wheel-slip, and stability controls. When wheels lock, the driver can no longer steer the vehicle. Antilock systems therefore have become part of the standard equipment of most vehicles in recent years (Fig. 16).

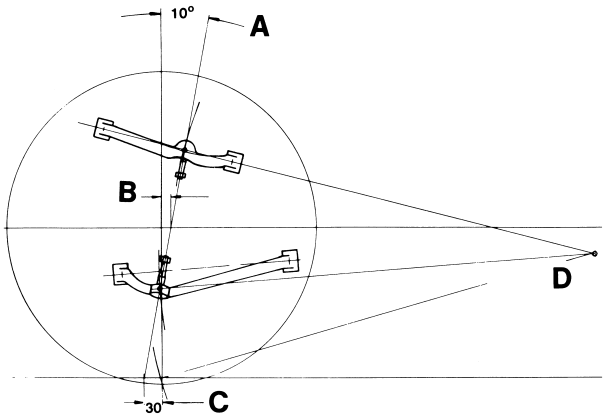


FIGURE 33 Suspension geometry, front wheel of passenger car, side view. (A) Caster angle; (B) caster offset; (C) caster distance; (D) instantaneous pole.

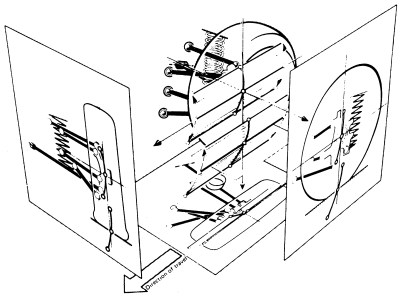


FIGURE 34 Rear-axle wheel movements in several planes, multilink independent suspension (Mercedes-Benz).

D. Lack of Distraction by Indicators Inside the Vehicle

Because uninterrupted observation of the traffic area is an essential condition for safe driving, every distraction that makes demands on the visual information channel through indicators in the vehicle should be avoided (reduction of indicators and warning buzzers). Current developments appear to run counter to this requirement as the latest electronics make possible the provision of increasingly sophisticated indicators, including route calculators, maps, and traffic guidance systems. However, electronic data processing will be used more and more to reduce the need for indicators, since, after all, they often report only technical faults.

E. Looking After the Driver's Condition

In order to maintain the control function of the driver as long and as well as possible, it is important to maintain the efficiency of the driver. Some considerations include:

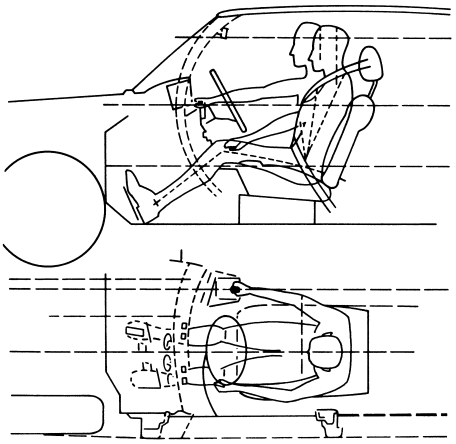


FIGURE 35 Ergonomic conditions for the driver of a passenger car.

1. Making vehicle vibrations endurable
2. Seats suitable for the body in shape and feel and with regard to frequency, amplitude, and damping of the vibrations system
3. Orthopedically shaped and possibly heated and ventilated seats
4. Devices for adapting the seats, backrests, and steering wheel positions to body dimensions, preferably power-operated
5. Physiologically endurable air conditioning of the passenger space by sufficient, nondrafty air circulation and heating
6. Ergonomically correct arrangement of all operating elements in relation to the driver's place, with special attention being paid to the body dimensions of the total population of potential drivers (Fig. 35)
7. Elimination of standing acoustic vibrations
8. Weakening of monotonous noise levels and optimization of useful sounds such as traffic information, radio, mobile phones, and conversation

VI. PASSIVE SAFETY

Because the ability to function and the safety of automobile traffic depend on the correct actions of the driver, the possibility of incorrect behavior (even accidents) must be accepted as reality. An important part of automobile technology consists, therefore, in preparing for accidents; this means that the automobile has to be designed to increase the chances of survival of the passengers and to reduce the degree of injury. This task is in fact a packaging problem. The container "vehicle" and the packaging of the "freight" (i.e., the occupants) must be arranged to ensure that they will survive the impact as much as possible without being damaged. In order to achieve this aim, a set of safety standards and test conditions has been prepared. Since head-on collisions are frequent, an important role is played by the mandatory head-on collision test at 48.3 km/h (= 30 mph) into a rigid barrier at an angle of either 0° or 30°. Of head-on collisions, 60% involve mainly the portion of the front of the vehicle facing oncoming traffic. Therefore, laterally offset frontal collisiontests with coverage of 40% of the vehicle width are also carried out.

The survival criteria (Fig. 36)—head injury criteria (HIC)—were established on the basis of biomechanical experiments that have also led to a description of the severity of the injuries in the form of the abbreviated injury scale (AIS), which ranges from AIS 0 (no injury) to AIS 6 (unsurvivable injury). Other standards and tests are impact from the rear, overturning, and side impact, and there are various standards and tests on parts of the cell.

The technical task consists of reducing the impact deceleration for the passengers so that there is a chance of surviving a collision. The technical requirements are as follows:

1. Rigid passenger cell as survival space for the occupants

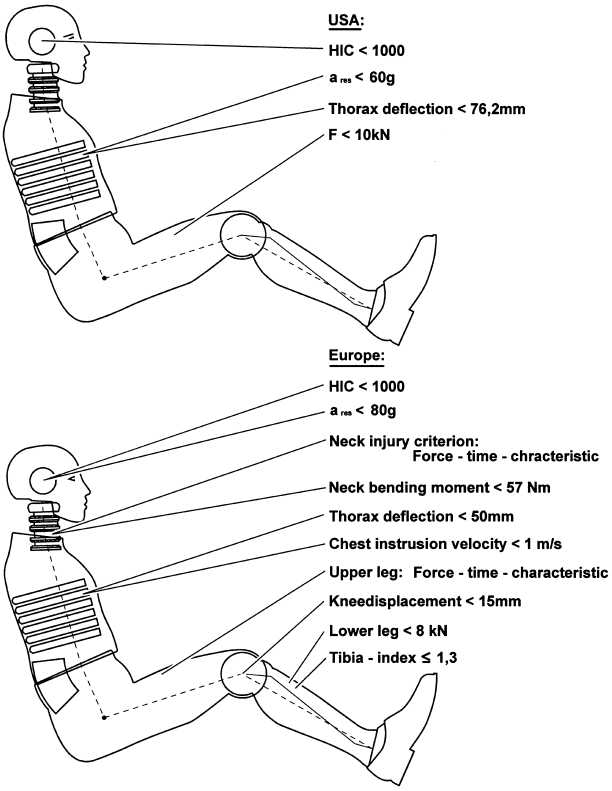


FIGURE 36 Dummy criteria (U.S. and Europe) for frontal impact. *HIC*, head-injury criteria; a_{res}, resulting acceleration; *F*, force.

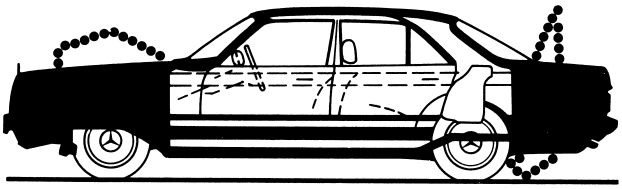


FIGURE 37 "Crashworthiness" of a passenger car according to B. Barény in German patent 864157 (1951). Passenger compartment is rigid, but front and rear segments deform under impact forces.

2. Deformable front and rear sections for the reduction of deceleration peaks and dissipation of impact energy. This is the only way of reducing high-impact decelerations to a degree passengers can tolerate (Fig. 37)
3. Blunting all hard edges on the inside and flush-mounting and rounding off all operating levels and knobs, deformable and padded surfaces, and columns in the interior (instrument panel, steering wheel hub, etc.)
4. Limiting the depth of penetration of the steering wheel (for example, position of steering gear behind the front axle, telescopic or buckling steering column) and foot pedals
5. Nonsplintering screens (friable or laminated glass) to avoid injury from cuts
6. Strong attachment of all the vehicle occupants to the passenger cell by a restraint system
7. Strength of seats and backrests and configuration of head restraints to withstand the supporting forces of the occupants
8. Installation of fuel tank and fuel pipes in places that are protected during an impact so no fuel leaks
9. Blunting the external contours of the vehicle. This is important to alleviate the results of a collision with unprotected road users, such as pedestrains, cyclists, and motorcyclists
10. Underrun protection on commercial vehicles to prevent other road users from getting underneath the vehicle

All these measures are important parameters for body construction; they determine the basic concept of the structure, reinforcements, and deforming sections; bumpers; doors and door locks; seals and backrests; folding internal and external mirrors; position and shape of operating levers; and much more. Automobile engineers employ complicated methods of calculation to determine the impact processes and to predetermine deformations and decelerations during collision. They also carry out extensive experiments, including crash tests with complete cars. However, only the analysis of actual accidents gives information about the results of these efforts.

While the technical conditions for the vehicle can be met, at least in principle, guaranteeing the other requirements of good packaging (namely, the support of the occupants in the passenger cell) presents a much greater problem, because the occupants will not allow themselves to be treated as objects to be packed in excelsior (even though in an accident they are in fact at the mercy of dynamic forces since humans cannot prevent forward movement by using their arms and legs at speeds greater than 12 mph). It is a question of finding a way of confining the occupants in a vehicle that is satisfactory technically and that the car occupants will use, even during short trips. In an emergency, neither the lower part of the body nor the torso must be allowed to fall too far forward.

The technical solution to this problem is a combination of different occupant restraint systems:

1. Automatic safety belt (manual systems): The most often used, three-point belt with an inertia reel device (automatic safety belt) is a sound compromise between effective safety, practicability, comfort, and cost. When a set deceleration value is reached, a built-in, quick-action lock blocks the belt reel.
2. Belt tensioning systems (Fig. 38): Belt tensioning systems should be regarded as a complementary improvement to the automatic safety belt systems.

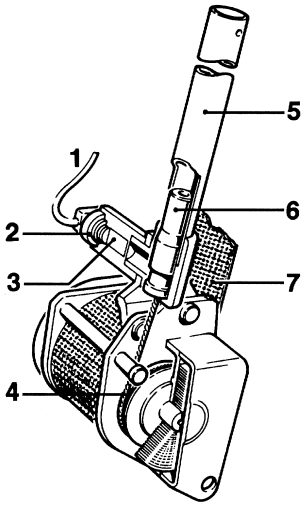


FIGURE 38 Belt tensioner. The sensor activates via the cable (1) to the priming cap (2), which then ignites the propellant. The piston (6) moves in the cylinder (5) and pulls the wire cable (4), which then turns the belt retractor spindle. The belt strap is finally tightened through rotation of the belt retractor spindle. (From SAE. "Bosch Automotive Handbook," Society of Automotive Engineers, Warrendale, PA. With permission.)

They help to reduce forward displacement of passengers due to belt slack in serious accidents. The system reduces the difference in speed between the passenger and the car and the displacement of passengers and thus also the injuries this causes.

3. Airbag (automatic systems) (Fig. 39): The airbag can reduce or prevent for the passengers the force of impact on parts of the vehicle interior (steering wheel, dashboard, doors, B-pillars).

In regard to airbags, when a crash occurs, a sensor ignites a pyrotechnic gas generator which instantly inflates the airbag so that it is completely inflated before the passenger comes into contact with it. On contact, the airbag deflates slightly, allowing the force of impact from the passenger to be absorbed softly with physiologically tolerable surface impact and deceleration values. This can prevent, or at least significantly reduce, injuries to the head and chest. The timely inflation of the airbag is crucial to the effectiveness of the protective function. The passenger must come into contact with the airbag only when it is fully inflated and has just begun to deflate. There are various airbag systems:

1. Front airbag system: The driver and front-seat passenger both have an airbag to protect against collisions with a rigid obstacle, as even a belt tensioner cannot always prevent the head from impacting with the steering wheel in major accidents. The whole process of collision and energy absorption is over in approximately 150 msec for the driver airbag.
2. Side airbag systems (sidebag and windowbag): The reduction of injuries, to the head and thorax during lateral collisions through the use of side- or windowbags, which inflate from the door, the backrest, or the roof frame, is difficult because of the barely existing crumple zone and the small distance separating passengers from intruding vehicle parts. Crash detection and airbag activation times must therefore be only 3 msec for hard lateral impact, and inflation time of the 12-liter windowbag should only be around 10 msec.

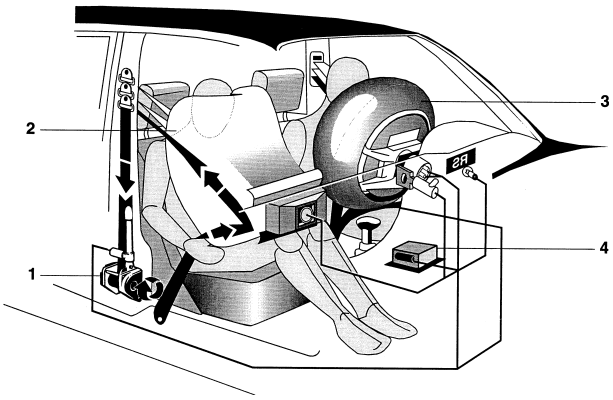


FIGURE 39 Occupant restraint system. (1) Belt tensioner; (2) driver-side airbag (triggered); (3) front-passenger air bag (triggered); (4) sensor for belt tensioner and airbag, plus additional energy reserve. (From SAE. "Bosch Automotive Handbook," Society of Automotive Engineers, Warrendale, PA. With permission.)

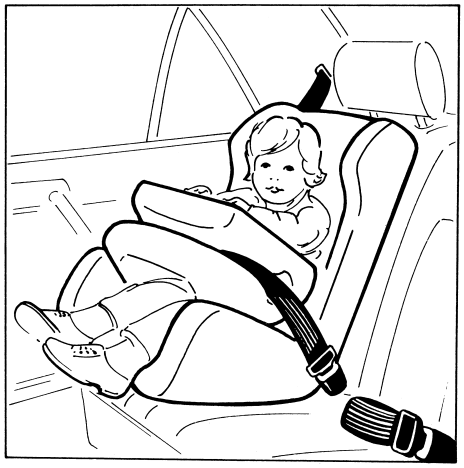


FIGURE 40 Restraining system for children 1–3 years old.

The number of protective functions is set to rise in the future. "Smart bags" will make it possible to bring triggering mechanisms and speed of inflation into line with passenger position, type of accident, use of belt, temperature, etc. However, children under 12 need special restraint systems that are suitable for their body size and weight with respect to their age. Figure 40 shows an example for children from 1–3 years.

VII. OUTLOOK

Is the development of automobile technology gradually approaching its peak, or even reaching an end, after more than 110 years, or can further advances be expected? In fact, development has never been more active than today. It has become clear that the application of electronics to the automobile will continue to increase, and though we do not yet have a clear view of all possible areas of applications, the challenge is evident and the potential for achievement is great.

In addition, opportunities for development will come from a quite different direction—namely, future production processes. Since the beginning of mass production, new processes for the production of automobiles have been developed continually. Individual handmade manufacture was followed by the conveyor belt; the multispindle, single-purpose machine; the transfer line; body presses; welding and paint-spraying lines; right up to the use of interlinked numerically controlled (NC) machines and complete handling systems. All these production

processes have been created in order to produce the automobile inexpensively and reliably. While in the past a prerequisite for profitable manufacture was the production of a large number of identical parts, electronic processing in machine controls, robots, assembly, and inspection will make the manufacture of much smaller numbers profitable and will permit modification to be made earlier in the prodcution process. This will enable the automobile engineer to design many variations that will be better adapted to the customer's needs and tastes.

The automobile of the future will be designed more in accordance with the differentiated dimensions, characteristics, abilities, and needs of humans; it will travel more safely in their hands; and, at the same time, it will do more justice to the demand for conservation of the environment.

SEE ALSO THE FOLLOWING ARTICLES

AUTOMOTIVE CONTROL SYSTEMS • ELECTRIC AND HYBRID VEHICLES • INTERNAL COMBUSTION ENGINES

BIBLIOGRAPHY

Enke, K. (1979). "Possibilities for Improving Safety within the Driver-Vehicle-Environment Control Loop," pp. 789–802, U.S. Dept. of Transportation, Washington, DC.

Forbes, T. W., ed. (1972). "Human Factors in Highway Traffic Safety Research," Wiley–Interscience, New York.

IME. (1983). "Road vehicle handling," Institution of Mechanical Engineers Conference Publication No. 3.

Jones, N., and Wierzbicki, T., eds. (1983). "Structural Crashworthiness," Butterworth, London.

Korff, W. H. (1980). "Designing Tomorrow's Cars," M-C Publishing, Burbank, CA.

Limpert, R. (1982). "Vehicle System Components: Design and Safety," Wiley, New York.

Newton, K., Steeds, W. I., and Garrett, T. K. (1983). "The Motor Vehicle," 10th ed., Butterworth, London.

SAE. "Bosch Automotive Handbook," Society of Automotive Engineers, Warrendale, PA.

Seiffert, U., and Walzer, P. (1984), "The Future for Automotive Technology," Pinter, Dover, N.H.

Taylor, C. F. (1985). "The Internal Combustion Engine in Theory," MIT Press, Cambridge, MA.

Weston, E. B. (1976). "Automobile Transmission Systems," Newnes Technical Books, Sevenoaks, G.B.

Concrete, Reinforced

R. Park
University of Canterbury

I. Introduction
II. Cement and Concrete
III. Development of Reinforced Concrete
IV. Design Approaches
V. Modern Reinforced Concrete Theory
VI. Prestressed Concrete
VII. Ferrocement and Fiber-Reinforced Concrete
VIII. General Applications of Structural Concrete

GLOSSARY

Aggregrate Inert granular material such as sand, gravel, or crushed stone that is mixed with Portland cement and water to form concrete.

Beam Horizontal structural member subjected primarily to loads producing flexure.

Cast-in-place concrete Concrete cast in its final position in the structure.

Column Vertical structural member subjected primarily to compressive loads.

Composite concrete members Concrete members of precast and cast-in-place concrete elements constructed in separate placements but so interconnected that all the elements respond to forces as a unit.

Deformed reinforcement Steel bars with surface deformations used for reinforcement in concrete.

Modulus of elasticity Ratio of normal stress to corresponding strain for tensile or compressive stresses below the limit of proportionality of the material.

Precast concrete Concrete cast in other than its final position in the structure.

Prestressed concrete Concrete in which tensioned steel elements are used to introduce internal stresses of such magnitude and distribution that the stresses resulting from external loading are counteracted to a desired degree.

Reinforced concrete Concrete containing steel reinforcement and designed and detailed so that the two materials act together in resisting forces.

Service loads Loads comprising the dead weight of the structure, live load representing external working loads, and other loads, such as wind, earthquake, and earth pressure, expected during normal working conditions.

Stirrup Reinforcement located perpendicular to and around the longitudinal reinforcement in beams used to resist shear, torsion, and buckling of the longitudinal bars and to confine the compressed concrete; similar to hoop and tie.

Strength design Members of structures are designed taking inelastic strains into account to reach ultimate (maximum) strength when an ultimate load equal to the sum of each service load multiplied by its respective load factor is applied to the structure.

Structure Assemblage of members that provide resistance to forces acting on a building or bridge.

Ultimate strength design Similar to Strength design.

Working stress design Members of structures are designed using elastic theory so that at the service loads the stresses in the reinforcing steel and the concrete do not exceed specified allowable stresses, which are fixed proportions of the ultimate or yield strength of the materials.

CONCRETE is a mixture of Portland cement or other hydraulic cement, fine aggregate, coarse aggregate, and water, with or without admixtures. It is a material that is relatively strong in compression but weak in tension. The successful use of concrete in structures has come about from the addition of steel reinforcement to provide tensile resistance. Reinforced concrete is designed on the assumption that the concrete and steel reinforcement act together in resisting forces. Reinforced concrete is now one of the most common materials from which structures are built, being widely used for the construction of buildings, bridges, and many other types of structures. Reinforced concrete structures may be of cast-in-place construction or formed from precast concrete components generally acting compositely with cast-in-place concrete. Advances in the analysis, design, and construction of reinforced concrete structures have been based on analytical and experimental studies and on practical experience.

I. INTRODUCTION

Concrete is relatively strong in compression but weak in tension. The compressive strength of concrete in structures is typically 3000–9000 psi (20–60 MPa), and the tensile strength is normally about 10–15% of the compressive strength. The successful use of concrete in structures has come about from the addition of metal reinforcement to provide tensile resistance. When properly reinforced, concrete is transformed from a relatively brittle material into a composite material with a reasonable and calculable strength in compression, tension, flexure, and shear. The use of reinforced concrete in structures has not come about as a result of a single vital discovery, but rather its use has increased gradually through the years as knowledge and experience have accumulated. The first use of iron and steel as reinforcement in concrete was made by practical engineers and was not based on the considerations of theorists. The success of many of these early reinforced concrete structures was based on sound judgment. However, there were many failures as well, due to the lack of an analytical approach and to errors in detailing reinforcement. Confidence in and development of reinforced concrete design and construction through the years have been achieved by combining the ideas of practitioners with the growing knowledge of the behavior of reinforced concrete obtained from analysis and experiment. Concrete is now one of the most common materials from which structures are built.

II. CEMENT AND CONCRETE

The development of concrete as a construction material has required the ready availability of a reliable cement. Cementing materials of various types were used by the ancient Egyptians, Greeks, and Romans. After a decline of knowledge in the Middle Ages, John Smeaton provided the first significant advance when he was commissioned in England in 1756 to rebuild the Eddystone Lighthouse off the Cornish Coast. For the construction of the lighthouse he used a pozzolana mixed with lime as a mortar to bind the stones together, having determined that the capacity of certain limes to harden under water was dependent on their clay contents. The first artificial cement was produced by the English mason and building contractor Joseph Aspin, who took out a patent in 1824 for cement made by burning a mixture of clay and limestone together. Aspin referred to this product as Portland cement because of its resemblance to Portland stone. The scientific basis for the production of modern Portland cement can be credited to Isaac Charles Johnston, a compatriot of Aspin, who in 1844 in England patented the process of continuing the burning process until sintering and, by experiment, obtained the best mixture of clay and chalk. Modern Portland cement is generally made from aluminum and silica as found in clay or shale and from a calcareous material such as limestone or chalk. The process of manufacture consists of grinding the raw materials, mixing them in certain proportions, burning in a large rotary kiln to a temperature approaching 1400°C until the material sinters and fuses into clinker, and grinding the clinkers when cool into a fine powder with some gypsum added to give Portland cement. In addition to ordinary Portland cement, a number of special-purpose cements are now made, for example, rapid-hardening Portland cement, high-alumina cement, and sulfate-resisting cement.

Concrete is a mixture of cement, fine aggregate, coarse aggregate, and water, with or without admixtures. The aggregate is inert and normally occupies a large proportion of the volume of the concrete. In concrete mix design, the distribution of the aggregate particle sizes and the relative proportion of cement, aggregate, and water are determined. These variables are normally expressed as the aggregate grading, aggregate/cement ratio, and water/cement ratio. In concrete mix design, values for these

variables are sought, bearing in mind that the mix should be as economical as possible, that the fresh concrete should be sufficiently workable to be properly placed, and that the hardened concrete must be durable and most of it must attain a specified compressive strength. A mean strength greater than the specified strength must be the aim of the mix design because of the inevitable variations in the actual strengths achieved, and this requires the application of statistical and probability theory.

The most important variables affecting the strength of concrete at a given age are the water/cement ratio and the degree of compaction. When concrete is fully compacted, its compressive strength is inversely proportional to the water/cement ratio, as was discovered by Duff Abrams in the United States in 1919. Abrams's finding was a special case of a more general law formulated by Feret in France in 1896 that expressed the concrete strength in terms of the cement/total voids ratio. The presence of air voids in concrete causes a reduction in the compressive strength, and hence adequate workability for proper compaction during placing is very important. The use of concrete in large-scale works was made possible by the development of power-driven mixers and was aided by the discovery by Freyssinet in France in 1917 of the value of mechanical high-frequency vibration for the compaction of concrete.

III. DEVELOPMENT OF REINFORCED CONCRETE

The principle of combining concrete and iron as a composite structural material received considerable attention in the latter half of the nineteenth century. In 1848 Lambot in France built a concrete rowing boat that was reinforced by a rectangular mesh of iron rods. In 1849 the Paris gardener Joseph Monier made for the first time concrete tubs with a reinforcement of wire mesh. The first patent for the use of reinforced concrete as a composite structural material was taken out in England in 1854 by Wilkinson. The patent was for embedding in floors or beams of concrete (either arched or flat) a network of flat iron bars or secondhand wire rope raised over the supports and sagging near the bottom of the beam or slab at midspan. Little was done in England to follow Wilkinson's lead. In France in 1852, Francois Coignet had made a building by encasing an iron skeleton framework in concrete. In 1855 Coignet took out an English patent that included the statement that floors could be made "by burying beams, iron planks or a square mesh of rods in concrete on falsework." Coignet foresaw most uses of reinforced concrete, but he produced no theory on which to base design. In the United States in 1878 a patent was granted to Thaddeus Hyatt for construction in reinforced concrete. Hyatt recognized the importance of the near equality of the coefficients of thermal expansion of concrete and iron and advocated the use of deformed bars for reinforcement. However, it was mainly through the expansion of Monier's activity that the engineering world at large became acquainted with reinforced concrete. Monier's numerous patents taken out from 1867 onward involved containers, floors, beams, pipes, bridges, and other elements. It is notable that Monier and the other early pioneers did not use any theoretical calculations to determine the dimensions of the concrete or reinforcement.

The first theoretical analysis of the behavior of reinforced concrete beams was published in 1886 in Germany by Koenen, who assumed that the steel alone took the tensile stresses, the neutral axis was at the middepth of the section, and the concrete above the neutral axis had a straight-line distribution of compressive stress. It was P. Neumann who first recognized the importance of the relationship between the moduli of elasticity of the steel and concrete on the position of the neutral axis. However, it was Edmund Coignet (son of Francois Coignet) and N. de Tedesco who in 1894 in France first presented the elastic theory for flexure in a form similar to that used today. Their theory assumed that plane sections before bending remained plane after bending, the material obeyed Hooke's law, the ratio of the moduli of elasticity of steel and concrete was a constant, and all tension in the concrete was ignored. There were many rival theories in the 1890s because it was known that concrete did not strictly obey Hooke's law, and not all theorists were convinced that it was reasonable to neglect the tensile resistance of concrete (Fig. 1).

Reinforced concrete construction developed particularly rapidly in France. In 1895 Considère had begun an extensive experimental study of reinforced concrete beams and columns, which led to a number of important concepts—for example, the use of spiral reinforcement in columns as a means of reducing the cross section of heavily loaded columns and the anchorage of the ends of main bars in beams by forming a hook with a diameter five times that of the bar. However, the greatest influence on reinforced concrete construction was that of Francois Hennebique (1843–1921), whose genius resulted in many innovations. For example, Hennebique is credited with having introduced the T beam, in which the cross-sectional shape makes the best use of the materials making up the composite section (Fig. 2a). Hennebique also combined columns, beams, and floors into completely monolithic structures. The Hennebique system was introduced to Britain in 1897. Figure 2b shows a typical continuous beam from his patent of that year. Hennebique's work marked the beginning of a new era of reinforced concrete construction.

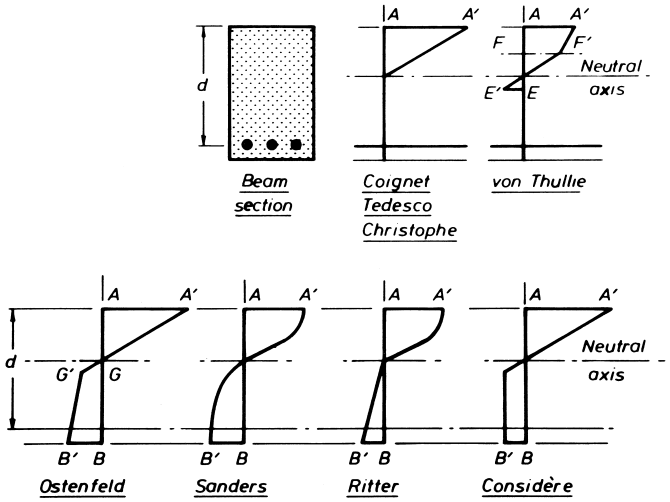


FIGURE 1 Some rival theories around 1900 of the distribution of longitudinal stress down the depth of a reinforced concrete section (AA', maximum compressive stress in concrete; BB', maximum tensile stress in concrete).

Other significant early contributions made by Europeans were those of Freyssinet in France, Emperger in Austria, and E. Mörsch in Germany. Mörsch's contributions included a classical textbook on reinforced concrete construction. The pioneer in the application of reinforced concrete in the United States was Ransome, who erected his first fully concrete-framed building in 1903.

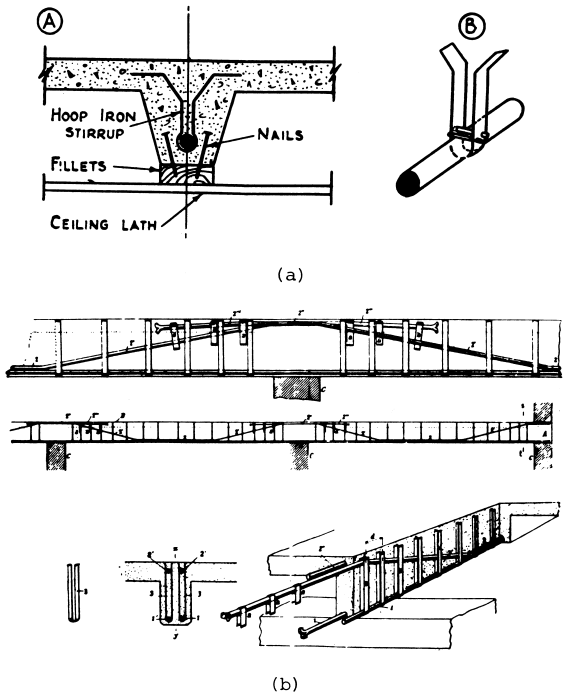


FIGURE 2 (a) Hennebique's T Beam of 1892 and details of the stirrup and cotters. (b) Hennebique's main British patent of 1897.

Early structural forms in reinforced concrete commonly followed traditional concepts. For example, some early floors were clearly direct imitations of floors made entirely of wood or of wood supported on iron or steel beams. However, new structural forms that owed their origin to the particular properties of reinforced concrete were developed. An example is the flat slab floor (a floor without beams), which was mainly an American innovation, being patented in the United States in 1902 by O. W. Norcross. The use of the flat slab floor for heavy loading was made possible by the addition of the drop panel (a slab thickening around the column) by A. R. Lord and the enlarged mushroom head to the column by C. A. P. Turner, both of the United States. It is of interest that there was no acceptable form of analysis available when the first flat slab floor was built by Turner in 1906 in Minneapolis. The building was erected at Turner's risk and load-tested before the owner would accept it. Other applications for which concrete proved suitable were frames, walls, and shell roofs for buildings, tanks, bridges, and other structures.

There has been a steady development in the theory of reinforced concrete since the start of the present century. The properties of the materials are now much better understood and have been greatly improved. The stresses in and the modes of failure of the common range of reinforced concrete structural elements and assemblages have received considerable investigation, both theoretically and experimentally. As a result calculations can now be made with increasing confidence, both in the service load range using elastic theory and at the ultimate load of the structure using ultimate strength theory.

The theory of ultimate flexural strength considers the behavior of a member when the concrete has reached its

compressive (crushing) strength and the reinforcing steel has generally reached its yield strength. Charles Whitney in the United States in 1942 suggested replacing the actual curved shape of the distribution of compressive stress in the concrete by an equivalent rectangle, resulting in a convenient analytical simplification when determining the ultimate flexural strength. Flexural strength calculations can be carried out with accuracy. Unfortunately, the theories developed for the ultimate strength in shear and torsion are still significantly dependent on empirical data, as are aspects of bond, anchorage, and cracking behavior.

Reinforced concrete structures have, if well constructed, shown very good durability and fire resistance through the years. The demonstrated durability is due mainly to the high alkalinity of Portland cement concrete, which inhibits the corrosion of embedded steel reinforcement, provided that a reasonable thickness of well-compacted concrete cover is present over the steel and that the crack widths are limited. A reasonable thickness of concrete cover over reinforcement also leads to very good fire resistance, since the thermal conductivity of concrete is relatively poor. Fire ratings have commonly been stated in terms of the concrete cover thickness necessary to keep the steel beneath a critical temperature for the prescribed duration of a fire.

IV. DESIGN APPROACHES

A. Working Stress Design

Several of the early studies of the flexural behavior of reinforced concrete members can be likened to ultimate strength theories, because they assume nonlinear distributions of concrete stress down the depth of the section (see, e.g., Fig. 1). However, when reinforced concrete theory was first developed, the elastic (straight-line) theory of Coignet and Tedesco became generally accepted, mainly because elastic theory was the conventional method of design for other materials and also because the straight-line distribution of stress led to mathematical simplification. In addition, tests had shown that the use of elastic theory with carefully chosen values for the working stresses led to a structure displaying satisfactory behavior at the service loads and having an adequate margin of safety against collapse. Thus elastic theory has formed the basis of reinforced concrete design for many years.

In the "working stress" design approach the stresses in the structure at the service loads as calculated by elastic theory are limited to allowable stresses, which are specified proportions of the concrete compressive strength and the steel yield strength. Typically the allowable stresses are taken to be about one-half of the material strengths. For example, in the building code of the American Concrete Institute (ACI) the allowable stresses for reinforced concrete beams designed by this method are $0.45 f'_c$ for concrete in compression and 0.4 to $0.5 f_y$ for steel reinforcement in tension, where f'_c is the specified compressive strength of concrete and f_y the specified yield strength of the reinforcement.

B. Ultimate Strength Design

Recently there has been renewed interest in ultimate strength theory as a basis of design. After more than half a century of practical experience and laboratory tests, the knowledge of the actual behavior of structural concrete has vastly increased and the deficiencies of the working stress (elastic theory) design method have become increasingly apparent. Design based on ultimate strength first became accepted as an alternative to working stress design in the building codes for reinforced concrete of the ACI in 1956 and of Britain in 1957. In "ultimate strength" design (or "load factor" design) the sections of the members are designed taking the inelastic (plastic) strains into account so as to have sufficient dependable ultimate strength to resist the design ultimate actions arising from the service loads factored so as to give an adequate margin of safety against collapse.

The design requirement can be written

$$U \leqq \phi S_u \tag{1}$$

where U is the design ultimate action at the section, ϕ the strength reduction factor, and S_u the nominal ultimate strength of the section. These terms are described more fully below.

In the building code of the ACI, the design ultimate actions are found from the service loads using load factors that are intended to ensure adequate safety. The design ultimate actions are found from the worst of the following load cases (the service loads are dead load D, live load L, and wind load W):

$$U = 1.4D + 1.7L \tag{2}$$

$$U = 0.75(1.4D + 1.7L + 1.7W) \tag{3}$$

where L may have its full value or zero, and

$$U = 0.9D + 1.3W \tag{4}$$

If design is to include earthquake loading E, the above load combinations apply, except that $1.1E$ is substituted for W. The nominal ultimate strength of a section of a member, S_u, is obtained from theory predicting failure behavior of the section and on given member dimensions and specified material strengths. The dependable ultimate strength is given by ϕS_u, where ϕ is the strength reduction factor to allow for approximations in calculations and variations in the material strengths, workmanship, and dimensions.

Values for ϕ recommended by the ACI code depend on the importance of the variable quantities. For the design of beams for flexure, $\phi = 0.9$. For the design of columns for flexure, if spirally reinforced, $\phi = 0.75$; otherwise, $\phi = 0.7$; except that for columns with axial compression approaching zero, $\phi \rightarrow 0.9$. For the design of members for shear and torsion, $\phi = 0.85$.

C. Trend toward Ultimate Strength Design

There has been a gradual trend toward the use of ultimate strength design for reinforced concrete since the 1950s. The issues behind this trend are numerous.

A main issue has concerned the modular ratio (ratio of the modulus of elasticity of the steel to that of the concrete) used in working stress design. The stress–strain behavior of concrete is time dependent, and the creep strains of concrete that occur during sustained stress may be several times the initial elastic strain. Therefore, any one value chosen for the modular ratio can be only a crude approximation. Creep strains can cause a substantial redistribution of stress in reinforced concrete sections, and this means that the stresses that actually exist at the service loads often bear little relation to the design stresses calculated by elastic theory for a particular value of the modular ratio. For example, the compression reinforcement in columns may reach the yield strength during the sustained application of service loads, although this occurrence is not evident from an elastic theory analysis using a normally recommended value for the modular ratio. Ultimate strength design does not require a value for the modular ratio.

Another issue concerns the economy of materials. Ultimate strength design utilizes reserves of strength resulting from a more efficient distribution of stresses allowed by plastic strains in the concrete and reinforcing steel, and at times it indicates the working stress method to be very conservative. For example, the steel reinforcement in a structural concrete beam usually reaches the yield strength at the ultimate load, but elastic theory may indicate a lower steel stress when the concrete is crushed. Also, ultimate strength design makes more efficient use of high-strength reinforcement, and smaller beam depths can be used without compression reinforcement.

A third issue centers around the proper assessment of risk against collapse. Elastic theory cannot reliably predict the ultimate strength of the reinforced concrete members because plastic strains are not taken into account. For structures designed by the working stress method, therefore, the exact load factor (ultimate load/service load) is unknown and varies from structure to structure. Also, ultimate strength design allows a rational selection of load factors, with values depending on the particular load combinations.

Finally, ultimate strength design is attractive because it also allows the designer to assess the ductility of the structure in the postelastic range. The available ductility is an important consideration, affecting the possible redistribution of bending moments in the design for gravity, earthquake, or blast loading.

The building code of the ACI permits the use of either working stress or ultimate strength design. However, the ACI code emphasizes the ultimate strength design approach. The working stress method is referred to as the alternative design method and has been relegated to an appendix of the code. In this alternative design method elastic theory is used to design beams for flexure, but factored-down ultimate strength equations are used to design members for all other actions. It is evident that this alternative method has been retained only in an attempt to keep what has been in the past the conventional design approach. Future ACI codes may omit this alternative procedure completely. Also of interest is a change in the terminology in ACI codes. The word *ultimate* rarely appears. For example, "ultimate strength" is written as "strength."

It is also worth noting that ultimate strength theory has been used for proportioning sections in the Russia and some other European countries for many years.

D. Limit State Design and Strength and Serviceability Design

It has been recognized internationally that the design approach for reinforced concrete ideally should combine the best features of ultimate strength and working stress design. This approach is desirable because, if the members are proportioned by ultimate strength requirements alone, there is a danger that, although the load factor is adequate, the cracking and the deflections at the service loads may be excessive.

In 1964 the European Concrete Committee produced its recommendations for an international code of practice for reinforced concrete. This document introduced the concept of limit state design, proposing that a structure be designed with reference to several limit states. The strength characteristics of the materials and the load factors were based on probability theory. The most important limit states were strength at ultimate load, deflections at service load, and crack widths at service load. This approach is being adopted by many codes.

The strength design procedure emphasized by the building code of the ACI could more properly be referred to as strength and serviceability design, since although the members are proportioned by strength requirements it is also required that the service load performance of the structure be shown to be satisfactory. Deflections at service load are controlled by using adequately stiff members

or by calculating the deflections by elastic theory and ensuring that limiting values are not exceeded. Crack widths at service load are controlled by using well-distributed arrangements of reinforcing steel. Thus the ACI strength and serviceability design approach is similar in principle to the European limit state design approach.

V. MODERN REINFORCED CONCRETE THEORY

Some examples of reinforced concrete theory are reviewed in this section. It is not intended that the review be comprehensive, but it will indicate some procedures for analysis and design.

A. Strength of Sections with Flexure and Axial Load

Figure 3a shows an eccentrically loaded, symmetrically reinforced concrete column section when the flexural strength is reached under the action of the ultimate load P_u and ultimate moment $M_u = P_u e$. The area of tension reinforcement is A_s and that of compression reinforcement is A'_s, and $A_s = A'_s$. The neutral axis depth is c. It is assumed that plane sections before bending remain plane after bending, and the tensile strength of the concrete is ignored. Values for the strain ε_c at the extreme compression fiber of the concrete at the flexural strength (maximum moment) have been found by experimental and analytical studies. The strain ε_c is commonly taken to be 0.0035 in European practice and 0.003 in North American and Australasian practice. The difference between these two ε_c values is of no great significance since most sections maintain nearly maximum moment capacity over curvatures corresponding to those strains. The actual curved shape of the concrete compressive stress distribution is commonly replaced by an equivalent rectangle, as originally suggested by Charles Whitney. Shown in Fig. 3a is the North American and Australasian equivalent rectangular stress distribution, which has a mean stress of $0.85 f'_c$, where f'_c is the concrete compressive cylinder strength, and a depth a, where the ratio a/c depends on the value of f'_c. Typically $a/c = \beta_1 = 0.85$ for $f'_c \leqq 30$ MPa (4350 psi) and is smaller for higher values of f'_c. These parameters were found so that the actual and equivalent distributions of concrete compressive stress have the same area and position of centroid. European practice also proposes an equivalent compressive stress distribution, generally either parabolic–rectangular or rectangular in shape.

The force and moment equilibrium equations that give the ultimate load and ultimate moment in terms of the internal stresses are

$$P_u = 0.85 f'_c ab + f'_s A'_s - f_s A_s \tag{5}$$

$$M_u = P_{ue} = 0.85 f'_c ab(0.5h - 0.5a) + f'_s A'_s(0.5h - d') + f_s A_s(0.5h - d') \tag{6}$$

where the strains in the compression and tension reinforcement are obtained from the strain diagram of Fig. 3a as

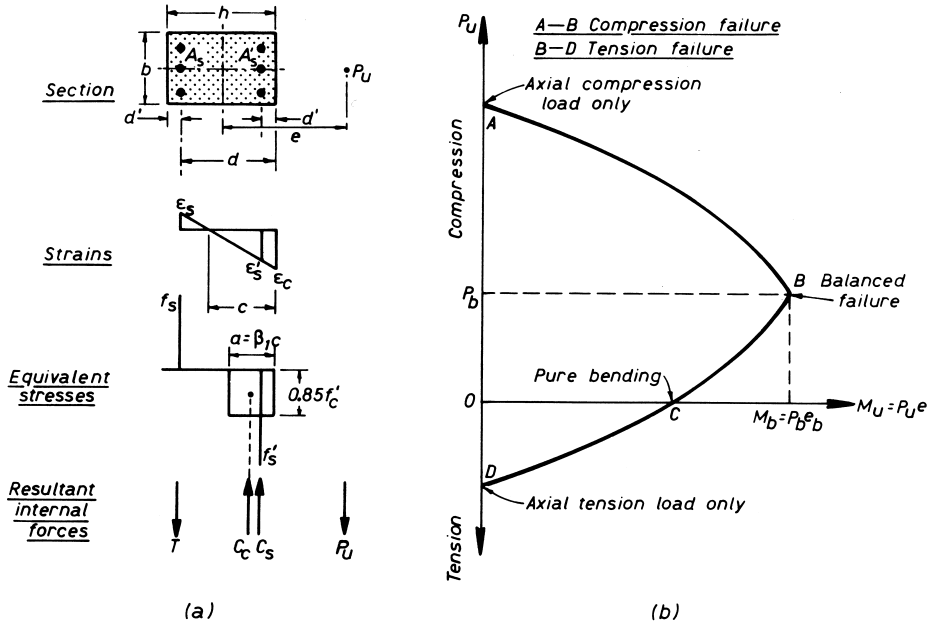


FIGURE 3 (a) Reinforced concrete short column at ultimate eccentrically applied load; (b) interaction diagram showing combinations of load and moment that cause failure.

$$\varepsilon'_s = \varepsilon_c(a - d'\beta_1)/a \quad (7)$$

$$\varepsilon_s = \varepsilon_c(d\beta_1 - a)/a \quad (8)$$

and the stresses of f'_s and f_s in the compression and tension reinforcement are obtained from those strains and the stress–strain relationships for the steel.

For a given ultimate load P_u, the depth of the equivalent rectangular stress distribution a can be found from Eqs. (5), (7), and (8) and the stress–strain relations for the steel, and then the ultimate moment M_u corresponding to that load can be found from Eq. (6). Alternatively, the P_u and M_u values corresponding to a range of a values can be found from Eqs. (5) to (8) and the stress–strain relationships for the steel, resulting in the interaction diagram shown in Fig. 3b. Any combination of load and moment that gives a point on the interaction diagram will cause failure of the column; any combination of load and moment giving a point within the area of the interaction diagram can be supported without failure.

If the ultimate load P_u is smaller than a particular load P_b (see Fig. 3b), a tension failure occurs in which the tension reinforcement yields ($f_s = f_y =$ steel yield strength), followed eventually by crushing of the concrete. If $P_u > P_b$, a compression failure occurs in which the tension steel does not yield ($f_s < f_y$). This behavior is a consequence of the increase in neutral axis depth and hence the reduction in tension steel strain, with increase in compressive load. The ductility displayed by a column at the ultimate load can be considerable during a tension failure but negligible during a compression failure. The ductility during a compression failure can be improved by the presence of closely spaced transverse reinforcement in the form of hoops or spirals, which improve the compression strain capacity of the concrete.

When the column is loaded concentrically in compression, the axial load strength (point A in Fig. 3b) is the sum of the compressive strengths of the concrete and steel. The axial tensile strength as a tie member (point D in Fig. 3b) is simply the yield strength of the steel multiplied by the total area of the longitudinal reinforcement.

B. Strength of Sections with Flexure

If the member acts as a beam, the ultimate moment M_u is found from Eqs. (5) to (8) with $P_u = 0$ substituted (point C in Fig. 3b). This gives, for a beam without compression reinforcement ($A'_s = 0$) and for the case of a tension failure ($f_s = f_y$), an ultimate moment of

$$M_u = 0.85 f'_c ab\left(d - 0.59\frac{A_s f_y}{f'_c b}\right) \quad (9)$$

When compression reinforcement is not present in a beam, a ductile failure can be achieved by limiting the quantity of tension reinforcement present to ensure that it reaches the yield strength before the concrete crushes.

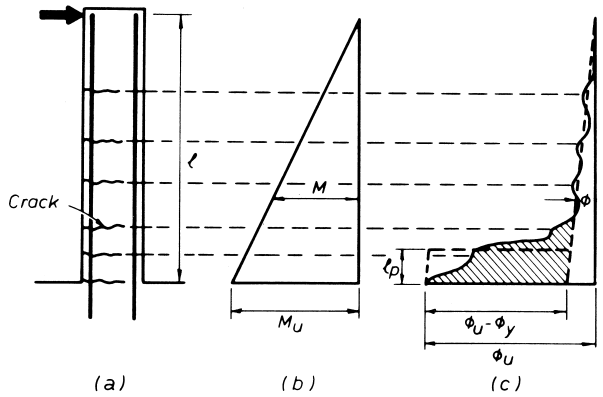


FIGURE 4 Member with ultimate moment and curvature reached. (a) Member; (b) bending moment diagram; (c) curvature diagram (full lines, actual; dashed lines, idealized; shaded area, plastic rotation).

C. Rotations and Deflections of Members Due to Flexure

Figure 4 shows a reinforced concrete member that, under transverse load, has reached the ultimate moment at the critical section. The region of inelastic curvature is spread over a length of member, this region being at least that over which the bending moment exceeds the moment when the tension steel first yields in the section. Local peaks of curvature occur at the cracks, because between the cracks the flexural rigidity of the member is increased as a result of the tension carried by the uncracked concrete. The actual distribution of curvature can be idealized as having elastic and inelastic regions (Fig. 4c).

In the elastic range the effective (average) flexural rigidity of a cracked reinforced concrete member is partway between the cracked transformed section and the gross (uncracked) section value. The smaller the first cracking moment is compared with the maximum moment in the member, the more closely the effective flexural rigidity of the sections approaches the cracked transformed section value. That is, the tension stiffening caused by the uncracked concrete between cracks becomes less important at high moments. For example, if the moment at first cracking is one-half of the maximum moment in a typical member at a particular load, the assumption of all concrete cracked will lead to less than 15% error in the deflection calculated in the elastic range.

The shaded area in Fig. 4c gives the inelastic rotation that can occur at a "plastic hinge" in the vicinity of the critical section. The shaded area can be replaced by an equivalent area of height $\phi_u - \phi_y$ and length l_p, where ϕ_y is

the curvature when the tension reinforcement first yields, ϕ_u the curvature when the compressed concrete reaches its maximum strain capacity, and l_p the equivalent length of the plastic hinge. When diagonal tension (inclined) cracks are present in a plastic hinge region, the inelastic curvature is spread farther along the member than the bending moment diagram implies. Also, any deterioration of bond between concrete and reinforcement occurring in a plastic hinge region tends to spread the yielding. Therefore, values for l_p have been proposed on the basis of experimental evidence. The value for l_p is at least 0.4 of the member depth and generally greater. The available plastic hinge rotation is given by the shaded area of Fig. 4c and is

$$\theta_p = (\phi_u - \phi_y) l_p \tag{10}$$

Deflections can be calculated from the curvatures.

D. Strength of Members with Shear

The combination of flexure and significant shear force in reinforced concrete members causes inclined cracks, referred to as diagonal tension cracks. Figure 5a shows part of a simply supported beam without shear reinforcement to one side of a diagonal tension crack. The shear force V is resisted by a shear force V_{cz} carried across the compression zone, a dowel force V_d transmitted across the crack by the flexural reinforcement, and the vertical component V_{av} of the force due to the shear stress transmitted across the inclined crack by means of interlocking of the aggregate particles. Thus

$$V = V_{cz} + V_d + V_{av} \tag{11}$$

If the resultant aggregate interlock force is assumed to pass through the intersection of the forces C and V_{cz} and if the contribution of V_d is neglected as being unreliable because of possible longitudinal splitting of concrete along the bars, then $M = T\,jd$, which means that the tensile force T at section 2 is due to the moment M at section 1. This shift in magnitude of the tensile force causes some spread of flexural yielding along the member and influences the plastic hinge length, as discussed previously. It must also be taken into account when the curtailment of longitudinal bars is being considered along spans. For example, a bar should be extended by at least a beam effective depth beyond the point at which the bending moment diagram indicates that the remaining bars can carry the moment.

In reinforced concrete beams of normal proportions without shear reinforcement, shear failure can occur by breakdown of the concrete cantilevers between the diagonal tension cracks due to bond forces arising from beam action. Although extensive analytical and experimental work has improved the understanding of the behavior of reinforced concrete members subjected to shear, particularly in recent years, no well-established theoretical equations for the ultimate shear strength have emerged. Therefore, relatively simple semiepirical design equations for the shear strength have been adopted by most codes. The shear force carried by the resisting mechanisms of Fig. 5a at shear failure can be written $v_c b_w d$, where b_w is the width of the beam web, d is the effective depth of the beam, and the nominal shear stress v_c is a function of the major variables affecting the shear resistance expressed in Eq. (11). These major variables are the tensile strength of the concrete, which is a function of the square root of the concrete compressive strength; the crack control, as measured by the ratio of longitudinal tension steel $A_s/b_w d$; and the shear span/depth ratio, as measured by M/Vd. A commonly assumed simplified value of v_c used in North America and Australasia is $2\sqrt{f'_c}$ psi ($0.17\sqrt{f'_c}$ MPa), where f'_c is the concrete compressive cylinder strength in pounds per square inch (or megapascals). Axial compression (as in columns) increases the shear carried by the concrete because of better crack control and the larger neutral axis depth.

Shear reinforcement, commonly in the form of steel stirrups, can carry shear force by truss action (Fig. 5b). In the truss analogy, first postulated by E. Mörsch in 1908, the web of the truss is considered to consist of stirrups in tension and concrete struts running parallel to the diagonal tension cracks. The top and bottom chords of the truss are formed by the compressed concrete due to flexure and the

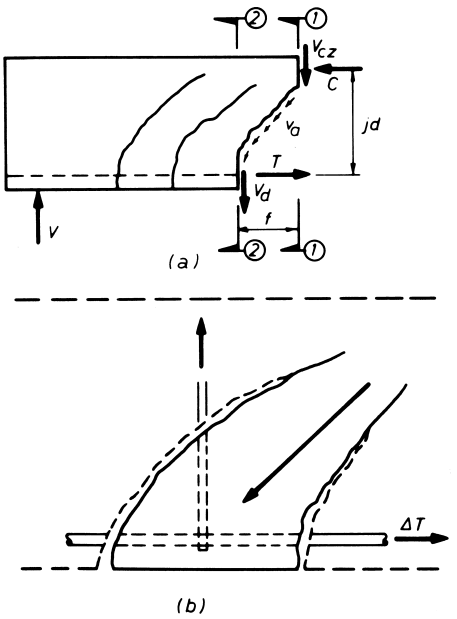


FIGURE 5 Shear-resisting mechanism in a beam. (a) Internal forces resisting shear without web reinforcement present; (b) concrete cantilever acting as strut when web reinforcement is present.

longitudinal tension reinforcement. The number of stirrups perpendicular to the axis of the member crossing an assumed 45° crack is approximately d/s, where s is the stirrup spacing. Therefore, the shear force carried by perpendicular stirrups is approximately $f_y A_v d/s$, where f_y is the stirrup steel yield strength and A_v the area of each stirrup.

The ultimate shear strength of a member with perpendicular stirrups is found by summing the contribution from the concrete mechanisms and the shear reinforcement as

$$V_u = v_c b_w d + f_y A_v d/s \tag{12}$$

E. Bond and Anchorage

The attainment of satisfactory bond between concrete and steel reinforcement is essential for the satisfactory performance of reinforced concrete. Local bond stresses are developed by the change in the force in reinforcing bars along their length. It has been observed, however, that, provided sufficient anchorage length is available at the ends of bars, failure originating from local bond stress does not occur. The anchorage requirement is that a bar must extend beyond any section by sufficient distance to transmit the force at that section to the concrete by bond. Existing code requirements for anchorage are entirely empirical. A straight anchorage length of 40 bar diameters, commonly used for many years for plain, round mild steel bars, has been much reduced by the use of bars with a deformed surface, which improves the bond resistance. For example, according to the building code of the ACI, the basic development length for a deformed bar of 1-in. (25.4-mm) diameter in concrete of compressive strength $f'_c = 3000$ psi (20.7 MPa) is 23 bar diameters if the steel yield strength f_y is 40,000 psi (275 MPa) or 34 bar diameters if $f_y = 60{,}000$ psi (414 MPa).

A bond failure of deformed bars generally occurs by concrete splitting due to the wedging action of the deformations, and hence the required anchorage length is a function mainly of the bar force at yield, the tensile strength of the concrete, the position of the bar in the member (top-cast bars perform adversely when compared with bottom-cast bars because of the effect of water gain and sedimentation under the top bars), lateral spacing between bars, concrete cover thickness, and the quantity of confining transverse steel present. Hooks formed at the end of bars can significantly reduce the required anchorage length.

F. Cracking at Service Load

The occurrence of cracks in reinforced concrete is inevitable because of the low tensile strength of concrete. With high service load steel stresses, particularly as a result of the use of high-strength reinforcement, significant cracking may occur at the service load. The cracking of a reinforced concrete structure at the service load should not be such as to spoil the appearance of the structure or lead to corrosion of the reinforcement. For example, the distribution of reinforcement permitted by the building code of the ACI is aimed at ensuring that the surface crack widths at service load will not exceed 0.016 in. (0.41 mm) in the case of members with interior exposure and 0.013 in. (0.33 mm) in the case of members with exterior exposure.

A number of semitheoretical and empirical approaches have been developed for the determination of the most probable maximum surface width of flexural cracks. Simplistically, the crack width can be postulated to be the elongation of the reinforcement between two adjacent cracks. Statistical studies of test results have shown crack widths to be primarily a function of the stress in the steel, the thickness of concrete cover, the distribution of the steel reinforcement, and the relative distances from the neutral axis of the reinforcing steel and the extreme concrete tension fiber. Crack widths are also influenced by shrinkage of concrete and other time-dependent effects and by repeated loading. Crack width measurements are inherently subject to large scatter, even in careful laboratory work. Thus great accuracy cannot be expected from existing equations for crack widths.

The best crack control at the service load is obtained when reinforcing bars are well distributed in the zone of concrete tension. Thus a number of small-diameter bars are preferable to the use of a few larger-diameter bars. It should also be emphasized that concrete with a low water/cement ratio and a reasonable thickness of well-compacted good-quality concrete are also essential for durable concrete structures.

G. Determination of Design Actions at Ultimate Load using Linear Elastic Structural Analysis

The main approach used in design to determine the bending moments and forces in a reinforced concrete structure at the ultimate load is linear elastic structural analysis. The advantages of using this approach are that the moments and forces can be found using relatively simple and well-established structural theory, that the stresses in the concrete and reinforcement at the service loads are kept as low as possible, thus minimizing the width of cracks and deflections, and that only a small amount of moment redistribution is required at the ultimate load. The last aspect means that the plastic rotations required at the critical sections will be small. Some redistribution of moments will almost always be necessary because, unless the structural analysis is based on the final complex distribution

of flexural stiffnesses after cracking, redistribution of moments will have to occur before all the critical sections can attain their ultimate flexural strength.

Most codes also allow the moments and forces derived from linear elastic structural analysis to be modified for a moderate amount of moment redistribution, the amount permitted being limited to ensure that the crack widths at the service loads are not excessive and that the critical sections are adequately ductile to achieve the required redistribution at the ultimate load. For example, the negative moments at the supports of continuous reinforced concrete flexural members may be increased or decreased by up to 20%, according to the building code of the ACI, the exact amount depending on the section ductility, provided that the positive moments within the spans are adjusted to maintain static equilibrium between internal forces and external loads.

H. Limit Design of Frames

The limit design approach for reinforced concrete frames, analogous to the "plastic theory" design of structural steel frames, allows any distribution of bending moments at the ultimate load to be used, provided that the plastic rotation capacity is sufficient to permit the assumed distribution of moments to be developed, and hence that the collapse mechanism can form at the ultimate load, and that the cracking and deflections at the service load are not excessive.

The major difficulty with this approach lies in the calculation of the required plastic rotations. This calculation is necessary because reinforced concrete sections, unlike structural steel, may in some cases fail in a brittle manner at a lower ultimate load before the collapse mechanism develops. A method that appears to have the most promise for reinforced concrete is that of A. L. L. Baker. To calculate the plastic rotations actually required at the plastic hinges to develop a particular limit moment diagram. Baker considers the structure to be in the state immediately before ultimate load, where sufficient plastic hinges have formed to make it statically determinate and the last plastic hinge or hinges are about to form and convert the structure to a collapse mechanism. In this state the structure has known ultimate moments at the plastic hinges and at the plastic hinge or hinges about to form, and the members are in the elastic range between the hinges. The rotations required of the plastic hinges can be calculated by the flexibility coefficient method. For a structure that is n times statically indeterminate with known ultimate moments $M_1, M_2, M_3, \ldots, M_n$ at the plastic hinges, the plastic rotation θ_i at hinge i is

$$-\theta_i = \delta_{i0} + \sum M_k \delta_{ik} \tag{13}$$

where δ_{i0} is rotation at hinge i due to the external ultimate loads acting with zero hinge moments, M_k the ultimate moment at hinge k, and δ_{ik} the rotation at hinge i due to $M_k = 1$ acting alone at hinge k. Thus the plastic rotations required for a selected limit moment diagram can be calculated and checked against the permissible rotations computed from Eq. (10) to ensure that full moment redistribution can take place. Baker's method is unwieldy in all but the simple cases, however, and although it has been recommended by some design codes, it can hardly be regarded as a design office technique in its present form.

I. Limit Design of Slabs

Reinforced concrete slabs are usually lightly reinforced because the span/depth ratios are normally governed by requirements of adequate stiffness to avoid excessive deflections. Hence the sections generally have sufficient plastic rotation capacity to allow full moment distribution to take place at ultimate load. It is therefore commonly assumed that limit design can be applied to reinforced concrete slabs without the need to check whether sufficient plastic rotation capacity is available. Lower- or upper-bound approaches can be used to obtain the design ultimate moments, assuming that a flexural failure occurs. The shear strength is checked separately.

The lower-bound method of reinforced concrete slab design involves finding a distribution of moments that satisfies the equilibrium equations and the boundary conditions when the ultimate load is applied and providing sufficient reinforcement to carry these moments without the ultimate moment being exceeded anywhere in the slab. This procedure can be fulfilled by using in design the distribution of moments given by thin-plate elastic theory. Alternatively, a commonly used lower-bound approach is Hillerborg's strip method. The simple strip method obtains the distribution of moments and shears by replacing the slab by two systems of strips, normally running in two directions at right angles, which share the external loading. The strips are not considered to carry any load by torsion, and the design bending moments are found by simple statics. A. Hillerborg has also introduced the advanced strip method, which features rectangular corner-supported elements, as well as triangular and rectangular edge-supported elements, for use in the design of beamless slabs.

The upper-bound method of reinforced concrete slab design is known as yield line theory and is due mainly to K. W. Johansen. Collapse mechanisms for slabs are known as yield line patterns and are composed of segments of slabs separated by lines of plastic hinging. The ultimate load of the slab is calculated from the postulated yield line pattern, and the position of the yield lines that gives

the smallest ultimate load is sought. The correct yield line patterns for all common cases are now well known. In design the procedure is to provide adequate reinforcement to ensure that the required ultimate load can be reached.

J. Complete Behavior of Structural Systems

The full analysis to determine the complete behavior of reinforced concrete structures at all levels of loading, where flexural deformations predominate, can be undertaken with the aid of a computer with large storage. For example, a given frame or slab system can be divided into small elements of the thickness of the member or slab, each with its known moment–curvature relation, including the effect of tension stiffening between the cracks. The distribution of moments throughout the structure as the external load is increased incrementally from zero can be determined using the stiffness method. At each level of external load an iterative procedure is used to obtain the distribution of moments that is compatible with the boundary conditions, the equilibrium equations, and the stiffnesses of the elements as given by the nonlinear moment–curvature relations. Eventually, with increasing load, the elements at the critical sections commence to reach their ultimate flexural strength, and plastic hinges spread throughout the structure. The ultimate load is reached either when a collapse mechanism forms or when a brittle failure occurs at a critical section due to lack of plastic rotation capacity. A computer program capable of this analysis procedure could be used in design only on a trial-and-error basis, but it would be a powerful tool for evaluating or checking structural performance over the full range of loading, including behavior at the service and ultimate loads. Geometric changes in the structure during loading could also be included in such a program to allow for the effect of deflections on internal actions and to indicate any instability effects. The full analytical approach is feasible now and some special-purpose computer programs are available, but are not commonly used.

VI. PRESTRESSED CONCRETE

A. Introduction

Prestressed concrete is defined as concrete that has been prestressed so that the induced internal actions counteract the external loading to a desired degree. Prestressed concrete members can be placed in one of two categories: pretensioned or posttensioned. Pretensioned prestressed concrete members are produced by stretching steel tendons between external buttresses before the concrete is placed. When the concrete has reached the required strength, the tendons are released from the buttresses and the prestressing force is transferred to the concrete of the member by the bond between the tendons and the concrete. In the case of posttensioned prestressed concrete members, steel tendons are stretched in ducts in the concrete, after the concrete has hardened and gained sufficient strength, by jacking against the concrete member itself, and then the tendons are anchored at the ends of the member.

The general principles of prestressing concrete were known for many years before it became a practical method of construction. Early attempts to prestress concrete were unsuccessful, mainly because high-strength steel was not used for the tendons and the low initial steel stress was lost with time as a result of the shrinkage and creep of concrete. The general concepts of prestressed concrete were first formulated in the 1880s by Doehring of Germany and Jackson of the United States. The modern development of prestressed concrete is credited to Freyssinet of France, who clearly showed the effect of concrete creep and shrinkage and in approximately 1928 demonstrated that by the use of high-tensile steel the losses in prestress could be kept down to reasonable proportions. The practical development of pretensioning was due mainly to Hoyer of Germany. The wide application of prestressed concrete became possible when reliable and economical methods of tensioning and end anchorage were developed. In 1939 Freyssinet developed double-acting jacks and conical end anchorages. In 1940 Magnel of Belgium developed metal wedges.

B. Basic Behavior

The usual concept of prestressed concrete is that of a material weak in tension that is compressed by prestress to avoid the occurrence of tension due to external loading. If the tensile strength of the concrete is not exceeded, the stresses can be determined by standard elastic theory for uncracked sections. Figure 6a shows a prestressed concrete beam with a symmetrical cross section and with a prestressing tendon placed in a curved profile with variable eccentricity e with respect to the centroid of the section of the member. The extreme fiber concrete stresses at section 1 due to prestress alone are

$$f = \frac{F}{A} \pm \frac{Fe}{Z} \tag{14}$$

where F is the prestressing force, A is the section area of the member, and Z is the section modulus of the member. When external loading is applied, the extreme fiber concrete stresses due to the external moment M at section 1 are $\pm M/Z$, and these stresses are added to those due to prestress. Figure 6c shows the distribution of concrete stresses at section 1 due to prestress alone and due

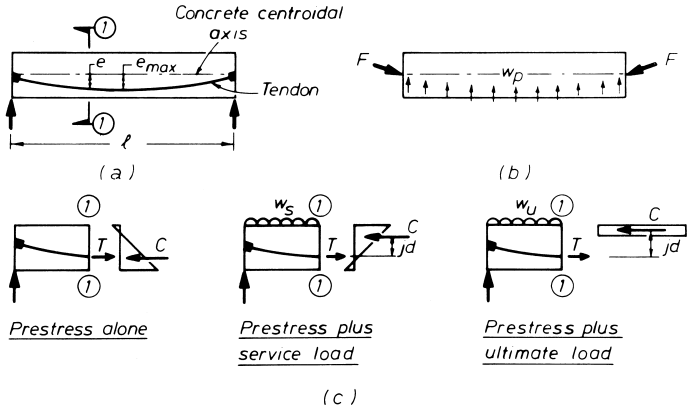


FIGURE 6 Behavior of a prestressed concrete member. (a) Beam and tendon; (b) forces acting on concrete due to parabolic tendon alone; (c) internal couple at a section when external moment at that section increases from zero to the ultimate moment.

to prestress plus service load (before cracking). In design a small tensile stress in the concrete, insufficient to cause cracking, is often allowed.

Another concept of prestressed concrete is that of a combination of steel and concrete in which the tensile force in the prestressing steel and the resultant compressive force in the concrete form an internal couple that resists the bending moment due to the external loading. Figure 6c illustrates the concept. The compressive force in the concrete C acts at the centroid of the compressive stress distribution. The tensile force in the tendon is T. For equilibrium, C must equal T at all stages. When there is no external loading on the member, C and T have the same line of action; that is, the internal couple is zero. When external loading is applied, the centroid of the compressive stress distribution shifts up the section because of the change in concrete compressive stress distribution, and the internal couple is equal to Cjd or Tjd, which exactly equals the bending moment acting at the section due to the external loading. As the external load varies within the service load range, T remains practically constant in a prestressed concrete member because a high initial steel stress is used and the increase in steel stress due to bending strains in the concrete represents only a small percentage of change in steel force. Hence as the service load varies, C and T may be considered to remain constant, but jd varies. By comparison, in a reinforced concrete member, as the external load varies within the service load range C and T vary but jd remains constant. If the external load exceeds the service load, the concrete will crack when it reaches its tensile strength. Cracking will result in an increase in the tensile stress in the steel and in the compressive stress in the concrete as the stresses readjust to carry the external moment. With further loading the concrete and steel enter the inelastic range and eventually the ultimate moment of the section is reached (Fig. 6c).

Reinforced concrete cannot effectively use very high tensile steel because the high steel stresses associated with the efficient use of such steel mean that large tensile strains must be induced in the steel, and this would cause the surrounding concrete to crack seriously at service load. In prestressed concrete this problem is overcome because the steel is prestretched. Also, high-strength concrete can be used to advantage. Hence prestressed concrete makes effective use of high-strength materials.

Prestressed concrete can also be regarded as an attempt to balance the external loading by the actions induced by the prestressing tendon. Figure 6a shows a simply supported beam prestressed by a tendon with force F and a parabolic profile. The tendon applies an upward uniform load w_p per unit length to the concrete (Fig. 6b), since it is forced to remain in the shape of a parabola, and from simple statics

$$w_p = 8Fe_{max}/l^2 \tag{15}$$

where e_{max} is the eccentricity of the tendon at midspan. Thus external loading on a member can be balanced by the upward loading from the tendon. If the external load is exactly balanced by the tendon, the beam will not be deflected and the concrete will be subjected to uniform stress F/A, since the moment due to external loading will be exactly balanced by the moment due to the prestress along the beam. Similarly, the net shear force will be zero. In practice it is not generally possible or desirable to balance the total service loads fully, mainly because such a design would be uneconomical, requiring a large amount of prestress, and because the live load fluctuates between zero and its full value during service conditions. However, by suitable design of the prestressing force, it is possible to offset a good deal of deflection due to external load. This allows the use of more slender spans in prestressed concrete than is possible with reinforced concrete. The load

balance approach was introduced by T. Y. Lin in 1963. It is a useful concept for design, especially in the case of posttensioned flat slab floor systems.

C. Statically Indeterminate Structures

When statically indeterminate structures are posttensioned during construction, the members deform elastically and secondary prestressing moments may arise as a result of the restraints offered by the support conditions. The secondary prestressing moments so induced are often not secondary in magnitude and must be added to the primary prestressing moments due to tendon eccentricity to find the resulting (total) prestressing moments. Concordant tendon profiles, for which the secondary prestressing moments are zero, can be found for continuous beams but are often uneconomical in design since full use cannot be made of the available section eccentricity. However, efficient nonconcordant (general) tendon profiles can readily be incorporated in design. In the case of a general tendon profile the resulting prestressing moment can be calculated directly, using a method of structural analysis, from the equivalent loads acting on the structure caused by the prestressing tendons. The equivalent loads are due to end anchorages and to the angular changes of the tendons along the members. Much of the pioneering theoretical work on the analysis of statically indeterminate structures was that of Y. Guyon, whose volumes still form a major reference source on prestressed concrete.

D. Design Approaches for Prestressed Concrete

Prestressed concrete structures are generally designed with reference to the limit states of allowable stresses and deflection at service loads using elastic theory, and adequate ultimate load using ultimate strength theory. The allowable concrete tensile stress at service load may be zero or a small value that will not lead to cracking, according to the concept of prestressing advocated by Freyssinet.

An alternative design approach that has been advocated in more recent years permits tensile stress to develop in the concrete, which may lead to cracking when the full service live load is applied. The acceptance of cracked prestressed concrete sections at service loads leads to the concept of partially prestressed concrete, as advocated by Abeles. The use of a lesser amount of prestressing steel in a partially prestressed concrete structure generally means that some nonprestressed steel reinforcement must also be present to help control the cracking at service load and to provide the additional flexural strength necessary to achieve an adequate ultimate load.

VII. FERROCEMENT AND FIBER-REINFORCED CONCRETE

Providing tensile resistance in concrete has conventionally involved embedded steel reinforcing bars or prestressing tendons. Nevertheless, significant progress has been made in obtaining a composite material formed from cement mortar or concrete acting with layers of small-diameter wire mesh or with short, randomly oriented fibers.

Ferrocement is the name given to the composite material formed by plastering several layers of steel wire mesh with cement mortar. The earliest recorded use of ferrocement was Lambot's concrete rowing boat, which was built in France in 1848. For many years ferrocement continued to be used mainly for boats. The most important developments in ferrocement were due to P. L. Nervi in Italy, who constructed several vessels in the early postwar years and was the first to make significant use of ferrocement in buildings. In recent years ferrocement has been used widely for boats and architectural elements of buildings. The close spacing of small-diameter wires results in very well controlled cracking of concrete, and only a small thickness of mortar cover is necessary for protection against corrosion. Also, the impact resistance is high. Complicated shapes can be readily constructed because the mortar is plastered directly onto the wire mesh, and formwork is unnecessary.

The idea of endowing concrete with a greater tensile strength and ductility by the use of fibers was proposed at the turn of the century. However, the practical use of short, randomly oriented fibers in concrete is relatively new. Much of the pioneering work on steel fiber reinforcement was conducted by J. P. Romualdi and G. B. Batson in the United States in the late 1950s and early 1960s. Ordinary glass fibers are unsuitable, since glass is attacked by the highly alkaline environment of concrete, but alkaline-resistant glass fibers have now been developed. The failure mechanism of fiber-reinforced concrete involves mainly fiber pull-out, and it is possible to improve the bond strength by deforming the fibers in various ways. The maximum size and quantity of aggregate in the concrete must be limited, because the fibers can only reinforce the matrix. The presence of fibers reduces the workability of the mix, and hence a maximum of $\sim$3% by volume of fibers can be added when mixing. The inclusion of fibers improves the toughness of concrete, and the load-carrying capacity can be sustained at relatively large tensile and compressive strains, well beyond peak load conditions. Fiber-reinforced concrete has possible applications when the toughness and energy-absorbing characteristics are important, such as for pavements and linings. Glass-fiber reinforced concrete has also been used as spray-on cladding for structures.

VIII. GENERAL APPLICATIONS OF STRUCTURAL CONCRETE

A. Buildings

Concrete floors are generally either of all-cast-in-place construction or formed from precast members, generally acting compositely with cast-in-place concrete.

Three basic types of all-cast-in-place reinforced concrete floors are shown in Fig. 7. Beams may or may not be present. The flat plate (Fig. 7a) with uniform thickness is the simplest form of beamless slab construction. The flat slab (Fig. 7b) has capitals at the tops of the columns and/or drop panels, which are thickened areas of slab surrounding each column, in order to provide the additional slab stiffness and shear strength necessary for larger spans and heavier live loads. The two-way slab (Fig. 7c) is supported on beams in two directions on the column lines. The beams can be replaced by walls. If the span between the column lines is large, secondary beams may also be present between the column lines in order to reduce the spans of the slabs. If the span of the slab in one direction is more than twice that in the other direction, the greater part of the slab loading will be carried in the direction of the short span, and the slab can be regarded as a one-way slab. One-way slab action will also occur if the beams or walls are present only in one direction.

Waffle slabs, containing recesses formed in the lower surface, can be used to reduce the dead load. Waffle slabs can also be used as a means of increasing the effective depth of a slab without an accompanying increase in dead load. Hollow clay tiles or other fillers can be employed in place of recesses as a further variation of a waffle slab.

The use of prestressing in flat slab floors has become common in many parts of the world, the posttensioned tendons allowing excellent crack and deflection control and longer spans.

Floor slabs may be cast in place on formwork in their final position. Some flat plates are built using lift–slab construction, in which all the slabs are cast in a stack at ground level and then lifted into their final positions after being cured.

The choice of slab type depends on the magnitude of the live load, the spans, the cost of materials and labor during construction, and the local preference. Cast-in-place two-way slab floors are more labor intensive during construction but require less reinforcing steel because of the large effective depths of the beams compared with beamless floors with the same spans between columns and live loads.

Precast concrete one-way floors consisting of double-T, hollow core plank, or other sections spanning beams or walls can be used as an alternative to an all-cast-in-place floor. Precast floors are generally used with a cast-in-place concrete topping slab. The precast units are either of reinforced or of pretensioned prestressed concrete. Precast floors are widely used in many parts of the world.

Concrete floors also connect parts of the vertical structural system to form a structure capable of transferring both gravity and horizontal loads to the foundations. The vertical structural system commonly comprises either continuous frames consisting of columns and beams rigidly connected together, or structural walls, which support the floors. Buildings can include various combinations of these vertical structural systems. Examples of these basic

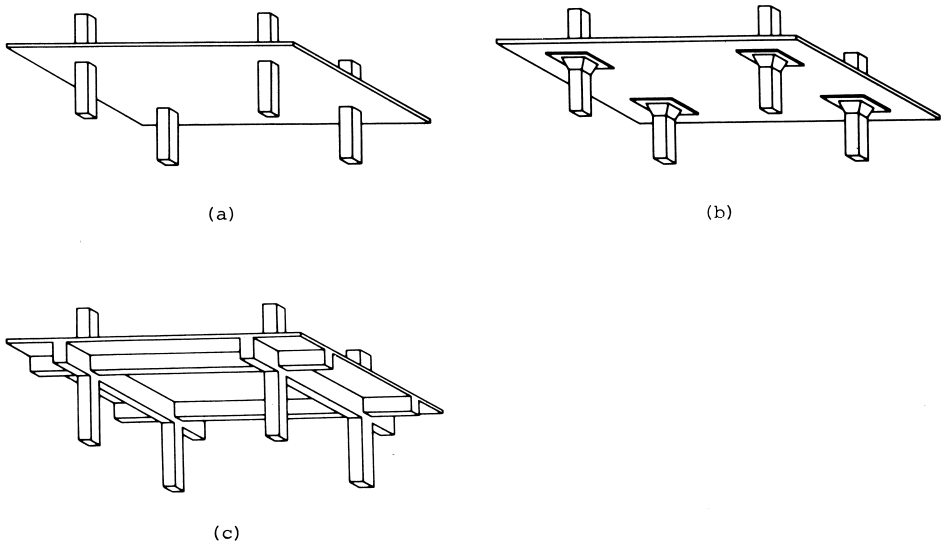


FIGURE 7 Types of all-cast-in-place reinforced concrete floors. (a) Flat plate; (b) flat slab; (c) two-way slab.

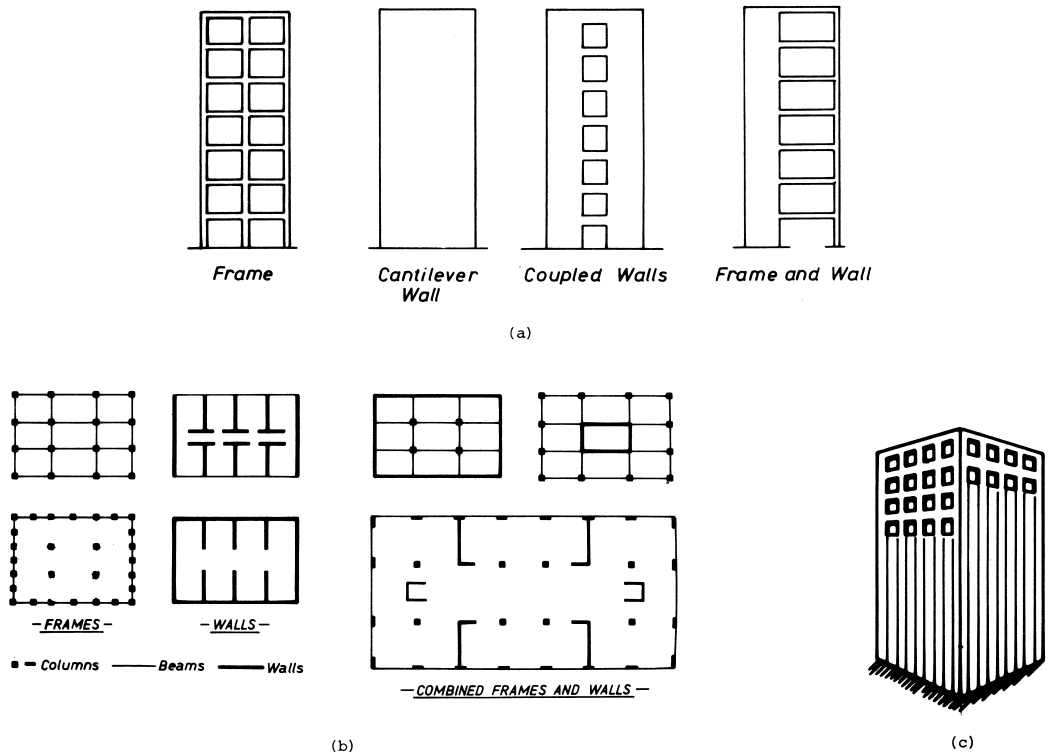


FIGURE 8 Basic structural concrete systems for resisting gravity and horizontal loads. (a) Frames and walls; (b) some building plans; (c) framed tube.

structural systems and combinations of them are shown in Fig. 8a,b.

A problem with relying entirely on frame action for strength and stiffness is that interstory horizontal deflections due to horizontal wind or earthquake loading may be significant. Codes generally require calculations to check that the horizontal drift is within tolerable limits, which should ensure that the discomfort caused to occupants and the damage caused to nonstructural components are not excessive. Structural walls, because of their greater stiffness, can be effectively combined with frames to help control behavior under horizontal loads. Indeed, several well-placed structural walls in a building can resist a large portion of the horizontal load and limit horizontal drift, leaving the more flexible frames to carry the majority of the gravity loading. Thus, combined systems of walls and frames can often be used to advantage.

Some buildings have frames with closely spaced columns around the perimeter, which simulate a perforated rectangular hollow tube, sometimes referred to as a framed tube (Fig. 8c). Cantilever walls and coupled walls can also make up tubes. Tubes are an efficient means of resisting lateral loading since part of the tube wall at right angles to the direction of lateral loading can act as the "flange" to the tube wall in the direction of the loading. An effective structural system for tall buildings is a tube within a tube, consisting of an inner tube formed by the walls enclosing the service core and an outer tube formed by frames with closely spaced columns around the perimeter of the building.

Design for gravity and wind loading is essentially a matter of ensuring that the building structure can withstand the greatest likely load applications without failure. However, it would be uneconomical to design a structure to withstand the greatest likely horizontal earthquake loading without damage. Dynamic analyses of building structures responding elastically to ground shaking recorded during severe earthquakes have shown that the theoretical response inertia loads are generally significantly greater than the static design horizontal loads recommended by codes. Hence structures designed for the horizontal earthquake loads normally recommended by codes can survive very severe earthquakes only if they have sufficient ductility to absorb and dissipate seismic energy by inelastic deformations. Hence buildings constructed in earthquake zones require special attention in design to ensure that in the event of a major earthquake they can deform in the inelastic range in a ductile manner. In practice, this means designing ductile plastic hinge regions in members and ensuring that failure does not occur by shear, bond, or column

failure. In frames a strong column–weak beam concept is used, and potential plastic hinge regions are made ductile by ensuring that the tension reinforcement content is not too high, that adequate compression steel is present, and that sufficient transverse reinforcement is available to prevent shear failure and compression bar buckling and to confine the concrete.

Conventionally, most concrete building structures have been of cast-in-place monolithic construction. However, the cost of skilled labor to build formwork and the need for speed of construction have made the use of precast concrete components, prefabricated off site, an attractive alternative to cast-in-place construction. The precast concrete components have followed the traditional patterns of beams, slabs, and columns. More recently they have also taken the form of precast bearing walls or panels and precast slab units. Care is needed in incorporating precast units into the design to ensure that the connections between them have both adequate strength and ductility. The progressive collapse of the Ronan Point building in England in 1968 after a domestic gas explosion was due to poor connections between precast concrete walls and floors and a poor design concept.

The first tall concrete building to be built with a reinforced concrete frame was the 16-story Ingalls Building, which was constructed in 1903 in Cincinnati, Ohio. The world's tallest concrete building has been Water Tower Place, built in 1975 in Chicago. This building is 859 ft (262 m), 76 stories tall and has high-strength 9000-psi (62-MPa) concrete at its base. The tower of the building consists of structural walls and closely spaced perimeter columns forming a tubular structure. Lightweight concrete has been used to reduce the dead weight of some tall concrete buildings. Figure 9 shows the 715-ft (218-m), 60-story-tall One Shell Plaza building built in Houston, Texas, in 1969 using high-strength 6500-psi (45-MPa) lightweight concrete. The CN Tower, in Toronto, completed in 1975, is the world's tallest free-standing tower (Fig. 10). The CN Tower serves for communications and observations and has a concrete structure 1464 ft (446 m) high topped by a steel structure to give a total height of 1815 ft (553 m).

As well as floors and vertical structural systems, concrete has had significant application for shell roof construction. The first concrete shell roof was built in Jena, Germany, in 1924. Since then, many developments in form have occurred and spans increased. For example, the roof of the building of the Centre Nationale des Industries et Technologies in Paris, completed in 1958, consists of three corrugated fan-shaped half-arched shells of reinforced concrete. This roof structure is supported at three points that form in plan an equilateral triangle, the side length of which is 675.8 ft (206 m). The Kings County

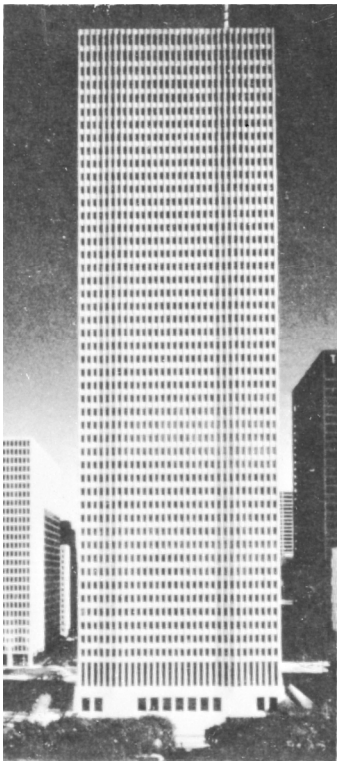


FIGURE 9 One Shell Plaza, Houston, Texas.

Stadium, situated in Seattle, Washington, was a dome of 661-ft (201.5-m) diameter consisting of a thin concrete shell stiffened by concrete ribs.

B. Bridges

Concrete bridges are generally either arch or beam in structural type. The beam is by far the most common structural element used in bridges. An arch carries the load of the bridge, mainly by axial compressive thrust in the member developed by the resistance against spread of the arch at the abutments. Some flexure and shear must also be resisted by the arch since the line of thrust for the loading generally does not follow the centroidal axis of the arch exactly. Beams are used in many structural configurations in bridges and carry load by flexure and shear. The beam can have a shaped cross section, such as an I or T section or a hollow box, and the depth can be varied along the length of the span to allow efficient use of materials. In multispan bridges either a series of simply supported beams can be used or the beam can be continuous. Cable-stayed bridges, in which cables attached to the beam transfer loading to towers at the bridge piers, are also used for long spans.

The first large-scale use of concrete, made with Portland cement, in bridges was probably the long series of concrete arches forming the Grand Maitre Aqueduct constructed in

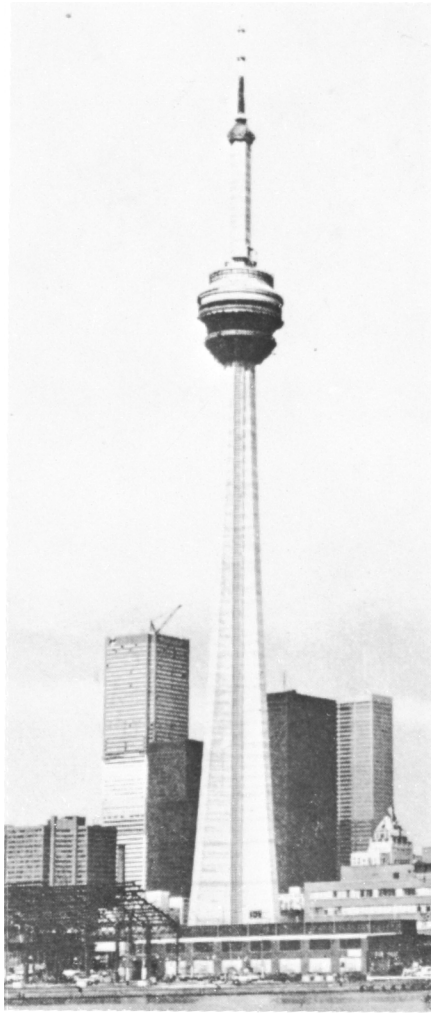


FIGURE 10 CN Tower, Toronto, Canada.

France around 1870. The first concrete bridge in Britain was an arch of 75-ft (23-m) span built at Earls Court, London, in 1867 by John Fowler. An early British reinforced concrete beam bridge is the Homersfield Bridge of 50-ft (15-m) span built in 1870 in Suffolk.

However, it was the French engineer Hennebique who had the greatest influence on early reinforced concrete bridge construction. Examples of Hennebique's work are his bridge built at Chatellerault in France in 1899, which is three arch spans of 131, 164, and 131 ft (40, 50, and 40 m), each with rise/span ratios of ∼0.1, and his Risorgimento Bridge built over the Tiber in Rome in 1911, which is a cellular arch of 328-ft (100-m) span with a rise/span ratio of 0.1. Although many of these early arch bridges are outstanding structures, the use of reinforced concrete in other bridge types was not particularly notable until the Swiss engineer Maillart (1872–1940) designed and built some outstanding bridge structures in the mountainous terrain of Switzerland. Significant use of reinforced concrete in bridge construction followed. The next era in bridge design and construction occurred after the Second World War when Eugene Freyssinet, with considerable courage and ingenuity, introduced the use of prestressed concrete. The first of Freyssinet's prestressed concrete bridges was completed in 1949 across the River Marne at Esbly, France, and was an arch of 243-ft (74-m) span. The ability to balance external loads with prestressing forces has given designers considerable scope for innovation in form and has led to slender, graceful structures. The St. James Park Footbridge in England built in 1958 (Fig. 11) is an excellent example of the elegance made possible by prestressed concrete. The 70-ft (21-m) main span of this bridge has a depth of only 14 in. (0.36 m) at midspan.

Bridge designs in many countries have led to the development of standard precast, prestressed concrete beams that are normally designed to act compositely with a cast-in-place reinforced concrete deck slab. For bridges with short spans the precast sections are normally voided rectangular, channel, or T shape. For bridges with medium spans of up to ∼130 ft (∼40 m), the precast sections are commonly of I shape. For long spans, segmental prestressed concrete hollow box girders are generally used. Typically in this form of construction the piers are cast in place. Then construction of the spans proceeds in each direction from the pier by the balanced cantileyer method, which involves placing precast or cast-in-place segments, each posttensioned to the previously completed construction. Alternatively, the span-by-span method can be used where the placing of the segments commences at one end of the structure and proceeds continuously to the other end and the segments are supported during construction by movable falsework. A box-girder bridge during segmental construction is shown in Fig. 12. Cable-stayed prestressed concrete box girders are also being used competitively for long spans. A novel form of bridge suitable for light loads and long spans is the stressed-ribbon bridge, consisting of

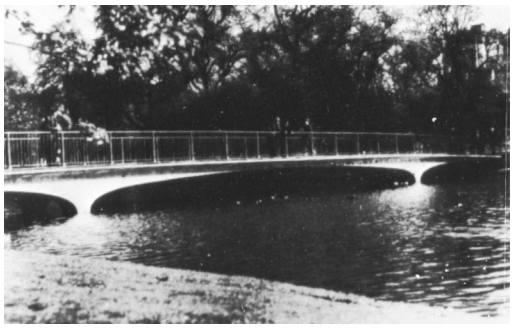


FIGURE 11 St. James Park Footbridge, London, England.

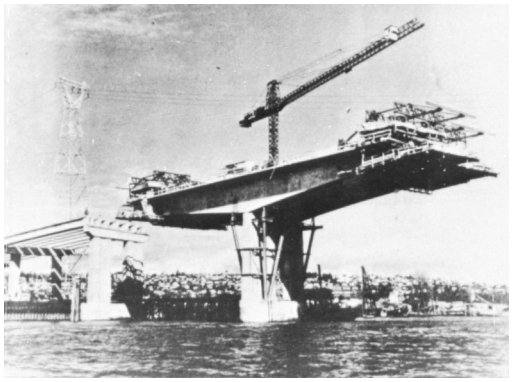


FIGURE 12 Knight Street Bridge, Vancouver–Richmond, Canada, during balanced cantilever construction.

a concrete deck supported directly on steel cables. Fatigue has not proved to be a problem in concrete bridges.

The world's longest span in concrete has been the Brotonne Bridge over the Seine in France, which is a cable-stayed structure with a prestressed concrete deck and a main span of 1050 ft (320 m). The longest span concrete arch bridge is the Gladesville Bridge in Australia, completed in 1964, which has a span of 1000 ft (305 m) and a rise of 134 ft (40.8 m). The longest main span for a cantilever bridge is 790 ft (241 m) and occurs in the Koror–Babelthuap Bridge in the Palau Islands in the Trust Territory of the Pacific Islands, which was completed in 1977. The longest main span of a stressed-ribbon bridge is 710 ft (216 m) and is the Holderbank–Wildegg Bridge in Switzerland built in 1964. It is technically feasible for these spans to be increased. For example, T. Y. Lin of the United States has proposed a scheme to build a bridge across the Bering Strait from Alaska to Siberia involving 250 precast prestressed cable-stayed spans, each with a length of 1200 ft (366 m).

C. Other Structures

Concrete has had significant use in such marine structures as docks, locks, barrages, barges, and towers. Perhaps the most interesting application in recent years has been the large concrete offshore towers that are in use in the North Sea for the drilling, production, and storage of gas or oil. These concrete towers are built in sheltered construction sites by land and then towed floating into position and sunk into place, leaving a working platform above water. The method of construction minimizes the amount of construction required in the open sea. These offshore towers are known as gravity structures because they are not pinned to the sea bottom by piles but simply sit on the sea bed. They have sufficient size and weight to give stability against overturning. The technology for concrete offshore structures has developed rapidly. The first structure of this type was the Ekofisk oil storage tank, with a capacity of 1 million barrels, which was installed in 230 ft (70 m) of water in the North Sea in 1973 (Fig. 13). In the period from 1973 to 1978, 16 offshore concrete towers were installed in the North Sea, in water depths ranging from 50 to 500 ft (15 to 152 m) with base diameters varying from 160 to 460 ft (50 to 140 m). These structures are indeed large in all respects. The Doris Ninian Tower required $\sim$185,000 yd^3 ($\sim$142,000 m^3) of concrete. Increasing use has also been made of reinforced and prestressed concrete for floating vessels. An example is a floating storage facility for liquid petroleum gas (LPG), 528 ft (161 m) long by 138 ft (42 m) wide, which was constructed in Tacoma, Washington, and towed almost 10,000 miles (16,000 km) to its working site in the Java Sea. On the whole, well-designed and -constructed concrete has proved to be a durable material in fixed and floating offshore structures.

Dams are impressive because of their size. The world's highest concrete dam is the Grande Dixence gravity dam of Switzerland, which was completed in 1962 and has a height of 935 ft (285 m). The Hoover Dam on the Colorado River was completed in 1936 and, with its height of 726 ft (221 m), was for 22 years the world's highest dam. The Hoover Dam is of the arch–gravity type, its structural action relying partly on arching between the valley walls and partly on self weight, and was a generation ahead of other large concrete dams.

Concrete is also widely used in airport and road pavements, foundation structures, cooling towers, tanks, tunnels, and other structures.

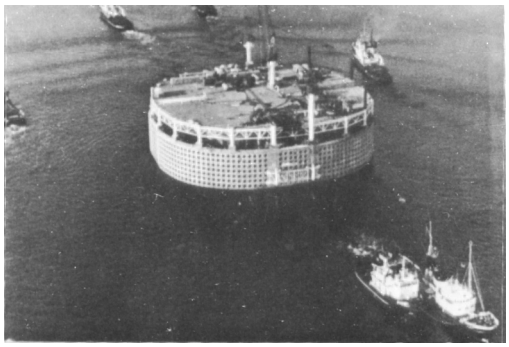


FIGURE 13 Efofisk oil storage tank under tow to the Ekofish oil field in the North Sea.

SEE ALSO THE FOLLOWING ARTICLES

BONDING AND STRUCTURE IN SOLIDS • FRACTURE AND FATIGUE • MASONRY • MECHANICS OF STRUCTURES

BIBLIOGRAPHY

Ah, P. W. (1967). *J. Am. Concr. Inst.* **64** (10).

American Concrete Institute (1999). "ACI 318-99 and 318R-99 Building Code Requirements for Structural Concrete and Commentary," CRC Press, Boca Raton, FL.

Carpinteri, A. (ed.) (1999). "Minimum Reinforcement in Concrete Members," Elsevier, Amsterdam.

Elfgren, L. (ed.) (1989). "Fracture Mechanics of Concrete Structures," Chapman and Hall, London.

Hamilton, S. B. (1956). "A Note on the History of Reinforced Concrete in Buildings," Natl. Build. Stud., Spec. Rep. No. 24, Dep. Sci. Ind. Res., H.M. Stationery Office, London.

Hanna, A. S. (1998). "Concrete Formwork Systems," Vol. 2, Civil and Environmental Engineering, Marcel Dekker, New York.

Hillerborg, A. (1975). "Strip Method of Design" (transl. from Swedish), Cement and Concrete Association, London.

Hopkins, H. J. (1970). "A Span of Bridges," David & Charles, Newton Abbot, England.

Institution of Civil Engineers Research Committee (1962). "Ultimate Load Design of Concrete Structures," *Proc. Inst. Civ. Eng.*, Vol. 21. ICE, London.

Johansen, K. W. (1962). "Yield Line Theory" (transl. from Danish), Cement and Concrete Association, London.

Lin, T. Y. (1963). "Design of Prestressed Concrete Structures," Wiley, New York.

Mehta, P. (1996). "Concrete: Microstructure, Properties, and Materials," 2nd ed., Primis Custom Publishing.

Neville, A. M. (1975). "Properties of Concrete," Pitman, London.

Oehlers, D. J., and Brodford, M. A. (1995). "Composite Steel and Concrete Structural Members," Elsevier, Amsterdam.

Park, R., and Gamble, W. L. (1980). "Reinforced Concrete Slabs," Wiley, New York.

Park, R., and Paulay, T. (1975). "Reinforced Concrete Structures," Wiley, New York.

Radomski, W. (2000). "Bridge Rehabilitation," World Scientific, Singapore.

Tanabe, T. (1999). "Comporative Performances of Seismic Design Codes for Concrete Structures," 2 vols., Elsevier, Amsterdam.



Constitutive Models for Engineering Materials

Kaspar J. Willam
University of Colorado at Boulder

I. Introductory Remarks
II. Elastic Models
III. Elastoplastic Models
IV. Analysis of Material Failure
V. Elastic Damage Models
VI. Conclusions

GLOSSARY

Elastic damage mechanics Degradation of elastic stiffness properties due to progressive damage.
Elasticity Linear and nonlinear stress–strain relations preserving reversibility in the small (hypoelasticity) and in the large (hyperelasticity).
Failure analysis Loss of stability, loss of uniqueness, and loss of ellipticity (localization) at the constitutive level.
Plasticity Yield constraint of stress (strain, energy) trajectory leading to irreversible deformations due to plastic flow.
Triaxial constitutive models Algebraic, differential, and integral equations relating three-dimensional stress and strain tensors.

CONSTITUTIVE MODELS describe the response behavior of natural and manufactured materials under different mechanical and environmental conditions. The focus of this overview is the mechanical performance of engineering materials expressed in terms of stress, strain, and internal state variables which describe the effect of the previous load history on the current properties. Thereby, the domain of traditional constitutive models encompasses the continuum concepts of elasticity, plasticity, viscosity, and their extension to include thermal and other environmental effects through continuum thermodynamics.

The constitutive model introduces or describes the physical properties of a given material. It connects the kinematic with the kinetic descriptions of motion, thus closing the formulation of the initial boundary value problem. In this context it is important to keep in mind the needs of modern computational analysis techniques, such as the finite element method, which demand triaxial constitutive models for realistic model-based simulations. In fact, it is the range of the underlying material models which delimits

the predictive value of large-scale simulations nowadays involving thousands of degrees of freedom. In short, our survey will concentrate on continuum-based material formulations currently used in engineering practice, in research and education, as well as in commercial finite element software packages for stress and deformation analyses. Primary applications cover life-cycle performance assessment of civil, mechanical, and aeronautical structures ranging from dams, bridges, and containment vessels to air- and spacecraft components as well as automotive crash simulations and micro-electro-mechanical systems.

I. INTRODUCTORY REMARKS

Traditionally, "material science" studies the behavior of materials at different scales in order to observe and quantify the chemo-physical processes at the underlying micromechanical, molecular, and atomistic levels. Multiscale material modeling upscales these processes onto the macroscopic level. The first step is to decompose the entire range of scales into several subranges. From the viewpoint of characterizing and designing engineering materials, we distinguish among the four scales illustrated in Fig. 1:

- *Meter level*: Practical problems in civil, mechanical, and aerospace structures such as the analysis and design of dams and containment vessels are solved at this level.
- *Millimeter level*: Most material properties are obtained from laboratory specimens at this level. In our terminology, this constitutes the macroscale—a level at which engineering materials may be treated as homogeneous continua after homogenizing the effect of the microstructural constituents into so-called *effective* properties.
- *Micrometer level*: Microstructural features such as microdefects, the grain size of polycrystals, and hydration products in cement-based materials are observed at this scale. In current terminology this constitutes the mesoscale—a level at which materials may be treated as heterogeneous composites (e.g., metal matrix and concrete composites) in which particle inclusions are bonded to the matrix by cohesive/frictional interface layers.
- *Nanometer level*: Molecular and atomistic processes take place at this level, which includes the molecular chaining of polymers and the behavior of single crystals. Many diffusion mechanisms for moisture and aggressive chemicals are considered to be active at this level, which includes, for example, the transport process of ions and chemical compounds. It should be noted that cause–effect relations in many cases reach beyond Newtonian mechanics, especially when subatomic processes are considered at the level of quantum mechanics.

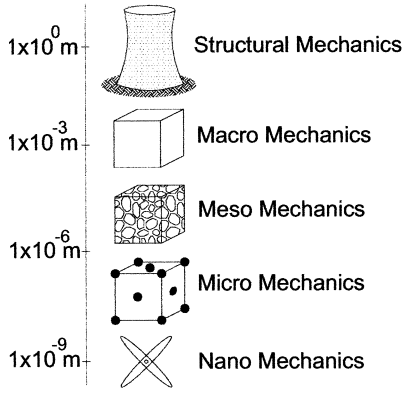


FIGURE 1 Multiscale material mechanics.

Historically, the current thinking of materials dates back to the "corpuscular" natural philosophy advanced by René Descartes (1596–1650), who postulated that the properties of matter emerge from multilevel microstructures comprised of molecules and voids. As early as 1722, René de Réaumur developed a fairly realistic picture of the anatomy and multilevel morphology of quench-hardened steel in his treatise on iron at different length scales. On a separate line of materials engineering, Joseph Aspdin patented in 1824 hydraulic Portland cement for manufacturing mortar and concrete. In the historical context, the great challenge is to bridge the gap between the atomistic thinking of the classical Greek school of natural philosophers around Democritus (460–370 BC), and the continuum world of differential calculus by Isaac Newton (1642–1727) and Gottfried Wilhelm Leibniz (1646–1716), which provides the framework of modern continuum mechanics founded by Augustin Cauchy (1789–1857).

Traditionally, engineering materials are considered to be macroscopically homogeneous (and often isotropic). While in most applications this approach may be adequate, progressive degradation processes can only be explained properly by considering microstructural features of the material. This requires characterization of each constituent and the interface bond conditions, in addition to the morphology of the specific meso- and microstructures, respectively. These types of studies are still very demanding in terms of manpower and computing power, in spite of the rapid development of computer simulations to investigate degradation processes at the micro-, meso-, and macrolevels in three-dimensional space and time.

In the long history of alchemy an ancient dream is to create precious materials by synthesizing and transmuting low-cost constituents through innovative chemo-physical processes. Translated into the world of high-tech materials

of today, this endeavor is at the core of "materials engineering," where optimization and material systems engineering meet well-defined performance objectives. Consequently, the design of new high-performance materials requires thorough understanding of the properties of each constituent and their interaction in chemo-physical processes. Thus, the role of constitutive relations is to quantify the performance of the resulting material compounds under mechanical and environmental load histories. For engineering purposes the length scale under consideration is the macroscale of traditional material testing laboratories, which provide the test data to calibrate the material parameters of the constitutive relation according to specifications of the American Society for Testing and Materials (ASTM) or other professional engineering societies.

In what follows, we will review established constitutive models for engineering materials, which may be broadly classified into three groups depending on their mathematical constructs:

1. Algebraic constitutive relations, such as linear and nonlinear elastic models
2. Differential constitutive relations, such as plastic and viscoplastic models
3. Integral constitutive relations, such as hereditary viscoelastic models

It is understood that the constitutive models are linear for performance analyses under service load conditions and highly nonlinear for limit state studies, when material failure is considered in the form of discontinuous fracture processes. For the sake of simplicity, this review will be restricted to small strains and local constitutive models in order to confine our attention to the interrelationship of stresses and strains, which are symmetric second-order tensors.

The review is organized as follows: Section II reviews primarily isotropic elastic models and examines three families of nonlinear elastic constitutive formulations using representation theorems of scalar and tensor functions. Section III summarizes the flow theory of elastoplasticity with a dicussion of volumetric–deviatoric coupling in the case of one-, two-, and three-invariant formulations. Section IV examines material failure at the constitutive level and introduces localization analysis with an application for loading in simple shear. Section V briefly reviews recent developments in continuum damage mechanics. Section VI concludes the state-of-the art report with seminal remarks on current research activities. Appendix 1 summarizes a few results of tensor algebra, while Appendix 2 contains some remarks on elastoplastic failure analysis subject to the kinematic constraint of plane strain.

II. ELASTIC MODELS

Linear elasticity is the main staple of material models in solids and structures. The statement "ut tensio sic vis," attributed to Robert Hooke (1635–1703), characterizes the behavior of a linear spring in which the deformations increase proportionally with the applied forces according to the anagram "ceiiinossstuv." The original format of Hooke's law included the geometric properties of the wire test specimens; therefore, the spring constant did exhibit a pronounced size effect. The definition of the modulus of elasticity E, where

$$\sigma = E\epsilon \tag{1}$$

is attributed to Thomas Young (1773–1829). He expressed the proportional material behavior through the notion of a normalized force density and a normalized deformation measure, though the original formulation also did not entirely eliminate the size effect.

The tensorial character of stress was established by Cauchy, who defined the triaxial state of stress by three traction vectors using the celebrated tetraeder argument of equilibrium. The state of stress is described in terms of Cartesian coordinates by the second-order tensor:

$$\boldsymbol{\sigma}(\boldsymbol{x}, t) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix} \tag{2}$$

The conjugate state of strain is a second-order tensor with Cartesian coordinates,

$$\boldsymbol{\epsilon}(\boldsymbol{x}, t) = \begin{bmatrix} \epsilon_{11} & \epsilon_{12} & \epsilon_{13} \\ \epsilon_{21} & \epsilon_{22} & \epsilon_{23} \\ \epsilon_{31} & \epsilon_{32} & \epsilon_{33} \end{bmatrix} \tag{3}$$

which is normally expressed in terms of the symmetric part of the displacement gradient, if we restrict our attention to infinitesimal deformations. In the case of nonpolar media we may confine our attention to stress measures, which are symmetric according to the axiom of L. Boltzmann,

$$\boldsymbol{\sigma} = \boldsymbol{\sigma}^t \quad \text{or} \quad \sigma_{ij} = \sigma_{ji} \tag{4}$$

and the conjugate strain measures,

$$\boldsymbol{\epsilon} = \boldsymbol{\epsilon}^t \quad \text{or} \quad \epsilon_{ij} = \epsilon_{ji} \tag{5}$$

where $i = 1, 2, 3$ and $j = 1, 2, 3$. As a result, the eigenvalues are real-valued and constitute the set of principal stresses and strains with zero shear components in the principal eigendirections of the second-order tensor. In contrast, nonsymmetric stress and strain measures may exhibit complex conjugate principal values and maximuum normal stress and strain components in directions with

nonzero shear components characteristic for micropolar Cosserat continua.

Restricting this exposition to symmetric stress and strain tensors, they may be cast into vector form using the Voigt notation of crystal physics:

$$\boldsymbol{\sigma}(\boldsymbol{x},t) = [\sigma_{11} \quad \sigma_{22} \quad \sigma_{33} \quad \tau_{12} \quad \tau_{23} \quad \tau_{31}]^t \tag{6}$$

and

$$\boldsymbol{\epsilon}(\boldsymbol{x},t) = [\epsilon_{11} \quad \epsilon_{22} \quad \epsilon_{33} \quad \gamma_{12} \quad \gamma_{23} \quad \gamma_{31}]^t \tag{7}$$

where $\tau_{ij} = \sigma_{ij}$, $\gamma_{ij} = 2\epsilon_{ij}$, $\forall i \neq j$. The vector form of stress and strain will allow us to formulate material models in matrix notation used predominantly in engineering, (some of the properties of second-order tensors and basic tensor operations are expanded in Appendix 1).

A. Linear Elastic Material Behavior

Generalization of the scalar format of Hooke's law is based on the notion that the triaxial state of stress is proportional to the triaxial state of strain through the linear transformation,

$$\boldsymbol{\sigma} = \boldsymbol{\mathcal{E}} : \boldsymbol{\epsilon} \quad \text{or} \quad \sigma_{ij} = \mathcal{E}_{ijkl}\epsilon_{kl} \tag{8}$$

Considering the symmetry of the stress and strain, the elasticity tensor involves in general 36 elastic moduli which may be further reduced to 21 elastic constants if we invoke major symmetry of the elasticity tensor; that is,

$$\boldsymbol{\mathcal{E}} = \boldsymbol{\mathcal{E}}^t \quad \text{or} \quad \mathcal{E}_{ijkl} = \mathcal{E}_{klij}$$

$$\text{with} \quad \mathcal{E}_{ijkl} = \mathcal{E}_{ijlk} \quad \text{and} \quad \mathcal{E}_{ijkl} = \mathcal{E}_{jikl} \tag{9}$$

The task of identifying 21 elastic moduli is simplified if we consider specific classes of symmetry, thus orthotropic elasticity involves nine, and transversely anisotropic elasticity five elastic moduli.

1. Isotropic Linear Elasticity

In the case of isotropy, the fourth-order elasticity tensor has the most general representation,

$$\boldsymbol{\mathcal{E}} = a_o\mathbf{1}\otimes\mathbf{1} + a_1\mathbf{1}\bar{\otimes}\mathbf{1} + a_2\mathbf{1}\underline{\otimes}\mathbf{1}$$

$$\text{or} \quad \mathcal{E}_{ijkl} = a_o\delta_{ij}\delta_{kl} + a_1\delta_{ik}\delta_{jl} + a_2\delta_{il}\delta_{jk} \tag{10}$$

where $\mathbf{1} = [\delta_{ij}]$ stands for the second-order unit tensor. The three-parameter expression may be recast in terms of symmetric and skew symmetric fourth-order tensor components as:

$$\boldsymbol{\mathcal{E}} = a_o\mathbf{1}\otimes\mathbf{1} + b_1\boldsymbol{\mathcal{I}} + b_2\boldsymbol{\mathcal{I}}^{skew} \tag{11}$$

where the symmetric fourth-order unit tensor reads:

$$\boldsymbol{\mathcal{I}} = \frac{1}{2}[\mathbf{1}\bar{\otimes}\mathbf{1} + \mathbf{1}\underline{\otimes}\mathbf{1}] \quad \text{or} \quad \mathcal{I}_{ijkl} = \frac{1}{2}[\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}] \tag{12}$$

and the skewed symmetric one,

$$\boldsymbol{\mathcal{I}}^{skew} = \frac{1}{2}[\mathbf{1}\bar{\otimes}\mathbf{1} - \mathbf{1}\underline{\otimes}\mathbf{1}] \text{ or } \mathcal{I}^{skew}_{ijkl} = \frac{1}{2}[\delta_{ik}\delta_{jl} - \delta_{il}\delta_{jk}] \tag{13}$$

Because of the symmetry of stress and strain the skewed symmetric contribution is inactive, $b_2 = 0$, and the isotropic linear elasticity of the material behavior is fully described by two independent elastic constants. In short, the fourth-order material stiffness tensor reduces to

$$\boldsymbol{\mathcal{E}} = \Lambda\mathbf{1}\otimes\mathbf{1} + 2G\boldsymbol{\mathcal{I}}$$

$$\text{or} \quad \mathcal{E}_{ijkl} = \Lambda\delta_{ij}\delta_{kl} + G[\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}] \tag{14}$$

where the two elastic constants Λ and G are named after Gabriel Lamé (1795–1870).

$$\Lambda = \frac{E\nu}{[1+\nu][1-2\nu]} \tag{15}$$

denotes the cross modulus, and

$$G = \frac{E}{2[1+\nu]} \tag{16}$$

designates the shear modulus which have a one-to-one relationship with the modulus of elasticity and Poisson's ratio, E, ν. In the absence of initial stresses and initial strains due to environmental effects, the linear elastic relation reduces to

$$\boldsymbol{\sigma} = \Lambda[tr\boldsymbol{\epsilon}]\mathbf{1} + 2G\boldsymbol{\epsilon} \quad \text{or} \quad \sigma_{ij} = \Lambda\epsilon_{kk}\delta_{ij} + 2G\epsilon_{ij} \tag{17}$$

Here, the trace operation is the sum of the diagonal entries of the second-order tensor corresponding to double contraction with the identity tensor $tr\boldsymbol{\epsilon} = \epsilon_{kk} = \mathbf{1} : \boldsymbol{\epsilon}$.

2. Matrix Form of Elastic Stiffness ($\boldsymbol{\sigma} = \boldsymbol{E}\boldsymbol{\epsilon}$)

Isotropic linear elastic behavior may be cast in matrix format using the Voigt notation of symmetric stress and strain tensors and the engineering definition of shear strain, $\gamma_{ij} = 2\epsilon_{ij}$. The elastic stiffness matrix may be written for isotropic behavior as:

$$\boldsymbol{E} = \left[\begin{array}{ccc|ccc} \Lambda+2G & \Lambda & \Lambda & & & \\ \Lambda & \Lambda+2G & \Lambda & & 0 & \\ \Lambda & \Lambda & \Lambda+2G & & & \\ \hline & & & G & & \\ & 0 & & & G & \\ & & & & & G \end{array}\right] \tag{18}$$

3. Matrix Form of Elastic Compliance ($\epsilon = C\sigma$)

In the isotropic case, the normal stress σ_{11} gives rise to three normal strain contributions, the direct strain $\epsilon_{11} = \frac{1}{E}\sigma_{11}$ and the normal strains $\epsilon_{22} = -\frac{\nu}{E}\sigma_{11}$ and $\epsilon_{33} = \frac{\nu}{E}\sigma_{11}$ because of the cross effect attributed to Siméon Denis Poisson (1781–1840). Using the principle of superposition, the additional strain contributions due to σ_{22} and σ_{33} enter the compliance relation for isotropic elasticity in matrix format:

$$\begin{bmatrix} \epsilon_{11} \\ \epsilon_{22} \\ \epsilon_{33} \end{bmatrix} = \frac{1}{E} \begin{bmatrix} 1 & -\nu & -\nu \\ -\nu & 1 & -\nu \\ -\nu & -\nu & 1 \end{bmatrix} \begin{bmatrix} \sigma_{11} \\ \sigma_{22} \\ \sigma_{33} \end{bmatrix} \tag{19}$$

In the isotropic case, the shear response is entirely decoupled from the direct response of the normal components. Thus, the compliance matrix expands into the partitioned form:

$$\boldsymbol{C} = \frac{1}{E} \left[\begin{array}{ccc|ccc} 1 & -\nu & -\nu & & & \\ -\nu & 1 & -\nu & & \mathbf{0} & \\ -\nu & -\nu & 1 & & & \\ \hline & & & 2[1+\nu] & & \\ & \mathbf{0} & & & 2[1+\nu] & \\ & & & & & 2[1+\nu] \end{array}\right] \tag{20}$$

where isotropy entirely decouples the shear response from the normal stress–strain response. This cross effect of Poisson is illustrated in Fig. 2, which shows the interaction of lateral and axial deformations under axial compression. It is intriguing that in his original work a value of $\nu = 0.25$ was proposed by Poisson based on molecular considerations. The elastic compliance relation reads in direct and indicial notations:

$$\mathcal{C} = -\frac{\nu}{E}\mathbf{1} \otimes \mathbf{1} + \frac{1}{2G}\mathcal{I}$$

$$\text{or} \quad C_{ijkl} = -\frac{\nu}{E}\delta_{ij}\delta_{kl} + \frac{1+\nu}{2E}[\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}] \tag{21}$$

4. Canonical Format of Isotropic Elasticity

Decomposing the stress and strain tensors into spherical and deviatoric components,

$$\boldsymbol{s} = \boldsymbol{\sigma} - \sigma_{vol}\mathbf{1} \quad \text{where} \quad \sigma_{vol} = \frac{1}{3}[tr\boldsymbol{\sigma}] \tag{22}$$

$$\boldsymbol{e} = \boldsymbol{\epsilon} - \epsilon_{vol}\mathbf{1} \quad \text{where} \quad \epsilon_{vol} = \frac{1}{3}[tr\boldsymbol{\epsilon}] \tag{23}$$

leads to the stress deviator

$$\boldsymbol{s}(\boldsymbol{x}, t) = \begin{bmatrix} \frac{2\sigma_{11} - \sigma_{22} - \sigma_{33}}{3} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \frac{2\sigma_{22} - \sigma_{33} - \sigma_{11}}{3} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \frac{2\sigma_{33} - \sigma_{11} - \sigma_{22}}{3} \end{bmatrix} \tag{24}$$

and the strain deviator

$$\boldsymbol{e}(\boldsymbol{x}, t) = \begin{bmatrix} \frac{2\epsilon_{11} - \epsilon_{22} - \epsilon_{33}}{3} & \epsilon_{12} & \epsilon_{13} \\ \epsilon_{21} & \frac{2\epsilon_{22} - \epsilon_{33} - \epsilon_{11}}{3} & \epsilon_{23} \\ \epsilon_{31} & \epsilon_{32} & \frac{2\epsilon_{33} - \epsilon_{11} - \epsilon_{22}}{3} \end{bmatrix} \tag{25}$$

which have the property $trs = 0$ and $tre = 0$. The decomposition decouples the volumetric from the distortional response, because of the underlying orthogonality of the spherical and deviatoric partitions, $\boldsymbol{s} : [\sigma_{vol}\mathbf{1}] = 0$ and $\boldsymbol{e} : [\epsilon_{vol}\mathbf{1}] = 0$. The decoupled response reduces the elasticity tensor to the scalar form,

$$\sigma_{vol} = 3K\epsilon_{vol} \quad \text{and} \quad \boldsymbol{s} = 2G\boldsymbol{e} \tag{26}$$

in which the bulk and the shear moduli,

$$K = \frac{E}{3[1-2\nu]} = \Lambda + \frac{2}{3}G$$

$$\text{and} \quad G = \frac{E}{2[1+\nu]} = \frac{3}{2}[K - \Lambda] \tag{27}$$

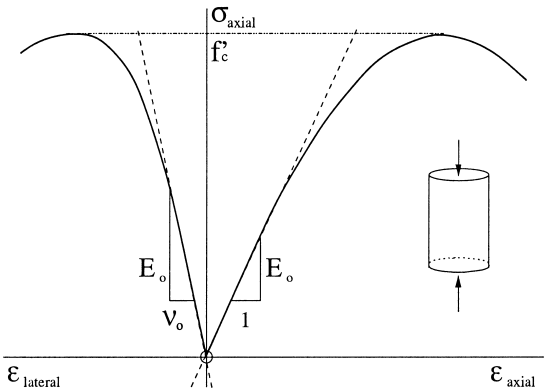


FIGURE 2 Poisson effect in axial compression.

define the volumetric and the deviatoric material stiffness. Consequently, the internal strain-energy density expands into the canonical form of two decoupled contributions,

$$2W = \boldsymbol{\sigma} : \boldsymbol{\epsilon} = [\sigma_{vol}\mathbf{1}] : [\epsilon_{vol}\mathbf{1}] + \boldsymbol{s} : \boldsymbol{e} = 9K\epsilon_{vol}^2 + 2G\boldsymbol{e} : \boldsymbol{e} \tag{28}$$

such that the positive strain energy argument delimits the range of possible values of Poisson's ratio to $-1 \leq \nu \leq 0.5$.

5. Isotropic Elasticity under Initial Volumetric Strain

In the case of isotropic material behavior, with no directional properties, the size of a representative volume element may change due to thermal effects or shrinkage and swelling, but it will not distort. Consequently, the expansion is purely volumetric (i.e., identical in all directions). Using direct and indicial notation, the additive decomposition of strain into elastic and initial volumetric components, $\boldsymbol{\epsilon} = \boldsymbol{\epsilon}_e + \boldsymbol{\epsilon}_o$, leads to the following extension of the elastic compliance relation:

$$\begin{aligned} \boldsymbol{\epsilon} &= -\frac{\nu}{E}[tr\boldsymbol{\sigma}]\mathbf{1} + \frac{1}{2G}\boldsymbol{\sigma} + \epsilon_o\mathbf{1} \\ \text{or} \quad \epsilon_{ij} &= -\frac{\nu}{E}\sigma_{kk}\delta_{ij} + \frac{1}{2G}\sigma_{ij} + \epsilon_o\delta_{ij} \end{aligned} \tag{29}$$

where $\boldsymbol{\epsilon}_o = \epsilon_o\mathbf{1}$ denotes the initial volumetric strain (e.g., due to thermal expansion). The inverse relation reads:

$$\begin{aligned} \boldsymbol{\sigma} &= \Lambda[tr\boldsymbol{\epsilon}]\mathbf{1} + 2G\boldsymbol{\epsilon} - 3\epsilon_o K\mathbf{1} \\ \text{or} \quad \sigma_{ij} &= \Lambda\epsilon_{kk}\delta_{ij} + 2G\epsilon_{ij} - 3\epsilon_o K\delta_{ij} \end{aligned} \tag{30}$$

Rewriting this equation in matrix notation, we have:

$$\begin{bmatrix} \sigma_{11} \\ \sigma_{22} \\ \sigma_{33} \end{bmatrix} = \begin{bmatrix} K + \frac{4}{3}G & K - \frac{2}{3}G & K - \frac{2}{3}G \\ K - \frac{2}{3}G & K + \frac{4}{3}G & K - \frac{2}{3}G \\ K - \frac{2}{3}G & K - \frac{2}{3}G & K + \frac{4}{3}G \end{bmatrix} \begin{bmatrix} \epsilon_{11} \\ \epsilon_{22} \\ \epsilon_{33} \end{bmatrix} - 3K\epsilon_o \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \tag{31}$$

Considering the special case of plane stress, $\sigma_{33} = 0$, the stress–strain relations reduce in the presence of initial volumetric strains to:

$$\begin{bmatrix} \sigma_{11} \\ \sigma_{22} \end{bmatrix} = \frac{E}{1-\nu^2}\begin{bmatrix} 1 & \nu \\ \nu & 1 \end{bmatrix}\begin{bmatrix} \epsilon_{11} \\ \epsilon_{22} \end{bmatrix} - \frac{E}{1-\nu}\epsilon_o\begin{bmatrix} 1 \\ 1 \end{bmatrix} \tag{32}$$

where the shear components are not affected by the temperature change in the case of isotropy.

6. Free Thermal Expansion

Under stress-free conditions, the thermal expansion $\boldsymbol{\epsilon}_o = \alpha[T - T_o]\mathbf{1}$ leads to $\boldsymbol{\epsilon} = \boldsymbol{\epsilon}_o$; that is,

$$\epsilon_{11} = \alpha[T - T_o] \tag{33}$$

$$\epsilon_{22} = \alpha[T - T_o] \tag{34}$$

$$\epsilon_{33} = \alpha[T - T_o] \tag{35}$$

Thus, the change of temperature results in free thermal expansion, while the mechanical stress remains zero under zero confinement, $\boldsymbol{\sigma} = \boldsymbol{\mathcal{E}} : \boldsymbol{\epsilon}_e = \mathbf{0}$.

7. Thermal Stress under Full Confinement

In contrast to the unconfined situation above, the thermal expansion is equal and opposite to the elastic strain $\boldsymbol{\epsilon}_e = -\boldsymbol{\epsilon}_o$ under full confinement, when $\boldsymbol{\epsilon} = 0$. In the case of plane stress, the temperature change $\Delta T = T - T_o$ leads to the thermal stresses:

$$\sigma_{11} = -\frac{E}{1-\nu}\alpha[T - T_o] \tag{36}$$

$$\sigma_{22} = -\frac{E}{1-\nu}\alpha[T - T_o] \tag{37}$$

B. Nonlinear Elasticity

In linear elasticity, the resulting constitutive relations do not depend on the point of departure, except for the symmetry argument, which is derived most naturally from the energy argument. In contrast, nonlinear elastic material models strongly depend on the setting of the constitutive formulation. Considering for a moment the uniaxial stress–strain relation shown in Fig. 3, it appears that the nonlinear material law may be formulated in terms of any one of the following three possibilities:

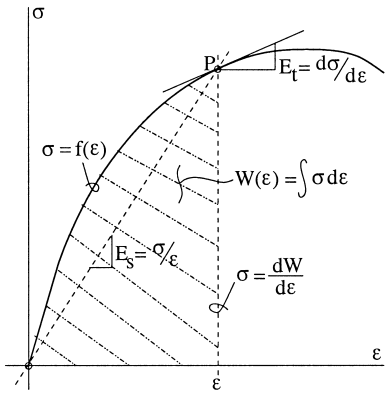


FIGURE 3 Conceptual material models of nonlinear elasticity.

1. *Algebraic format* ($\boldsymbol{\sigma} = \boldsymbol{f}(\boldsymbol{\epsilon})$): The algebraic format extends Hooke's law $\boldsymbol{\sigma} = \boldsymbol{\mathcal{E}} : \boldsymbol{\epsilon}$ into the nonlinear regime. The triaxial generalization of the scalar-valued algebraic function leads in the simplest case to the concept of a secant stiffness, $\boldsymbol{\sigma} = \boldsymbol{E}_s : \boldsymbol{\epsilon}$. In the triaxial situation, this leads to the question of how to represent a symmetric tensor-valued function of the symmetric strain tensor, $\boldsymbol{\sigma} = \boldsymbol{f}(\boldsymbol{\epsilon})$. This type of constitutive model is limited to rate- and history-independent material behavior, such as linear and nonlinear elasticity, which do not exhibit hysteretic effects.

2. *Integral format* ($\boldsymbol{\sigma} = \frac{\partial W}{\partial \boldsymbol{\epsilon}}$): The integral format is the repository of functional representations which include linear and nonlinear viscoelasticity in which creep and relaxation are considered in the form of fading memory effects, $\boldsymbol{\sigma}(t) = \boldsymbol{\mathcal{F}}_{t_o \le \tau \le t} \frac{\partial W(\boldsymbol{\epsilon}(\tau), \boldsymbol{q})}{\partial \boldsymbol{\epsilon}}$. As long as the rate sensitivity does not include any history effects, $\boldsymbol{q} = \boldsymbol{0}$, the functional representation does not exhibit hysteretic effects. In this case and in the absence of rate dependence, we recover the format of instantaneous hyperelasticity, in which the stress is derived from the gradient of an elastic potential, $\boldsymbol{\sigma} = \frac{\partial W}{\partial \boldsymbol{\epsilon}}$.

3. *Differential format* ($d\boldsymbol{\sigma} = \boldsymbol{E}_t : d\boldsymbol{\epsilon}$): In its elementary format the differential form reduces to the tangential stress–strain relationship. In a broader sense it provides also a repository for internal variables, which memorize inelastic changes of the material properties in plasticity and viscoplasticity. In this case, hysteretic effects are included to account for material dissipation and energy release in the case of fracture. Restricting our attention to nonlinear hypolasticity, the differential format leads to the question of how to represent the stress rate in terms of a tensor-valued function of two symmetric tensors, $\dot{\boldsymbol{\sigma}} = \boldsymbol{g}(\boldsymbol{\sigma}, \dot{\boldsymbol{\epsilon}})$.

In what follows we will see that the three versions of nonlinear elasticity lead to constitutive formulations which exhibit fundamental differences when we consider triaxial conditions.

1. Algebraic Format

The so-called "total" stress–strain relation expresses stress in terms of a nonlinear function of strain. As an alternative, the secant relation, $\boldsymbol{\sigma} = \boldsymbol{\mathcal{E}}_s : \boldsymbol{\epsilon}$ provides a pseudolinear format, which simply redefines the nonlinear function $\boldsymbol{f}(\boldsymbol{\epsilon})$ in terms of the nonlinear secant modulus $\boldsymbol{\mathcal{E}}_s$.

The triaxial setting, however, leads to two algebraic formats of nonlinear elasticity which differ in a fundamental manner:

a. Cauchy elasticity ($\boldsymbol{\sigma} = \boldsymbol{f}(\boldsymbol{\epsilon})$). In this case, the triaxial state of stress is a nonlinear tensor function of the strain tensor—in indicial notation, $\sigma_{ij} = f_{ij}(\epsilon_{kl})$. Using the representation theorems of second-order symmetric tensor functions the possibilities are restricted to a small set of possible choices when isotropy is invoked. In this case, the most general format of Cauchy elasticity may have one of the two representations,

$$\boldsymbol{\sigma} = \Phi_1 \boldsymbol{1} + \Phi_2 \boldsymbol{\epsilon} + \Phi_3 \boldsymbol{\epsilon}^2 \quad \text{or} \quad \boldsymbol{\sigma} = \Psi_1 \boldsymbol{1} + \Psi_2 \boldsymbol{\epsilon} + \Psi_3 \boldsymbol{\epsilon}^{-1} \tag{38}$$

because of the Cayleigh–Hamilton theorem. Thus, the three response functions Φ_i and Ψ_i are scalar functions of the three invariants of either stress or strain. It is important to keep in mind that Cauchy elasticity is based on a second-order symmetric tensor function of a second-order symmetric tensor.

b. Secant or pseudoelasticity ($\boldsymbol{\sigma} = \boldsymbol{\mathcal{E}}_s : \boldsymbol{\epsilon}$). In the pseudoelastic format of elasticity, the nonlinearity is incorporated into the fourth-order secant stiffness tensor. This format has been used early on in different engineering disciplines to develop nonlinear extensions of simple classes of linear elasticity. The so-called variable stiffness models retain the format of linear elasticity and simply replace the two elastic constants of linear elasticity by nonlinear functions. We will see later on that the nonlinear K–G models are theoretically sound if the nonlinear response decouples the volumetric from the deviatoric response; that is, $K = K(\epsilon_{vol})$ and $G = G(tr\boldsymbol{e}^2)$ where $tr\boldsymbol{e}^2 = \boldsymbol{e} : \boldsymbol{e}$.

It is not very surprising that elastic damage models did start from this secant format of nonlinear elasticity using arguments of effective stiffness properties, which are in some sense equivalent to distributed micromechanical defects. In fact, the original proposal of scalar damage was nothing but a nonlinear pseudoelastic model in which the secant stiffness is in matrix notation:

$$\boldsymbol{\sigma} = \boldsymbol{E}_s \boldsymbol{\epsilon} \quad \text{where} \quad \boldsymbol{E}_s = [1-d]\boldsymbol{E}_o \quad \text{and} \quad d = 1 - \frac{E_s}{E_o} \tag{39}$$

In its basic format, the secant matrix of elastic damage retains the structure of the initial elastic stiffness except for the factor $[1-d]$:

$$\boldsymbol{E}_s = \frac{[1-d]E_o}{[1+\nu_o][1-2\nu_o]} \times \left[\begin{array}{ccc|ccc} 1-\nu_o & \nu_o & \nu_o & & & \\ \nu_o & 1-\nu_o & \nu_o & & 0 & \\ \nu_o & \nu_o & 1-\nu_o & & & \\ \hline & & & \frac{1-2\nu_o}{2} & & \\ & 0 & & & \frac{1-2\nu_o}{2} & \\ & & & & & \frac{1-2\nu_o}{2} \end{array}\right] \tag{40}$$

which measures the remaining integrity of the material when the damage variable increases from zero to one, $0 \leq d \rightarrow 1$. Restricting damage to the format of isotropic linear elasticity, it is natural to decompose degradation into volumetric and deviatoric damage components such that:

$$K_s = [1 - d_{vol}]K_o \quad \text{and} \quad G_s = [1 - d_{dev}]G_o \tag{41}$$

The corresponding secant relation of elastic damage reads in matrix notation

$$\boldsymbol{E}_s = \left[\begin{array}{ccc|ccc} K_s + \frac{4}{3}G_s & K_s - \frac{2}{3}G_s & K_s - \frac{2}{3}G_s & & & \\ K_s - \frac{2}{3}G_s & K_s + \frac{4}{3}G_s & K_s - \frac{2}{3}G_s & & 0 & \\ K_s - \frac{2}{3}G_s & K_s - \frac{2}{3}G_s & K_s + \frac{4}{3}G_s & & & \\ \hline & & & G_s & & \\ & 0 & & & G_s & \\ & & & & & G_s \end{array}\right] \tag{42}$$

From this expression we observe that the secant format of isotropic elastic damage is very simple because of the underlying decoupling of volumetric from deviatoric degradation. As long as we are only interested in a given state of damage, this isotropic pseudoelastic formulation suffices to describe the response behavior using effective material properties based on the principle of stress or strain equivalence. However, the constitutive formulation becomes far more intricate when a thermodynamically consistent damage formulation is needed for progressive damage simulations. In this case, loading and unloading have to be defined for the general case of triaxial conditions, and evolution laws are required to describe the two independent processes of volumetric and deviatoric damage which will be described in Section 5 in more detail.

2. Integral Format

The integral description of elasticity starts from the postulate of a strain energy function from which the stress is derived by differentiation. This hyperelastic format of elasticity dates back to the original work of George Green (1793–1841).

a. Green elasticity ($\boldsymbol{\sigma} = \frac{\partial W}{\partial \boldsymbol{\epsilon}}$). Given a strain energy density potential, $W = W(\boldsymbol{\epsilon})$, the hyperelastic stress–strain relation is simply the gradient of that potential function with respect to strain,

$$\boldsymbol{\sigma} = \frac{\partial W}{\partial \boldsymbol{\epsilon}} \quad \text{or} \quad \sigma_{ij} = \frac{\partial W}{\partial \epsilon_{ij}} \tag{43}$$

The compliance relation derives from the dual complementary strain energy potential. The important aspect of this integral formulation is the path-independence of the line integral which defines the internal strain energy,

$$W(\boldsymbol{\epsilon}) = \int_{\epsilon} \boldsymbol{\sigma} : d\boldsymbol{\epsilon} = \int_{\epsilon} \frac{\partial W}{\partial \boldsymbol{\epsilon}} : d\boldsymbol{\epsilon} = \int_{\epsilon} dW \tag{44}$$

in terms of the total differential dW. The traditional notions of elasticity, such as reversibility and lack of energy dissipation under closed cycles of strain, are a direct consequence of path independence; that is,

$$W(\boldsymbol{\epsilon}) = \oint_{\epsilon} dW = 0 \tag{45}$$

In other terms, the hyperelastic material description is nondissipative and preserves energy under arbitrary strain histories.

The corresponding hyperelastic stiffness tensor is a measure of the curvature of the strain energy function involving the second derivatives of $W = W(\boldsymbol{\epsilon})$,

$$\dot{\boldsymbol{\sigma}} = \boldsymbol{\mathcal{E}}_t : \dot{\boldsymbol{\epsilon}} \quad \text{where} \quad \boldsymbol{\mathcal{E}}_t = \frac{\partial^2 W}{\partial \boldsymbol{\epsilon} \otimes \partial \boldsymbol{\epsilon}} \tag{46}$$

Consequently, the elasticity tensor is symmetric, $\boldsymbol{\mathcal{E}}_t = \boldsymbol{\mathcal{E}}_t^t$, if $W(\boldsymbol{\epsilon})$ is sufficiently smooth. This reduces the 36 elastic constants to 21 in the general case of anisotropic elasticity, and to two in case of isotropy. The positive definiteness of the hyperelastic tangent operator, $\det \boldsymbol{\mathcal{E}}_t > 0$, is directly connected to the convexity of the strain energy functional and the uniqueness argument of boundary value problems in elasticity.

b. Isotropic hyperelastic models. In the case of isotropy, frame objectivity requires that the strain energy density function must be independent from the coordinate system of an observer. Thus, the potential function must be a function of strain invariants and combinations thereof. In the past, two different families of nonlinear isotropic hyperelastic material models have been proposed; one leads to the format of nonlinear K–G models, while the other results in a pseudo-orthotropic format based on principal coordinates of stress and strain.

1. *Hyperelastic K–G models*: Starting from the invariant representation of the strain energy density function, $W(\boldsymbol{\epsilon}) = W(I_1, I_2, I_3)$, the stress–strain relation follows from the chain rule of differentiation as:

$$\boldsymbol{\sigma} = \frac{\partial W}{\partial \boldsymbol{\epsilon}} = \frac{\partial W}{\partial I_1}\frac{\partial I_1}{\partial \boldsymbol{\epsilon}} + \frac{\partial W}{\partial I_2}\frac{\partial I_2}{\partial \boldsymbol{\epsilon}} + \frac{\partial W}{\partial I_3}\frac{\partial I_3}{\partial \boldsymbol{\epsilon}} \tag{47}$$

Using the moment invariants for the sake of convenience, where $I_i = \frac{1}{i} tr \boldsymbol{\epsilon}^i$, the derivatives simplify to $\frac{\partial I_i}{\partial \boldsymbol{\epsilon}} = \boldsymbol{\epsilon}^{i-1}$. Consequently, the nonlinear isotropic stress–strain relation results in the general hyperelastic format for nonlinear isotropic behavior,

$$\boldsymbol{\sigma} = W_1 \mathbf{1} + W_2 \boldsymbol{\epsilon} + W_3 \boldsymbol{\epsilon}^2 \tag{48}$$

The three scalar functions $W_i = \frac{\partial W}{\partial I_i} = W_i(I_j)$ describe the degree of nonlinearity in terms of three irreducible invariants, while the tensorial properties are characterized by the three irreducible basis tensors $\mathbf{1}$, $\boldsymbol{\epsilon}$, $\boldsymbol{\epsilon}^2$. Note the interrelationship between the scalar functions,

$$\frac{\partial W_i}{\partial I_j} = \frac{\partial W_j}{\partial I_i} \quad \text{since} \quad \frac{\partial^2 W}{\partial I_i \partial I_j} = \frac{\partial^2 W}{\partial I_j \partial I_i} \tag{49}$$

when they are derived from a sufficiently differentiable scalar potential, $W = W(I_1, I_2, I_3)$. In contrast, the scalar functions Φ_i and Ψ_i of the pseudoelastic formulation do not satisfy this form of reciprocity if the underlying integrability conditions are not enforced. On the issue of volumetric–deviatoric coupling we observe that the trace operation leads to:

$$tr\boldsymbol{\sigma} = 3W_1 + W_2 tr\boldsymbol{\epsilon} + W_3 tr\boldsymbol{\epsilon}^2 \tag{50}$$

Considering a simple shear deformation, with $\epsilon_{12} = 0.5\gamma$, W_3 and thus the dependence of $W = W(I_1, I_2, I_3)$ on the third invariant are responsible for volumetric–deviatoric interaction. On another note, the quadratic expansion of the strain energy density function leads exactly to the two Lamé constants of linear isotropic elasticity since $W_3 = 0$. Thus, it was Green hyperelasticity which firmly established the bimodular theory of isotropic elasticity in contrast to the unimodular theory which was widely accepted by the French school of engineering scientists in the last century. In fact, this bimodular format also holds for nonlinear hyperelasticity if the strain energy density function can be decomposed into two independent functions, one defining the volumetric and the other the deviatoric behavior:

$$W(\boldsymbol{\epsilon}) = W_{vol}(tr\boldsymbol{\epsilon}) + W_{dev}(tr\mathbf{e}^2) \tag{51}$$

This infers, however, that the influence of the third invariant I_3 remains negligible, since it is this term which is responsible for coupling the volumetric and deviatoric response behavior. The decomposition of the strain energy function leads to the appealing concept of nonlinear K–G models because of their inherent simplicity which retains the two-modular form of linear elasticity. Figures 4 and 5 illustrate the secant stiffness relations, which may be expressed best in the form of the octahedral components of stress σ_o, τ_o and strain ϵ_o, γ_o:

$$\begin{bmatrix} \sigma_o \\ \tau_o \end{bmatrix} = \begin{bmatrix} 3K_s & 0 \\ 0 & 2G_s \end{bmatrix} \begin{bmatrix} \epsilon_o \\ \gamma_o \end{bmatrix} \tag{52}$$

where $K_s = K(tr\boldsymbol{\epsilon})$ and $G_s = G(tr\boldsymbol{e}^2)$. The so-called K–G models shown in Figs. 4 and 5 play a prominent role for modeling nonlinear material behavior. Combining the volumetric and deviatoric scalar relations leads to the secant stiffness relation:

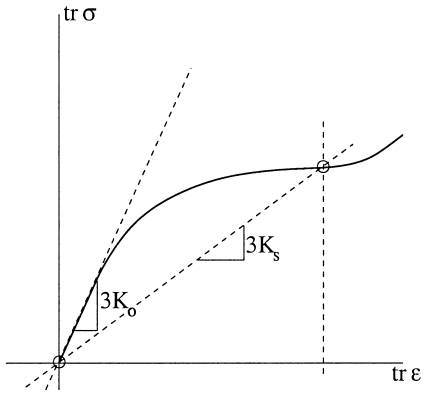


FIGURE 4 Nonlinear volumetric response behavior.

$$\boldsymbol{\sigma} = \boldsymbol{\mathcal{E}}_s : \boldsymbol{\epsilon} \quad \text{where} \quad \boldsymbol{\mathcal{E}}_s = \left[K_s - \frac{2}{3}G_s\right]\mathbf{1} \otimes \mathbf{1} + 2G_s\boldsymbol{\mathcal{I}} \tag{53}$$

where the elastic material constants have been simply replaced by their secant values. The corresponding tangential stiffness relation, however, introduces an additional fourth-order tensor term because of the chain rule of differentiation when taking derivatives of $G_s = G_s(\sqrt{\boldsymbol{e} : \boldsymbol{e}})$,

$$\dot{\boldsymbol{\sigma}} = \boldsymbol{\mathcal{E}}_t : \dot{\boldsymbol{\epsilon}} \quad \text{where} \quad \boldsymbol{\mathcal{E}}_t = \left[K_t - \frac{2}{3}G_s\right]\mathbf{1} \otimes \mathbf{1} + 2G_s\boldsymbol{\mathcal{I}} + \frac{4[G_t - G_s]}{tr\boldsymbol{e}^2}\boldsymbol{e} \otimes \mathbf{e} \tag{54}$$

The dyadic tensor product $\boldsymbol{e} \otimes \boldsymbol{e}$ originates from the nonlinearity and reflects the state of the deviator strain on the incremental material law. This term is the source of strain-induced anisotropy, which characterizes the tangential stiffness relations of all elastoplastic and elastic damage models.

2. *Hyperelastic model in principal coordinates*: In this case, the strain energy function is expressed in terms of the principal strain values,

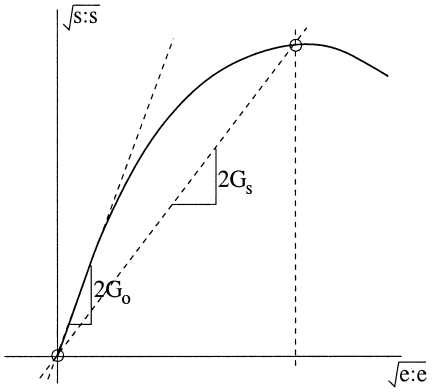


FIGURE 5 Nonlinear deviatoric response behavior.

$$W = W(\epsilon_1, \epsilon_2, \epsilon_3) = W(\epsilon_2, \epsilon_3, \epsilon_1) = W(\epsilon_3, \epsilon_1, \epsilon_2) \tag{55}$$

whereby the underlying isotropy infers cyclic permutation of indices. Because of the underlying coaxiality of the principal axes of stress and strain, the stress–strain relation involves only their principal values; that is,

$$\sigma_i = \frac{\partial W(\epsilon_1, \epsilon_2, \epsilon_3)}{\partial \epsilon_i} \tag{56}$$

The tangential stress–strain relation leads to the following matrix representation of the principal coordinates:

$$\begin{bmatrix} \dot{\sigma}_1 \\ \dot{\sigma}_2 \\ \dot{\sigma}_3 \end{bmatrix} = \begin{bmatrix} \frac{\partial^2 W}{\partial \epsilon_1 \partial \epsilon_1} & \frac{\partial^2 W}{\partial \epsilon_1 \partial \epsilon_2} & \frac{\partial^2 W}{\partial \epsilon_1 \partial \epsilon_3} \\ \frac{\partial^2 W}{\partial \epsilon_2 \partial \epsilon_1} & \frac{\partial^2 W}{\partial \epsilon_2 \partial \epsilon_2} & \frac{\partial^2 W}{\partial \epsilon_2 \partial \epsilon_3} \\ \frac{\partial^2 W}{\partial \epsilon_3 \partial \epsilon_1} & \frac{\partial^2 W}{\partial \epsilon_3 \partial \epsilon_2} & \frac{\partial^2 W}{\partial \epsilon_3 \partial \epsilon_3} \end{bmatrix} \begin{bmatrix} \dot{\epsilon}_1 \\ \dot{\epsilon}_2 \\ \dot{\epsilon}_3 \end{bmatrix} \tag{57}$$

This illustrates the symmetry of the tangential stiffness properties if the strain energy function is sufficiently smooth. Moreover, the tangential stiffness properties are positive definite if the strain energy function remains convex. Although the tangential relation appears to be anisotropic in principal coordinates, the nonlinear stress–strain rate relation maintains coaxiality between $\boldsymbol{\sigma} - \boldsymbol{\epsilon}$ as long as the tangential shear terms satisfy the condition (Ogden, 1984):

$$\begin{bmatrix} \dot{\tau}_{12} \\ \dot{\tau}_{23} \\ \dot{\tau}_{31} \end{bmatrix} = \begin{bmatrix} \frac{\sigma_2 - \sigma_1}{\epsilon_2 - \epsilon_1} & 0 & 0 \\ 0 & \frac{\sigma_3 - \sigma_2}{\epsilon_3 - \epsilon_2} & 0 \\ 0 & 0 & \frac{\sigma_1 - \sigma_3}{\epsilon_1 - \epsilon_3} \end{bmatrix} \begin{bmatrix} \dot{\gamma}_{12} \\ \dot{\gamma}_{23} \\ \dot{\gamma}_{31} \end{bmatrix} \tag{58}$$

In the past, the principal coordinate representation of nonlinear stress–strain relations has been popularized under the name of "orthotropic" material models (see Chen and Han, 1988), though strictly speaking the constitutive format is isotropic in which the strain-induced orthotropy appears because of the different levels of nonlinearity in the principal coordinates.

3. Differential Format

The differential description of elasticity starts from representation theorems of nonlinear tensor functions which lead in the simplest case of grade-one materials to an incrementally linear format of tangential stiffness. The hypoelastic terminology is attributed to Clifford Truesdell (1919–2000), although the linear format of incremental elasticity is normally referred to as the tangential stiffness model.

a. Truesdell elasticity ($\dot{\boldsymbol{\sigma}} = \boldsymbol{g}(\boldsymbol{\sigma}, \dot{\boldsymbol{\epsilon}})$). In the differential format of elasticity, the stress rate is expanded into a symmetric tensor function of two second-order tensors. In the case of a stress-based formulation, the two independent arguments of the tensor function are the stress and the rate of strain tensors. Invoking the argument of material objectivity and frame indifference, the irreducible set of base tensors encompasses the following terms:

$$\begin{aligned} &\mathbf{1}, \boldsymbol{\sigma}, \boldsymbol{\sigma}^2, \dot{\boldsymbol{\epsilon}}, \dot{\boldsymbol{\epsilon}}^2, (\boldsymbol{\sigma} \cdot \dot{\boldsymbol{\epsilon}} + \dot{\boldsymbol{\epsilon}} \cdot \boldsymbol{\sigma}), (\boldsymbol{\sigma} \cdot \dot{\boldsymbol{\epsilon}}^2 + \dot{\boldsymbol{\epsilon}}^2 \cdot \boldsymbol{\sigma}), \\ &(\boldsymbol{\sigma}^2 \cdot \dot{\boldsymbol{\epsilon}} + \dot{\boldsymbol{\epsilon}} \cdot \boldsymbol{\sigma}^2), (\boldsymbol{\sigma}^2 \cdot \dot{\boldsymbol{\epsilon}}^2 + \dot{\boldsymbol{\epsilon}}^2 \cdot \boldsymbol{\sigma}^2) \end{aligned} \tag{59}$$

Conseqently, an isotropic tensor function of two symmetric tensors involves in the most general case nine response functions, $\phi_1, \ldots, \phi_9$, which depend in turn on the six moment invariants of the stress and strain rate tensors below,

$$\begin{aligned} &\bar{I}_1^{\sigma} = tr\boldsymbol{\sigma}, \qquad \bar{I}_2^{\sigma} = tr\boldsymbol{\sigma}^2, \qquad \bar{I}_3^{\sigma} = tr\boldsymbol{\sigma}^3; \\ &\bar{I}_1^{\dot{\epsilon}} = tr\dot{\boldsymbol{\epsilon}}, \qquad \bar{I}_2^{\dot{\epsilon}} = tr\dot{\boldsymbol{\epsilon}}^2, \qquad \bar{I}_3^{\dot{\epsilon}} = tr\dot{\boldsymbol{\epsilon}}^3 \end{aligned} \tag{60}$$

as well as on the four joint invariants:

$$\begin{aligned} &\bar{J}_1 = tr(\boldsymbol{\sigma} \cdot \dot{\boldsymbol{\epsilon}}), \qquad \bar{J}_2 = tr(\boldsymbol{\sigma}^2 \cdot \dot{\boldsymbol{\epsilon}}), \\ &\bar{J}_3 = tr(\boldsymbol{\sigma} \cdot \dot{\boldsymbol{\epsilon}}^2), \qquad \bar{J}_4 = tr(\boldsymbol{\sigma}^2 \cdot \dot{\boldsymbol{\epsilon}}^2) \end{aligned} \tag{61}$$

The general format results in the general stress-strain rate relation,

$$\begin{aligned} \dot{\boldsymbol{\sigma}} = {} & \phi_1 \mathbf{1} + \phi_2 \boldsymbol{\sigma} + \phi_3 \boldsymbol{\sigma}^2 + \phi_4 \dot{\boldsymbol{\epsilon}} + \phi_5 \dot{\boldsymbol{\epsilon}}^2 \\ & + \phi_6 (\boldsymbol{\sigma} \cdot \dot{\boldsymbol{\epsilon}} + \dot{\boldsymbol{\epsilon}} \cdot \boldsymbol{\sigma}) + \phi_7 (\boldsymbol{\sigma} \cdot \dot{\boldsymbol{\epsilon}}^2 + \dot{\boldsymbol{\epsilon}}^2 \cdot \boldsymbol{\sigma}) \\ & + \phi_8 (\boldsymbol{\sigma}^2 \cdot \dot{\boldsymbol{\epsilon}} + \dot{\boldsymbol{\epsilon}} \cdot \boldsymbol{\sigma}^2) + \phi_9 (\boldsymbol{\sigma}^2 \cdot \dot{\boldsymbol{\epsilon}}^2 + \dot{\boldsymbol{\epsilon}}^2 \cdot \boldsymbol{\sigma}^2) \end{aligned} \tag{62}$$

For rate independence, the expansion must be homogeneous of order one, thus the rate terms of the tensor function must be restricted to the first order. In other terms, the hypoelastic material law is rate independent if and only if:

$$\boldsymbol{g}(\boldsymbol{\sigma}, \alpha \dot{\boldsymbol{\epsilon}}) = \alpha \boldsymbol{g}(\boldsymbol{\sigma}, \dot{\boldsymbol{\epsilon}}) \tag{63}$$

Among the numerous possibilities, two classes of hypoelastic constitutive models may be distinguished:

1. *Incrementally linear hypoelastic models*: The linear restriction of the hypoelastic stress–strain relations leads to the classical tangential stiffness format:

$$\dot{\boldsymbol{\sigma}} = \mathcal{E}_t : \dot{\boldsymbol{\epsilon}} \quad \text{where} \quad \mathcal{E}_t = \mathcal{E}(\boldsymbol{\sigma}) \tag{64}$$

This stress-based format is reversible in the small, but not in the large. In other words, the classical hypoelastic formulation leads to path-dependence:

$$\boldsymbol{\sigma} = \int_{\epsilon} \mathcal{E}_t(\boldsymbol{\sigma}) : \frac{d\boldsymbol{\epsilon}}{dt} dt \tag{65}$$

This infers energy dissipation and irreversible behavior for arbitrary load histories as opposed to hyperelasticity.

In fact, the hyperelastic property of path independence is recovered only if appropriate integrability conditions are satisfied which assure that the stress is the gradient of a single potential function (i.e., $\boldsymbol{\sigma} = \frac{\partial W}{\partial \boldsymbol{\epsilon}}$). The most general format of the hypoelastic tangent operator involves 12 hypoelastic response functions which depend in general on all ten stress invariants, $C_i = C_i(I_j, J_k)$. The tensorial structure involves 12 fourth-order tensor products between the second-order unit tensor and stress tensors up to the fourth power:

$$\boldsymbol{\mathcal{E}}_t = \begin{bmatrix} C_1\mathbf{1}\otimes\mathbf{1} & +C_2\boldsymbol{\sigma}\otimes\mathbf{1} & +C_3\boldsymbol{\sigma}^2\otimes\mathbf{1} \\ +C_4\mathbf{1}\otimes\boldsymbol{\sigma} & +C_5\boldsymbol{\sigma}\otimes\boldsymbol{\sigma} & +C_6\boldsymbol{\sigma}^2\otimes\boldsymbol{\sigma} \\ +C_7\mathbf{1}\otimes\boldsymbol{\sigma}^2 & +C_8\boldsymbol{\sigma}\otimes\boldsymbol{\sigma}^2 & +C_9\boldsymbol{\sigma}^2\otimes\boldsymbol{\sigma}^2 \\ +C_{10}[\mathbf{1}\bar{\otimes}\mathbf{1}+\mathbf{1}\underline{\otimes}\mathbf{1}] & +C_{11}[\boldsymbol{\sigma}\bar{\otimes}\mathbf{1}+\mathbf{1}\underline{\otimes}\boldsymbol{\sigma}] & +C_{12}[\boldsymbol{\sigma}^2\bar{\otimes}\mathbf{1}+\mathbf{1}\underline{\otimes}\boldsymbol{\sigma}^2] \end{bmatrix} \tag{66}$$

Consequently, the tangential stiffness operator of the nonlinear K–G model in Section II.B.2 forms a very limited subclass as it activates only three out of the 12 terms in the general framework of incrementally linear hypoelasticity. Under the name of variable moduli models, a good number of simplified hypoelastic material models have been proposed and are still used in structural and geotechnical engineering.

2. *Incrementally nonlinear hypoelastic models*: Another rate-independent restriction leads to a class of incrementally nonlinear models which have been proposed under the name of hypoplastic models. Because of the incremental nonlinearity they are capable of distinguishing between different loading and unloading stiffness properties in analogy to the endochronic time model introduced by Valanis (1975). In the absence of a loading function, it is understood that the irreversible contribution leads to continuous energy dissipation under repeated load cycles in contrast to unload–reload cycles in elastoplasticity.

III. ELASTOPLASTIC MODELS

There are several points of departure to develop a consistent theory for elastoplastic constitutive models. Traditionally, we start from the underlying rheological model shown in Fig. 6 which lends itself to develop the elastoplastic stress-strain relations for uniaxial conditions. The basic concepts may be generalized using the internal variable theory of Coleman and Gurtin (1967) based on continuum thermodynamics which provides an elegant repository for history-dependent inelasticity. In the sequel, we will confine our discussion to rate-independent material behavior. In this case, the starting point of the internal

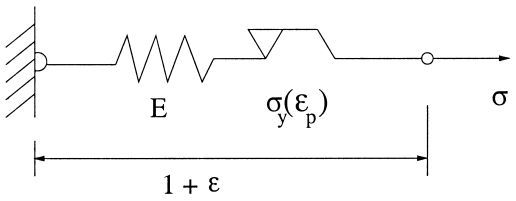


FIGURE 6 Elastoplastic serial model of spring and friction element.

variable models is the assumption that the stress tensor may be represented as a function of the current strain tensor and a finite number of parameters, $q_1, q_2, \ldots, q_n$:

$$\boldsymbol{\sigma} = \boldsymbol{f}(\boldsymbol{\epsilon}, q_1, q_2, \ldots, q_n) \tag{67}$$

These internal variables $q_1, q_2, \ldots, q_n$ represent the material state that depends on the process history. Their current values is defined in terms of evolution equations which form in general a system of ordinary first-order differential equations:

$$\dot{q}_i(t) = g_i(\boldsymbol{\sigma}, \boldsymbol{\epsilon}, q_1(t), q_2(t), \ldots, q_n(t)) \quad \text{where} \quad i = 1, 2, \ldots, n \tag{68}$$

The evolution equations specify the temporal change of the internal variables depending on their current value and input histories either in the form of stress or strain control. For strain control, the general integral of the functional

$$q_i(t) = \mathcal{F}_{t_o \le \tau \le t}(\boldsymbol{\epsilon}(\tau), q_1(t_o), q_2(t_o), \ldots, q_n(t_o)) \tag{69}$$

needs to be evaluated numerically. The internal variables represent the memory properties of the inelastic material behavior. Their physical meaning is to represent the dynamical process which take place because of inelasticity at the material microstructure and to make visible their macroscopic effects. In plasticity, they constitute the plastic strain tensor and the reduced variables for isotropic and kinematic hardening. In elastic damage mechanics they form the damage tensor and the scalar damage variables for volumetric and deviatoric elastic stiffness degradation. Thus, the evolution equations are additional constitutive equations which augment the stress–strain relation.

In the sequel we start from the uniaxial relations of the the rheological model and extend them to triaxial conditions using an engineering approach. For illustration we consider the one-invariant von Mises plasticity model and its extension to the two-invariant Drucker and Prager (1952) model as well as the three-invariant extension along the line of the five parameter model by Willam and Warnke (1975). We will conclude this section with the example

problem of simple shear which will illustrate the dramatic effect of volumetric–deviatoric interaction on the overall response behavior.

A. Elastoplastic Rheological Model

The serial arrangement of an elastic spring and a friction element dates back to Reuss and the additive decomposition of strain into elastic and plastic components. The combination of an elastic spring and the friction element leads to two formats of elastoplasticity, one is the "deformation theory" of Hencky, and the other is the "rate theory" of Prandtl–Reuss. The deformation theory is essentially a secant-type formulation of plasticity along the nonlinear K–G model which is augmented by a plastic load-elastic unload condition. This leads to discontinuities in the transition region from elasticity to plasticity under repeated loading and unloading cycles. Thus, we will concentrate on the flow theory of plasticity, although the continuity leads to overly stiff elastic response predictions when plastic loading takes place to the side of a smooth yield surface without corners.

In small deformation problems, the flow theory is based on the additive decomposition of the total strain rate into elastic and plastic components,

$$\dot{\epsilon} = \dot{\epsilon}_e + \dot{\epsilon}_p \tag{70}$$

The serial arrangement of an elastic spring and a plastic friction element infers stress equivalence in the two elements. In other terms, the state of stress in the elastic spring is limited by the slip conditions of the friction element. The behavior in the elastic response regime $\sigma \leq \sigma_y$ is described by Hooke's law:

$$\dot{\epsilon}_e = \frac{\dot{\sigma}}{E} \tag{71}$$

while the plastic response is active when the stress reaches the yield condition, such that under persistent plastic loading when $\sigma = \sigma_y$,

$$\dot{\epsilon}_p = \frac{\dot{\sigma}}{E_p} \tag{72}$$

Consequently,

$$\dot{\epsilon} = \frac{\dot{\sigma}}{E} + \frac{\dot{\sigma}}{E_p} = \frac{\dot{\sigma}}{E_{ep}} \tag{73}$$

with the elastoplastic tangent stiffness,

$$E_{ep} = \frac{EE_p}{E + E_p} \tag{74}$$

Figure 7 illustrates the elastoplastic response for a load–unload–reload input history. Note that the tangent modulus $E_t = E_{ep}$ may range from positive to negative values

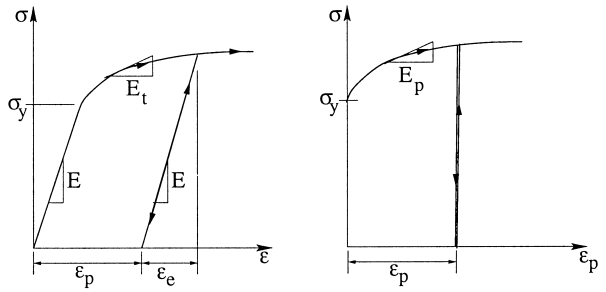


FIGURE 7 Load–unload–reload response of elastoplastic hardening solid.

according to:

$$E_{ep} \begin{cases} >0, & \text{hardening} \\ =0, & \text{perfectly plastic} \\ <0, & \text{softening} \end{cases}$$

with the understanding that a critical value for softening is reached when $E_p^{crit} = -E$.

We also note the indeterminancy of the plastic strain rate for perfectly plastic behavior, when the plastic modulus is zero, $\dot{\epsilon}_p = \frac{0}{0}$. This indicates that strain rather than stress control must be used to maintain uniqueness for elastic–perfectly plastic conditions as well as for strain softening:

$$\dot{\epsilon}_p = \frac{\dot{\sigma}}{E_p} = \frac{E}{E + E_p}\dot{\epsilon} \tag{75}$$

This assures that the plastic strain rate increases with increasing strain rate and vice versa. Moreover, the intuitive loading condition must be replaced by a more precise yield condition of the form of the scalar-valued yield function of stress:

$$F(\sigma) = |\sigma| - \sigma_y = 0 \tag{76}$$

which acts as a threshold condition when the stress demand $|\sigma|$ reaches the yield capacity of the elastoplastic material. We note that plastic loading requires that (1) the stress path has to reach the yield strength, which is here assumed to be the same in uniaxial tension and compression; and (2) the rate of strain must be such that under persistent plastic action $\epsilon\,\dot{\epsilon} > 0$ under strain control.

B. General Format of the Flow Theory of Elastoplasticity

Starting from the small strain setting, where the strain tensor is the symmetric part of the displacement gradient, $\boldsymbol{\epsilon} = \frac{1}{2}[\nabla \boldsymbol{u} + \nabla^t \boldsymbol{u}]$, the serial arrangement of the rheological model in Fig. 6 decomposes the strain into the sum of elastic and plastic parts:

$$\boldsymbol{\epsilon} = \boldsymbol{\epsilon}_e + \boldsymbol{\epsilon}_p \tag{77}$$

1. Elastic Behavior

Extending the hyperelastic concept of a strain energy potential, we start this time from the free energy function Ψ according to Helmholtz, in which the plastic strains and the internal variable κ account for the hardening behavior:

$$\Psi = \Psi(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}_p, \kappa) \tag{78}$$

With the classical arguments of Clausius–Duhem, the inequality of the second law of thermodynamics yields the stress tensor $\boldsymbol{\sigma}$ as a thermodynamically conjugate variable to the elastic strains,

$$\boldsymbol{\sigma} = \frac{\partial \Psi}{\partial \boldsymbol{\epsilon}_e} \quad \text{and} \quad \dot{\boldsymbol{\sigma}} = \boldsymbol{\mathcal{E}} : [\dot{\boldsymbol{\epsilon}} - \dot{\boldsymbol{\epsilon}}_p] \tag{79}$$

Thereby the relation between the stress rates and the elastic strain rates is defined in terms of the fourth-order elasticity tensor $\boldsymbol{\mathcal{E}}$ discussed earlier on in Section II.

2. Plastic Yield Condition

The yield function

$$F(\boldsymbol{\sigma}, \kappa) = f(\boldsymbol{\sigma}) - r_y(\kappa) \leq 0 \quad \text{with} \quad \boldsymbol{n} = \frac{\partial F}{\partial \boldsymbol{\sigma}} \tag{80}$$

delimits the elastic domain with the normal $\boldsymbol{n}$ with respect to the tangent plane at the yield surface. The geometric interpretation of the yield function in the form of a yield surface helps to visualize the elastic region which may expand or shrink according to the underlying hardening/softening model. Here, the yield resistance r_y is conjugate to the internal variable for isotropic hardenening κ:

$$r_y = \frac{\partial \Psi}{\partial \kappa} \quad \text{and} \quad \dot{r}_y = H_p \dot{\kappa}, \tag{81}$$

Thus, the relation between the rate of the yield resistance and the hardening variable defines the hardening modulus H_p.

3. Plastic Flow Rule

The evolution of plastic strains is governed by the plastic flow rule,

$$\dot{\boldsymbol{\epsilon}}_p = \dot{\lambda} \boldsymbol{m} \quad \text{with} \quad \boldsymbol{m} = \frac{\partial Q}{\partial \boldsymbol{\sigma}} \tag{82}$$

whereby $\boldsymbol{m}$ denotes the normal to the plastic potential Q which differs from the yield function F in the nonassociated case. In the above equation, $\dot{\lambda}$ denotes the plastic multiplier, and $Q = Q(\boldsymbol{\sigma})$ is the plastic flow potential. We speak of associated flow when $\boldsymbol{m} \parallel \boldsymbol{n}$ which infers normality of plastic flow in the direction of the gradient of the yield function. In the case of nonassociated flow, $Q \neq F$, one speaks of loss of normality.

4. Plastic Consistency Condition

For plastic behavior not only must the yield condition hold ($F = 0$), but the plastic consistency condition must also be satisfied under persistent plastic action. The so-called Prager-consistency condition $\dot{F} = 0$ enforces the stress path to remain on the yield surface (at least in the differential sense).

The set of constitutive equations is completed by the Kuhn–Tucker condition for plastic loading:

$$F \leq 0 \quad \dot{\lambda} \geq 0 \quad F\dot{\lambda} = 0 \tag{83}$$

Together with the plastic consistency condition,

$$\dot{F} = 0. \tag{84}$$

the plastic multiplier results in:

$$\dot{\lambda} = \frac{1}{h_p} \boldsymbol{n} : \boldsymbol{\mathcal{E}} : \dot{\boldsymbol{\epsilon}} \quad \text{with} \quad h_p = H_p + \boldsymbol{n} : \boldsymbol{\mathcal{E}} : \boldsymbol{m} \tag{85}$$

5. Elastoplastic Stiffness Relation

Substituting the plastic multiplier into the elastoplastic stress–strain relation yields:

$$\dot{\boldsymbol{\sigma}} = \boldsymbol{\mathcal{E}} : [\dot{\boldsymbol{\epsilon}} - \dot{\lambda} \boldsymbol{m}] = \boldsymbol{\mathcal{E}} : \left[\dot{\boldsymbol{\epsilon}} - \boldsymbol{m} \frac{\boldsymbol{n} : \boldsymbol{\mathcal{E}} : \dot{\boldsymbol{\epsilon}}}{H_p + \boldsymbol{n} : \boldsymbol{\mathcal{E}} : \boldsymbol{m}} \right] \tag{86}$$

Rearranging results in the elastoplastic tangent operator $\boldsymbol{\mathcal{E}}_{ep}$ which relates the stress and strain rates:

$$\dot{\boldsymbol{\sigma}} = \boldsymbol{\mathcal{E}}_{ep} : \dot{\boldsymbol{\epsilon}} \tag{87}$$

in the form of a rank-one update of the elastic material operator,

$$\boldsymbol{\mathcal{E}}_{ep} = \boldsymbol{\mathcal{E}} - \frac{1}{h_p} \boldsymbol{\mathcal{E}} : \boldsymbol{m} \otimes \boldsymbol{n} : \boldsymbol{\mathcal{E}} = \boldsymbol{\mathcal{E}} - \frac{1}{h_p} \bar{\boldsymbol{m}} \otimes \bar{\boldsymbol{n}} \tag{88}$$

where $\bar{\boldsymbol{m}} = \boldsymbol{\mathcal{E}} : \boldsymbol{m}$ and $\bar{\boldsymbol{n}} = \boldsymbol{n} : \boldsymbol{\mathcal{E}}$. The update notation emphasizes that the elastic reference tensor is being reduced by the plastic rank-one modification $\bar{\boldsymbol{m}} \otimes \bar{\boldsymbol{n}}$. Note that the critical softening modulus is reached when $h_p^{crit} = 0$ or, in other words, when $H_p^{crit} = -\boldsymbol{n} : \boldsymbol{\mathcal{E}} : \boldsymbol{m}$. We also note the loss of symmetry (i.e., $\boldsymbol{\mathcal{E}}_{ep} \neq \boldsymbol{\mathcal{E}}_{ep}^t$) for nonassociated flow when $\boldsymbol{n} \neq \boldsymbol{m}$.

C. Special case of J_2-Plasticity

In J_2-plasticity we combine the von Mises condition of plastic yielding with the plastic flow rule and the additive decomposition of Prandtl–Reuss. In view of the deviatoric setting of plasticity we may restrict the development of J_2-plasticity to a deviatoric stress–strain relation. For the sake of simplicity we describe the resistance of the material by the yield strength in uniaxial tension, $r_y = \sigma_y$,

and we assume elastic perfectly plastic behavior to eliminate the need for internal variables to describe hardening/softening.

1. *J_2-yield function*:

$$F(\boldsymbol{s}) = \frac{1}{2}\boldsymbol{s} : \boldsymbol{s} - \frac{1}{3}\sigma_y^2 = 0 \tag{89}$$

2. *Associated plastic flow rule*:

$$\dot{\epsilon}_p = \dot{\lambda}\boldsymbol{s} \quad \text{where} \quad \boldsymbol{m} = \frac{\partial F}{\partial \boldsymbol{s}} = \boldsymbol{s} \tag{90}$$

3. *Plastic consistency condition*:

$$\dot{F} = \frac{\partial F}{\partial \boldsymbol{s}} : \dot{\boldsymbol{s}} = \boldsymbol{s} : \dot{\boldsymbol{s}} = 0 \tag{91}$$

The deviatoric stress–strain rates are related by:

$$\dot{\boldsymbol{s}} = 2G[\dot{\boldsymbol{e}} - \dot{\boldsymbol{e}}_p] = 2G[\dot{\boldsymbol{e}} - \dot{\lambda}\boldsymbol{s}] \tag{92}$$

Substituting into the consistency condition,

$$\dot{F} = 2G\boldsymbol{s} : [\dot{\boldsymbol{e}} - \dot{\lambda}\boldsymbol{s}] = 0 \tag{93}$$

leads to the plastic multiplier:

$$\dot{\lambda} = \frac{\boldsymbol{s} : \dot{\boldsymbol{e}}}{\boldsymbol{s} : \boldsymbol{s}} \tag{94}$$

4. *Deviatoric stress–strain relation*: The deviatoric stress–strain rate expression may be written as:

$$\dot{\boldsymbol{s}} = 2G\left[\dot{\boldsymbol{e}} - \frac{\boldsymbol{s} : \dot{\boldsymbol{e}}}{\boldsymbol{s} : \boldsymbol{s}}\boldsymbol{s}\right] = 2G\left[\boldsymbol{\mathcal{I}} - \frac{\boldsymbol{s} \otimes \boldsymbol{s}}{\boldsymbol{s} : \boldsymbol{s}}\right] : \dot{\boldsymbol{e}} \tag{95}$$

or in short form as:

$$\dot{\boldsymbol{s}} = \boldsymbol{\mathcal{G}}_{ep} : \dot{\boldsymbol{e}} \quad \text{with} \quad \boldsymbol{\mathcal{G}}_{ep} = 2G\left[\boldsymbol{\mathcal{I}} - \frac{\boldsymbol{s} \otimes \boldsymbol{s}}{\boldsymbol{s} : \boldsymbol{s}}\right] \tag{96}$$

5. *Tangent stiffness operator of J_2-elastoplasticity*: Including the elastic volumetric response, $[tr\dot{\boldsymbol{\sigma}}] = 3K[tr\dot{\epsilon}]$, the elastoplastic stress–strain relationship leads to:

$$\dot{\boldsymbol{\sigma}} = \frac{1}{3}[tr\dot{\boldsymbol{\sigma}}]\mathbf{1} + \dot{\boldsymbol{s}} = K[tr\dot{\epsilon}]\mathbf{1} + \boldsymbol{\mathcal{G}}_{ep} : \dot{\boldsymbol{e}} \tag{97}$$

Considering

$$\dot{\boldsymbol{\sigma}} = K[tr\dot{\epsilon}]\mathbf{1} + \boldsymbol{\mathcal{G}}_{ep} : \left[\dot{\epsilon} - \frac{1}{3}[tr\dot{\epsilon}]\mathbf{1}\right] \tag{98}$$

the combined elastoplastic tangent relationship results in:

$$\dot{\boldsymbol{\sigma}} = \boldsymbol{\mathcal{E}}_{ep} : \dot{\epsilon} \quad \text{with} \quad \boldsymbol{\mathcal{E}}_{ep} = \Lambda\mathbf{1} \otimes \mathbf{1} + 2G\left[\boldsymbol{\mathcal{I}} - \frac{\boldsymbol{s} \otimes \boldsymbol{s}}{\boldsymbol{s} : \boldsymbol{s}}\right] \tag{99}$$

which has the same constitutive structure as the nonlinear K–G model in Eq. (54).

D. Example Problem of Simple Shear

In frictional materials the so-called Reynolds effect of shear dilatancy is one of the central mechanisms responsible for the interaction of the volumetric and deviatoric components. In the case of plastic loading, the elastoplastic tangential stiffness relationship may be written as

$$\boldsymbol{\mathcal{E}}_{ep} = \boldsymbol{\mathcal{E}} - \frac{1}{h_p}\boldsymbol{\mathcal{R}}_p \quad \text{where} \quad \boldsymbol{\mathcal{R}}_p = \boldsymbol{\mathcal{E}} : \boldsymbol{m} \otimes \boldsymbol{n} : \boldsymbol{\mathcal{E}} \tag{100}$$

Here, the plastic dyadic tensor product $\boldsymbol{\mathcal{R}}_p$ is the sole repository of stress- or strain-induced anisotropy and for coupling between the direct and shear behavior in the case of isotropic elasticity.

For definiteness let us examine the relation between shear strains and normal stresses under simple shear, a strain-controlled test illustrated in Fig. 8. The objective is to determine the source of coupling of the direct stresses and shear strains in the case of elastoplastic formulations of increasing complexity.

Figure 9 illustrates the fundamentally different response predictions of the pressure-independent von Mises and a pressure-sensitive and dilatant two-invariant plasticity formulation in terms of $I_1 = [tr\boldsymbol{\sigma}]$ and $J_2 = \frac{1}{2}[trs^2]$ as proposed by Drucker and Prager (1952). A parabolic generalization of the original two-parameter model may be developed in terms of the friction angle α_F and the cohesive resistance $r_y = \tau_y$ to define the shear strength in the form:

$$F(I_1, J_2) = J_2 + \alpha_F I_1 - \tau_y^2 = 0 \tag{101}$$

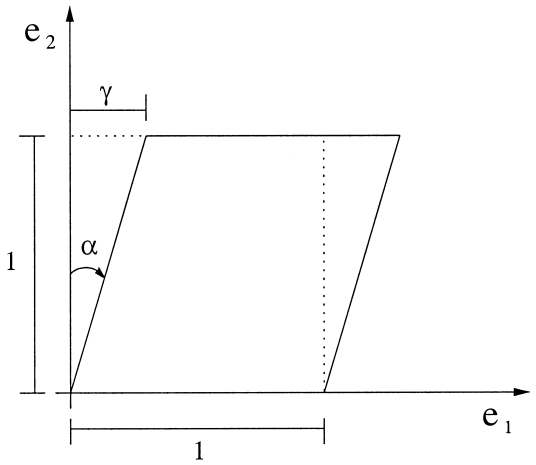


FIGURE 8 Simple shear problem.

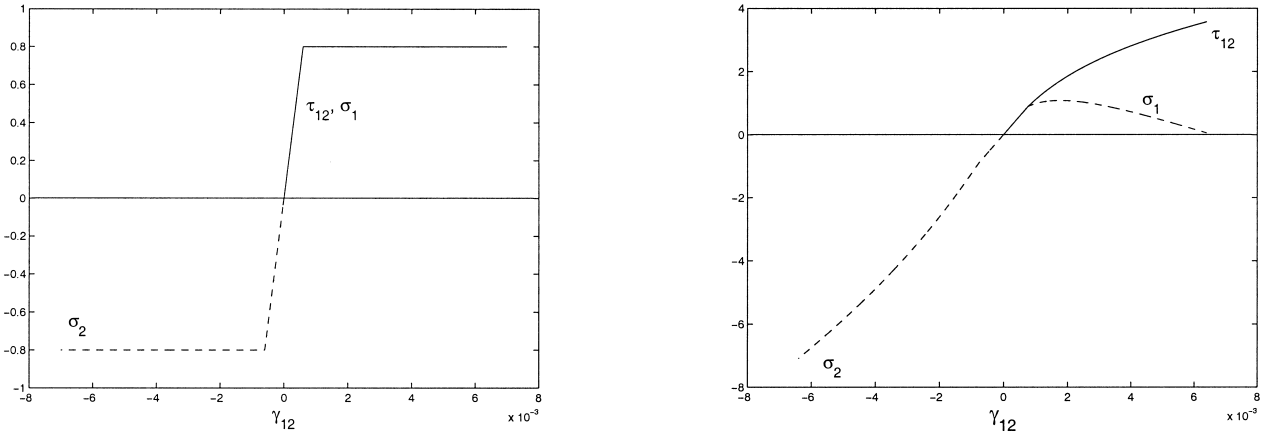


FIGURE 9 Simple shear response behaviour of one-invariant von Mises model (left) and two-invarient parabolic Drucker–Prager model (right).

The two parameters may be identified from uniaxial tension and compression test data f_t' and f_c' as follows:

$$\alpha_F = \frac{f_c' - f_t'}{3} \quad \text{and} \quad \tau_y^2 = \frac{f_c' f_t'}{3} \tag{102}$$

The von Mises yield condition in shear, $\sqrt{J_2} = \tau_y = 0.8\ ksi$, reproduces the elastic–perfectly plastic input, whereby the shear stress–strain relation coincides with the principal stress response. Note, strain control coincides with the stress-controlled situation of pure shear, $\tau_{12} = \sigma_1 = -\sigma_2 = \tau_y$, if there is no coupling between the volumetric and deviatoric response behavior. In contrast, the simple shear test exhibits apparent hardening, when an associated flow rule is used in conjunction with the parabolic Drucker–Prager yield condition. Calibration of the two-parameter representation with the uniaxial tensile and compressive strength values $f_c' = 4\ ksi$ and $f_t' = 0.6\ ksi$ leads to the same incipient shear strength as the von Mises model. However, in contrast to the pressure-independent plasticity model, the parabolic Drucker–Prager model exhibits very different behavior in the plastic regime because of the apparent hardening under persistent plastic flow in spite of the underlying assumption of perfectly plastic behavior. We observe a large increase of shear stress, which is accompanied by an equivalent increase of the minor principal stress $-\sigma_2 \gg \tau_y$ at the cost of reducing the major principal tensile stress from tension into compression.

In view of these fundamental differences of pressure-sensitive vs. pressure-insensitive plasticity models, we further investigate the source of these discrepancies. To this end we systematically examine the influence of the three invariants in the plastic yield condition on the simple shear response and the shear dilatancy in particular.

1. One-Invariant Plasticity Formulation

To start with, we consider the effect of the second invariant in the yield condition, $F = F(\rho)$, where the deviator $\rho = \sqrt{\boldsymbol{s} : \boldsymbol{s}}$ defines the radial distance of the yield condition from the hydrostatic axis of the cylindrical Haigh–Westergaard coordinates ξ, ρ, θ of stress. In this case, the expression for the gradient of the yield function defines the normal as:

$$\boldsymbol{n} = \frac{\partial F}{\partial \boldsymbol{\sigma}} = \frac{\partial F}{\partial \rho}\frac{\partial \rho}{\partial \boldsymbol{\sigma}} = \frac{F_\rho}{\rho}\boldsymbol{s} \quad \text{where} \quad \frac{\partial \rho}{\partial \boldsymbol{\sigma}} = \frac{1}{\rho}\boldsymbol{s} \tag{103}$$

The partial derivative, $F_\rho = \partial F / \partial \rho$, is a positive scalar which determines the magnitude of the normal shown in Fig. 10.

The plastic flow direction is defined by the gradient of the plastic potential. In the case of pressure-insensitive plastic flow

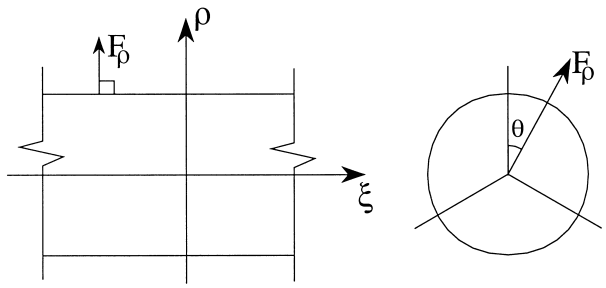


FIGURE 10 Gradient of generic single-invariant yield surface.

$$\boldsymbol{m} = \frac{\partial Q}{\partial \boldsymbol{\sigma}} = \frac{\partial Q}{\partial \rho}\frac{\partial \rho}{\partial \boldsymbol{\sigma}} = \frac{Q_\rho}{\rho}\boldsymbol{s} \tag{104}$$

The expressions for $\boldsymbol{n}$ and $\boldsymbol{m}$ simplify considerably for loading in simple shear when $\dot{\gamma}_{12} > 0$ and when all other strain components remain zero, $\epsilon_{11} = \epsilon_{22} = \epsilon_{33} = \gamma_{23} = \gamma_{13} = 0$. At incipient yielding the elastic shear stress is the only deviatoric component which is nonzero, $s_{12} = \tau_{12} = G\gamma_{12}$. With $\rho = \sqrt{\boldsymbol{s} : \boldsymbol{s}} = \sqrt{2}\tau_{12}$, the gradients of the yield function and plastic potential $\boldsymbol{n}$ and $\boldsymbol{m}$ in vector notation reduce to:

$$\boldsymbol{n} = \begin{bmatrix} 0 & 0 & 0 & \frac{1}{\sqrt{2}}F_\rho & 0 & 0 \end{bmatrix}^t$$

$$\text{and} \quad \boldsymbol{m} = \begin{bmatrix} 0 & 0 & 0 & \frac{1}{\sqrt{2}}Q_\rho & 0 & 0 \end{bmatrix}^t \tag{105}$$

For plastic loading in simple shear the elastoplastic tangent matrix has the simple format:

$$\begin{Bmatrix} \dot{\sigma}_{11} \\ \dot{\sigma}_{22} \\ \dot{\sigma}_{33} \\ \cdots \\ \dot{\tau}_{12} \\ \dot{\tau}_{23} \\ \dot{\tau}_{13} \end{Bmatrix} =$$

$$\left[\begin{array}{ccc:ccc} K+\frac{4}{3}G & K-\frac{2}{3}G & K-\frac{2}{3}G & 0 & 0 & 0 \\ K-\frac{2}{3}G & K+\frac{4}{3}G & K-\frac{2}{3}G & 0 & 0 & 0 \\ K-\frac{2}{3}G & K-\frac{2}{3}G & K+\frac{4}{3}G & 0 & 0 & 0 \\ \hdashline 0 & 0 & 0 & G\left[1-\frac{1}{2h_p}GF_\rho Q_\rho\right] & 0 & 0 \\ 0 & 0 & 0 & 0 & G & 0 \\ 0 & 0 & 0 & 0 & 0 & G \end{array}\right] \begin{Bmatrix} 0 \\ 0 \\ 0 \\ \cdots \\ \dot{\gamma}_{12} \\ 0 \\ 0 \end{Bmatrix} \tag{106}$$

For von Mises plasticity the dyadic tensor product adds only a single term to E_{44}^{ep} on the principal diagonal. Note that the off-diagonal partitions of the elastoplastic tangent operator are zero, similar to isotropic elasticity exhibiting no coupling between the normal stresses and shear strains and vice versa.

For completeness let us examine under what conditions the elastoplastic tangent operator $\boldsymbol{E}_{ep}$ becomes singular. To this end, we partition the elastoplastic tangent matrix into the four components and use the Schur theorem of determinants which states:

$$\det \boldsymbol{E}_{ep} = \det \boldsymbol{E}_{ss} \det\left(\boldsymbol{E}_{nn} - \boldsymbol{E}_{ns}\boldsymbol{E}_{ss}^{-1}\boldsymbol{E}_{sn}\right) \tag{107}$$

A necessary and sufficient condition for $\det \boldsymbol{E}_{ep} = 0$ is reached when either $\det \boldsymbol{E}_{ss} = 0$ or when the determinant of the so-called Schur complement is zero, $\det\left(\boldsymbol{E}_{nn} - \boldsymbol{E}_{ns}\boldsymbol{E}_{ss}^{-1}\boldsymbol{E}_{sn}\right) = 0$. In the case of no coupling, $\boldsymbol{E}_{ns} = \boldsymbol{E}_{sn} = \boldsymbol{0}$, and with the positive partition of the elastic stiffness, $\det \boldsymbol{E}_{nn} > 0$, the only possibility that the elastoplastic tangent operator turns singular arises when $\det \boldsymbol{E}_{ss} = 0$. This condition is true if and only if the diagonal term vanishes (i.e., $E_{44}^{ep} = 0$). Substituting the normals $\boldsymbol{n}$ and $\boldsymbol{m}$ in Eq. (105) into the denominator $h_p = H_p + \boldsymbol{n} : \boldsymbol{E} : \boldsymbol{m} = H_p + \frac{1}{2}GF_\rho Q_\rho$, the plastic degradation of the elastic stiffness results in:

$$E_{44}^{ep} = G\left[1 - \frac{1}{1 + \frac{2H_p}{GF_\rho Q_\rho}}\right] \tag{108}$$

As the shear modulus is strictly positive, $G > 0$, and as $F_\rho Q_\rho > 0$ in the case of the von Mises yield cylinder, the diagonal shear stiffness goes to zero, $E_{44}^{ep} \to 0$, only if the hardening parameter goes to zero, $H_p \to 0$. Consequently,

$$\det \boldsymbol{E}_{ep} = 0 \quad \text{iff} \quad H_p = 0 \tag{109}$$

2. Two-Invariant Plasticity Formulation

In the second stage, we examine the dilatational stress relation of a two-invariant Drucker–Prager yield condition, which includes both hydrostatic and deviatoric plastic effects as shown in Fig. 11.

The normal gradient of the two-invariant formulation reads in the case of two-invariant plasticity:

$$\boldsymbol{n} = \frac{\partial F}{\partial \boldsymbol{\sigma}} = \frac{\partial F}{\partial \xi}\frac{\partial \xi}{\partial \boldsymbol{\sigma}} + \frac{\partial F}{\partial \rho}\frac{\partial \rho}{\partial \boldsymbol{\sigma}} = \frac{F_\xi}{\sqrt{3}}\boldsymbol{1} + \frac{F_\rho}{\rho}\boldsymbol{s} \tag{110}$$

where $\xi = \frac{1}{\sqrt{3}}\boldsymbol{\sigma} : \boldsymbol{1}$ denotes the dependence on the first stress invariant $I_1 = [tr\boldsymbol{\sigma}]$. The scalar derivative is in this case $F_\xi = \partial F/\partial \xi$, while the tensor derivatives are $\partial \xi/\partial \boldsymbol{\sigma} = \frac{1}{\sqrt{3}}\boldsymbol{1}$.

In analogy to the normal $\boldsymbol{n}$, the gradient of the nonassociated plastic potential is expressed as:

$$\boldsymbol{m} = \frac{\partial Q}{\partial \boldsymbol{\sigma}} = \frac{\partial Q}{\partial \xi}\frac{\partial \xi}{\partial \boldsymbol{\sigma}} + \frac{\partial Q}{\partial \rho}\frac{\partial \rho}{\partial \boldsymbol{\sigma}} = \frac{Q_\xi}{\sqrt{3}}\boldsymbol{1} + \frac{Q_\rho}{\rho}\boldsymbol{s} \tag{111}$$

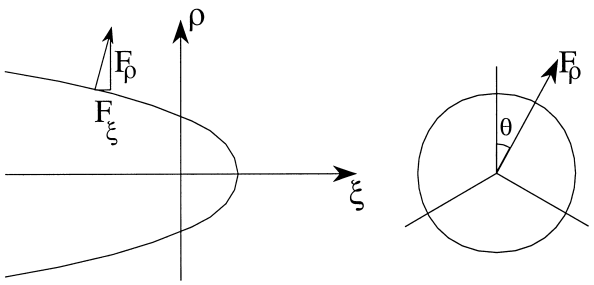


FIGURE 11 Gradient of generic two-invariant yield surface.

In vector form, the two gradients read for simple shear:

$$\boldsymbol{n} = \left[\frac{F_\xi}{\sqrt{3}}, \frac{F_\xi}{\sqrt{3}}, \frac{F_\xi}{\sqrt{3}}, \frac{1}{\sqrt{2}}F_\rho, 0, 0\right]^t$$
$$\text{and} \quad \boldsymbol{m} = \left[\frac{Q_\xi}{\sqrt{3}}, \frac{Q_\xi}{\sqrt{3}}, \frac{Q_\xi}{\sqrt{3}}, \frac{1}{\sqrt{2}}Q_\rho, 0, 0\right]^t \quad (112)$$

The dyadic product $\boldsymbol{\mathcal{R}}_p = \boldsymbol{\mathcal{E}} : \boldsymbol{m} \otimes \boldsymbol{n} : \boldsymbol{\mathcal{E}}$ of the plastic tangent stiffness leads to loss of symmetry if $F_\rho = Q_\rho$ and/or $F_\xi \neq Q_\xi$. The factor 2 in the normal-shear partition is due to the use of the technical shear strain definition in matrix notation.

$$E_{44}^{ep} = G\left[1 - \frac{1}{1 + \frac{2H_p + 6KF_\xi Q_\xi}{GF_\rho Q_\rho}}\right] \quad (115)$$

The elastic bulk and shear moduli (K, G) and the derivatives F_ξ, F_ρ and Q_ξ, Q_ρ are all positive in the conical region of the yield surface and the plastic potential. Consequently, the diagonal term will remain positive, $E_{44}^{ep} > 0$, and the simple shear response exhibits "apparent" hardening in the case of perfectly plastic behavior when $H_p = 0$. The tangent operator becomes singular only when the plastic modulus reaches the limiting softening value:

$$\boldsymbol{R}_p = \begin{bmatrix} 3K^2F_\xi Q_\xi & 3K^2F_\xi Q_\xi & 3K^2F_\xi Q_\xi & \vdots & \sqrt{\frac{3}{2}}KGF_\rho Q_\xi & 0 & 0 \\ 3K^2F_\xi Q_\xi & 3K^2F_\xi Q_\xi & 3K^2F_\xi Q_\xi & \vdots & \sqrt{\frac{3}{2}}KGF_\rho Q_\xi & 0 & 0 \\ 3K^2F_\xi Q_\xi & 3K^2F_\xi Q_\xi & 3K^2F_\xi Q_\xi & \vdots & \sqrt{\frac{3}{2}}KGF_\rho Q_\xi & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \sqrt{\frac{3}{8}}KGF_\xi Q_\rho & \sqrt{\frac{3}{8}}KGF_\xi Q_\rho & \sqrt{\frac{3}{8}}KGF_\xi Q_\rho & \vdots & \frac{1}{2}G^2F_\rho Q_\rho & 0 & 0 \\ 0 & 0 & 0 & \vdots & 0 & 0 & 0 \\ 0 & 0 & 0 & \vdots & 0 & 0 & 0 \end{bmatrix} \quad (113)$$

The hardening parameter in the denominator is comprised of three terms:

$$h_p = H_p + \frac{1}{2}GF_\rho Q_\rho + 3KF_\xi Q_\xi \quad (114)$$

whereby the third contribution introduces an additional term, which is positive as long as $F_\xi Q_\xi > 0$. As expected, now there is coupling between the normal stresses and the shear strain. This shows that plastic loading in the two-invariant formulation results in normal stresses, which are negative as long as $F_\rho Q_\xi > 0$ and vice versa, $F_\xi Q_\rho > 0$. In other terms, pressure-sensitive plastic loading in simple shear induces compromise confinement in the case of Einematic constraints. Thus, the volumetric–deviatoric interaction depends to a large extent on the volumetric component of the plastic flow rule, which is positive in the conical region of the plastic potential and which diminishes to zero (i.e., $Q_\xi \to 0$) when the plastic potential approaches the von Mises condition.

The singularity of the elastoplastic tangent operator $\boldsymbol{E}_{ep}$ is again determined in terms of singular partitions using the Schur theorem of determinants. The diagonal format of the shear partition suggests examination of the diagonal term:

$$H_p^{limit} = -3KF_\xi Q_\xi \quad (116)$$

In summary, softening is needed to compensate for the apparent hardening effect of the volumetric dependence of the yield function and the plastic potential. Therefore, the singularity depends critically on the volumetric components of the yield condition and the plastic potential (i.e., $H_p \to 0$ as $F_\xi Q_\xi \to 0$).

3. Three-Invariant Plasticity Formulation

In the third and final study we generalize the plasticity formulation and include the effect of the third invariant in the yield function and the plastic potential. The effect of the third invariant leads to the triple-symmetric shape of the yield function in the deviatoric plane as illustrated in Fig. 12.

In this case, the gradient operators are becoming considerably more involved in view of the definition of the third invariant, which is here expressed in terms of the angle of similarity:

$$\theta = \frac{1}{3}\cos^{-1}\left[\frac{3\sqrt{3}J_3}{2J_2^{1.5}}\right] \quad (117)$$

where $J_3 = \det \boldsymbol{s}$.

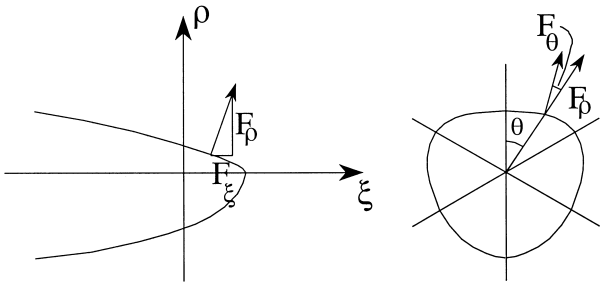


FIGURE 12 Gradients of generic three-invariant yield surface.

The gradients $\boldsymbol{n}$, $\boldsymbol{m}$ are comprised of three contributions; that is,

$$\boldsymbol{n} = \frac{\partial F}{\partial \boldsymbol{\sigma}} = \frac{\partial F}{\partial \xi}\frac{\partial \xi}{\partial \boldsymbol{\sigma}} + \frac{\partial F}{\partial \rho}\frac{\partial \rho}{\partial \boldsymbol{\sigma}} + \frac{\partial F}{\partial \theta}\frac{\partial \theta}{\partial \boldsymbol{\sigma}}$$
$$= \frac{F_\xi}{\sqrt{3}}\mathbf{1} + \frac{F_\rho}{\rho}\boldsymbol{s} + F_\theta\frac{\partial \theta}{\partial \boldsymbol{\sigma}} \tag{118}$$

and

$$\boldsymbol{m} = \frac{\partial Q}{\partial \boldsymbol{\sigma}} = \frac{\partial Q}{\partial \xi}\frac{\partial \xi}{\partial \boldsymbol{\sigma}} + \frac{\partial Q}{\partial \rho}\frac{\partial \rho}{\partial \boldsymbol{\sigma}} + \frac{\partial Q}{\partial \theta}\frac{\partial \theta}{\partial \boldsymbol{\sigma}}$$
$$= \frac{Q_\xi}{\sqrt{3}}\mathbf{1} + \frac{Q_\rho}{\rho}\boldsymbol{s} + Q_\theta\frac{\partial \theta}{\partial \boldsymbol{\sigma}} \tag{119}$$

The additional scalar derivatives include $F_\theta = \frac{\partial F}{\partial \theta}$ and $Q_\theta = \frac{\partial Q}{\partial \theta}$, while the tensor derivative is complicated:

$$\frac{\partial \theta}{\partial \boldsymbol{\sigma}} = \frac{\sqrt{3}}{2\sqrt{1-\cos^2 3\theta}}\left[\frac{1}{J_2^{1.5}}\frac{\partial J_3}{\partial \boldsymbol{\sigma}} - \frac{3J_3}{2J_2^{2.5}}\frac{\partial J_2}{\partial \boldsymbol{\sigma}}\right]$$

$$\text{with} \quad \frac{\partial J_3}{\partial \boldsymbol{\sigma}} = \boldsymbol{s}\cdot\boldsymbol{s} - \frac{2}{3}J_2\mathbf{1} \quad \text{and} \quad \frac{\partial J_2}{\partial \boldsymbol{\sigma}} = \boldsymbol{s} \tag{120}$$

In vector form, the gradients for simple shear appear as:

$$\boldsymbol{n} = \left[\frac{F_\xi}{\sqrt{3}} + \frac{F_\theta}{2\sqrt{3}\tau_{12}}, \frac{F_\xi}{\sqrt{3}} + \frac{F_\theta}{2\sqrt{3}\tau_{12}}, \frac{F_\xi}{\sqrt{3}} - \frac{F_\theta}{\sqrt{3}\tau_{12}}, \frac{1}{\sqrt{2}}F_\rho, 0, 0\right]^t \tag{121}$$

and

$$\boldsymbol{m} = \left[\frac{Q_\xi}{\sqrt{3}} + \frac{Q_\theta}{2\sqrt{3}\tau_{12}}, \frac{Q_\xi}{\sqrt{3}} + \frac{Q_\theta}{2\sqrt{3}\tau_{12}}, \frac{Q_\xi}{\sqrt{3}} - \frac{Q_\theta}{\sqrt{3}\tau_{12}}, \frac{1}{\sqrt{2}}Q_\rho, 0, 0\right]^t \tag{122}$$

In short, the plastic contributions of the third invariant enter the dyadic product $\boldsymbol{\mathcal{R}}_p = \boldsymbol{\mathcal{E}} : \boldsymbol{m} \otimes \boldsymbol{n} : \boldsymbol{\mathcal{E}}$, affecting the coupling partitions $\boldsymbol{\mathcal{R}}^p_{ns}$ and $\boldsymbol{\mathcal{R}}^p_{sn}$ as well as the direct partition $\boldsymbol{\mathcal{R}}^p_{nn}$. To understand their effect, we study the individual terms in the different 3×3 partitions below:

$$\boldsymbol{R}^p_{nn} = \begin{bmatrix} R^p_{11} & R^p_{12} & R^p_{13} \\ R^p_{21} & R^p_{22} & R^p_{23} \\ R^p_{31} & R^p_{32} & R^p_{33} \end{bmatrix} \tag{123}$$

where

$$R^p_{11} = R^p_{22} = R^p_{12} = R^p_{21} = 3K^2F_\xi Q_\xi + \frac{G^2F_\theta Q_\theta}{3\tau_{12}^2} + \frac{KG}{\tau_{12}}(F_\theta Q_\xi + F_\xi Q_\theta) \tag{124}$$

$$R^p_{33} = 3K^2F_\xi Q_\xi + \frac{4G^2F_\theta Q_\theta}{3\tau_{12}^2} - \frac{2KG}{\tau_{12}}(F_\theta Q_\xi + F_\xi Q_\theta) \tag{125}$$

$$R^p_{13} = R^p_{23} = 3K^2F_\xi Q_\xi - \frac{2G^2F_\theta Q_\theta}{3\tau_{12}^2} + \frac{KG}{\tau_{12}}(F_\xi Q_\theta - 2F_\theta Q_\xi) \tag{126}$$

$$R^p_{31} = R^p_{32} = 3K^2F_\xi Q_\xi - \frac{2G^2F_\theta Q_\theta}{3\tau_{12}^2} + \frac{KG}{\tau_{12}}(F_\theta Q_\xi - 2F_\xi Q_\theta) \tag{127}$$

The individual terms indicate that the dependence on the third invariant, $F_\theta Q_\theta$, introduces different terms in the in-plane and out-of-plane stiffness contributions as opposed to the dependence on the first invariant, $F_\xi Q_\xi$. In contrast, the shear partition maintains the same format as the two-invariant model,

$$\boldsymbol{R}^p_{ss} = \begin{bmatrix} \frac{1}{2}G^2F_\rho Q_\rho & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{128}$$

The coupling partitions exhibit a distinct difference between in-plane and out-of-plane stiffness contributions in simple shear:

$$\boldsymbol{R}^p_{ns} = \begin{bmatrix} \sqrt{\frac{3}{2}}KGF_\rho Q_\xi + \frac{G^2}{\sqrt{6}}\frac{F_\rho Q_\theta}{\tau_{12}}, & 0, 0 \\ \sqrt{\frac{3}{2}}KGF_\rho Q_\xi + \frac{G^2}{\sqrt{6}}\frac{F_\rho Q_\theta}{\tau_{12}}, & 0, 0 \\ \sqrt{\frac{3}{2}}KGF_\rho Q_\xi - \frac{2G^2}{\sqrt{6}}\frac{F_\rho Q_\theta}{\tau_{12}}, & 0, 0 \end{bmatrix} \tag{129}$$

and

$$\boldsymbol{R}^p_{sn} = \frac{1}{2}\begin{bmatrix} \sqrt{\frac{3}{2}}KGF_\xi Q_\rho + \frac{G^2}{\sqrt{6}}\frac{Q_\rho F_\theta}{\tau_{12}}, 0, 0 \\ \sqrt{\frac{3}{2}}KGF_\xi Q_\rho + \frac{G^2}{\sqrt{6}}\frac{Q_\rho F_\theta}{\tau_{12}}, 0, 0 \\ \sqrt{\frac{3}{2}}KGF_\xi Q_\rho - \frac{2G}{\sqrt{6}}\frac{Q_\rho F_\theta}{\tau_{12}}, 0, 0 \end{bmatrix}^t \tag{130}$$

We note that the entries of the dyadic product are no longer the same as in the two-invariant formulation. In fact, the negative sign in the third entry introduces additional structure and distinguishes in-plane from out-of-plane action, which reduces significantly the out-of-plane confining stress σ_{33} under simple shear. We also observe the loss of symmetry in all partitions except for the shear partition, when the plastic dilatancy differs from the frictional resistance of the yield condition, $Q_\xi \neq F_\xi$. It might surprise the reader that the shear partition $\boldsymbol{R}^p_{ss}$ remains unchanged except for the denominator h_p which is augmented by an apparent hardening term from the contribution of the third invariant; that is,

$$h_p = H_p + \frac{1}{2}GF_\rho Q_\rho + 3KF_\xi Q_\xi + \frac{3}{2}K\frac{F_\theta Q_\theta}{\tau_{12}^2} \tag{131}$$

Analysis of the singularity in simple shear leads to the question of when the diagonal term vanishes (i.e., $E^{ep}_{44} \to 0$). The answer follows the previous argument of the two invariant formulation. In fact, $\det \boldsymbol{E}_{ep} = 0$ when the plastic softening modulus reaches the limiting value:

$$H_p^{limit} = -3K\left[F_\xi Q_\xi + \frac{1}{2}\frac{F_\theta Q_\theta}{\tau_{12}^2}\right] \tag{132}$$

This indicates that dependence on the third invariant further stabilizes the elastoplastic tangent operator beyond the limit point condition of the two-invariant format.

E. Concrete Plasticity Model

To illustrate the observations above we resort to numerical simulation of the simple shear test with the help of the three-invariant plasticity model by Kang and Willam (1999). In this case, the curvilinear loading surface, $F(\xi, \rho, \theta, q_h, q_s) = 0$, is C^1-continuous except at the vertex in equitriaxial tension (see Fig. 13a. The triaxial concrete model exhibits pressure sensitivity of the deviatoric strength as a function of the third stress-invariant, inelastic dilatancy during shearing and brittle–ductile transition from fragile behavior in tension to ductile behavior with increasing confinement.

The plastic loading function is comprised of three components: the triaxial failure envelope and hardening or softening contributions, which are mobilized alternatively in the pre- and post-peak regions:

$$F(\xi, \rho, \theta, q_h, q_s) = F(\xi, \rho, \theta)_{\text{fail}} + F(\xi, \rho, \theta, k(q_h))_{\text{hardg}} + F(\xi, \rho, \theta, c(q_s))_{\text{softg}} \tag{133}$$

The failure envelope,

$$F(\xi, \rho, \theta)_{\text{fail}} = \frac{\rho r(\theta, e)}{f'_c} - \frac{\rho_1}{f'_c}\left[\frac{\xi - \xi_o}{\xi_1 - \xi_o}\right]^\alpha = 0 \tag{134}$$

fixes the triaxial strength in stress space in terms of a curvilinear triple-symmetric cone. Figure 13a depicts the failure envelope in terms of the meridians in triaxial compression and extension, and Fig. 13b shows the failure envelope in terms of the deviatoric traces at different levels of hydostatic stress.

The simple shear response is shown in Fig. 14 for the three-invariant concrete model. In comparison to the two-invariant formulation shown in Fig. 9, the shear strength

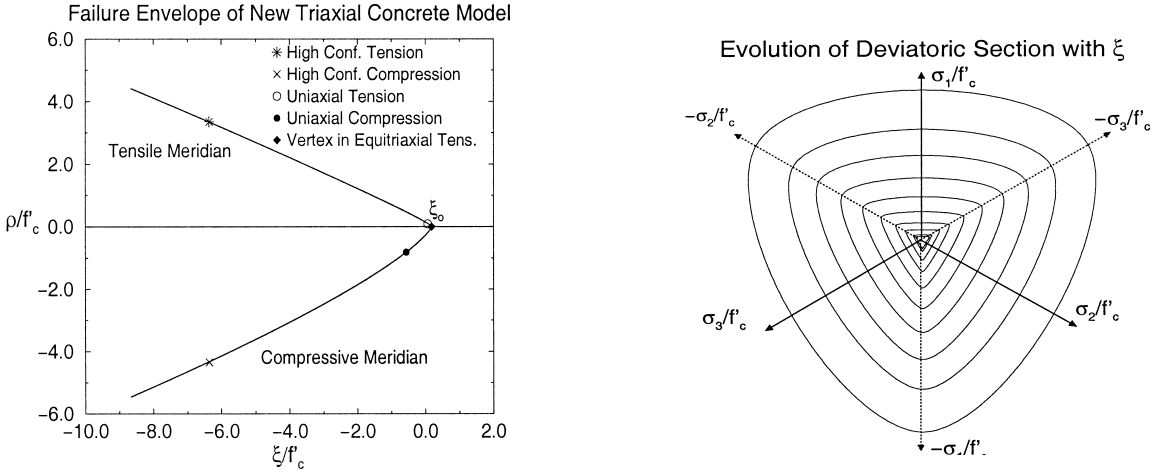


FIGURE 13 (a) Tension and compression meridians, (b) deviatoric contours of triaxial concrete envelope. [From Kang, H., and Willam, K. (1199). *ASCE J. Eng. Mech.* **125,** 941–950. With permition.]

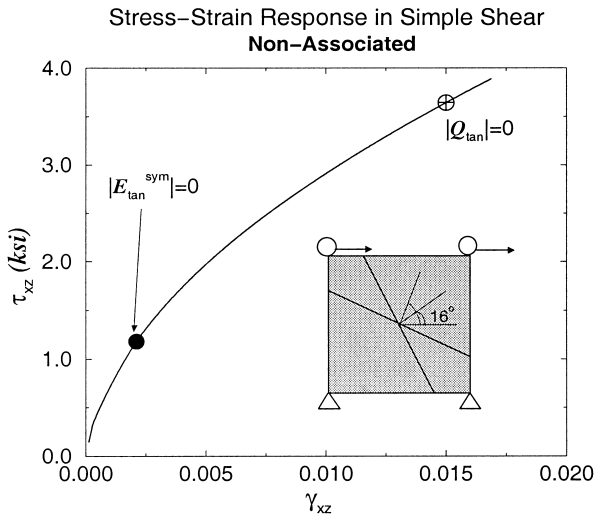


FIGURE 14 Simple shear response of three-invariant concrete model. [From Kang, H., and Willam, K. (1999). *ASCE J. Eng. Mech.* **125,** 941–950. With permission.]

and shear ductility are reduced, because of the diminished out-of-plane confinement of the three-invariant model.

Note that the nonassociated plasticity model exhibits loss of stability ($\det \mathcal{E}_{ep}^{sym} = \det \mathcal{E}_{tan}^{sym} = 0$) and discontinuous bifurcation in the form of localized failure ($\det \boldsymbol{Q}_{ep} = \det \boldsymbol{Q}_{tan} = 0$) in the apparent hardening regime before a limit point condition is reached. For an understanding of the different failure diagnostics, the reader is referred to Section IV and to Appendix 2 for additional comments on elastoplastic failure in plane strain.

IV. ANALYSIS OF MATERIAL FAILURE

In this section, we examine different criteria for initiating failure at the constitutive level of materials. The traditional approach in strength of materials is to probe demand versus resistance with the aid of a limit state condition. This leads to the geometrical visualization of the triaxial state of stress and its proximity to the envelope condition according to Otto Mohr (1835–1918).

Note that the geometrical interpretation of the triaxial strength hypothesis $F = F(\boldsymbol{\sigma}) = 0$ leads to a critical combination of normal and shear tractions σ, σ_T on a critical failure plane $\pm\theta$ which determines the mode of failure. Thus, the Mohr envelope criterion tells us not only when failure takes place, but also what kind of failure mode develops. The main drawback is that it does not depend on the intermediate principal stress since the major principal circle is governed by the maximum and minimum values of stress, $R_{max} = \frac{1}{2}|\sigma_1 - \sigma_3|$. Representative strength criteria for cohesive-frictional materials include the maximum shear stress conditions of Mohr–Coulomb, Tresca, and Leon, as well as the maximum normal stress condition of Rankine for tensile cracking. Alternatively to strength of materials, failure has also been described by deformation hypotheses $F(\boldsymbol{\epsilon}) = 0$ such as the Saint Venant criterion of maximum normal strain for tensile cracking. Moreover, energy criteria $F(\boldsymbol{\sigma} : \boldsymbol{\epsilon}) = 0$ have been proposed to describe material failure such as the Beltrami condition of maximum strain energy and the Huber criterion of maximum distortional energy. In fact, fracture criteria of a critical stress intensity factor or an equivalent strain energy release rate may also be included in this list of failure criteria, $F(K_I, K_{II}, K_{III}) = 0$ and $F(G_f) = 0$, though discrete fracture mechanics normally starts from the existence of a crack or a notched defect and defines crack propagation in terms of fracture concepts. In this sense, failure criteria differ from fracture criteria in a fundamental fashion, the former indicate initiation while the latter monitor propagation.

In the original envelope concept of Mohr (Fig. 15), the actual mechanism behind material failure was left open in terms of its kinematic or static manifestation. There

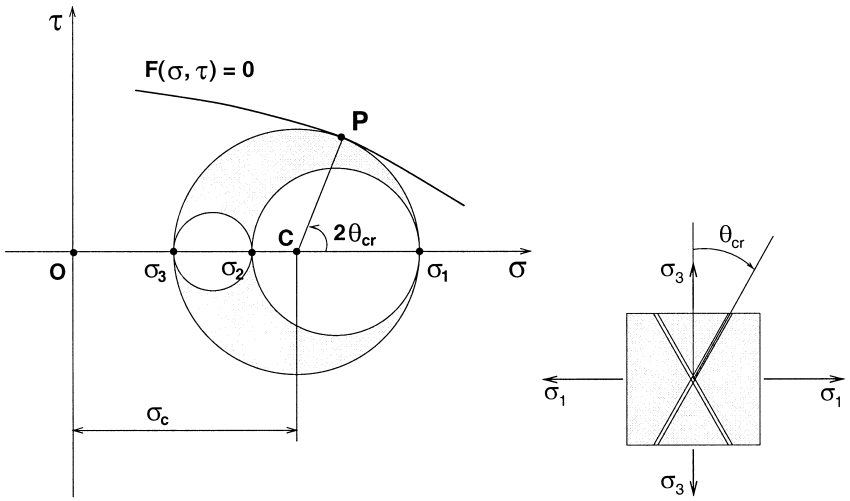


FIGURE 15 Mohr concept of universal failure envelope.

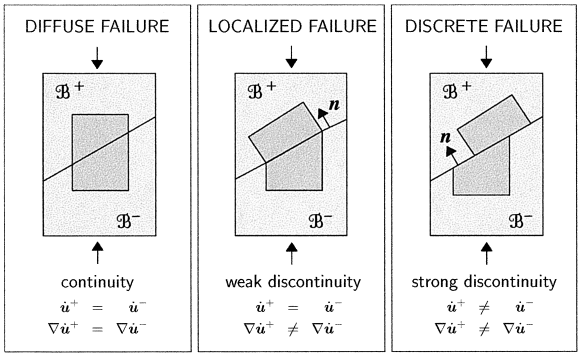


FIGURE 16 Kinematic deterioration of the continuum into a discontinuum.

was the conceptual distinction between tensile cracking and separation of material interfaces in mode I, and frictional slip among adjacent material interfaces in mode II. However, nothing was said about what happens after failure initiates. In fact, failure initiation was assumed to be critical and trigger collapse of the entire structure.

In view of recent progress in localization analysis it is important to keep in mind that failure is a process of events which start small at the material level and lead to progressive deterioration of the continuum into a discontinuum. Figure 16 illustrates how his process leads to a progression of kinematic deterioration. It initiates in the form of diffuse failure and leads through localized, weakly discontinuous failure to discrete or strongly discontinuous failure. The following discussion will be restricted to the first two stages of failure, whereby diffuse failure is characterized by loss of material stability and loss of uniqueness, while localized failure results in the formation of weak discontinuities, which signal the onset of kinematic degradation of the continuum into a discontinuum synonymous with loss of ellipticity.

A. Loss of Stability/Material Instability

The loss of material stability was identified early on by Drucker and Hill with the loss of positive internal work. In fact, the exclusion functional of positive second-order work density,

$$d^2W = \frac{1}{2}\dot{\boldsymbol{\sigma}} : \dot{\boldsymbol{\epsilon}} = \frac{1}{2}\dot{\boldsymbol{\epsilon}} : \boldsymbol{\mathcal{E}}_t : \dot{\boldsymbol{\epsilon}} > 0, \qquad \forall \dot{\boldsymbol{\epsilon}} \neq 0 \quad (135)$$

is widely accepted as a sufficient condition for material stability.

In the case of nonassociated elastoplasticity, where $\boldsymbol{\mathcal{E}}_{ep} \neq \boldsymbol{\mathcal{E}}^t_{ep}$, this criterion leads to:

$$\begin{aligned} d^2W &= \frac{1}{2}\dot{\boldsymbol{\epsilon}} : \boldsymbol{\mathcal{E}}_{ep} : \dot{\boldsymbol{\epsilon}} = \frac{1}{4}\dot{\boldsymbol{\epsilon}} : \left[\boldsymbol{\mathcal{E}}_{ep} + \boldsymbol{\mathcal{E}}^t_{ep}\right] : \dot{\boldsymbol{\epsilon}} \\ &= \frac{1}{2}\dot{\boldsymbol{\epsilon}} : \boldsymbol{\mathcal{E}} : \dot{\boldsymbol{\epsilon}} - \frac{1}{4h_p}\dot{\boldsymbol{\epsilon}} : [\bar{\boldsymbol{m}} \otimes \bar{\boldsymbol{n}} + \bar{\boldsymbol{n}} \otimes \bar{\boldsymbol{m}}] : \dot{\boldsymbol{\epsilon}} \quad (136) \end{aligned}$$

In this context, it is important to recall that the energy functional extracts only the symmetric part of the tangent operator which leads to the following observation based on the Bromwich eigenvalue bounds of nonsymmetric matrices:

$$\lambda_{min}\left(\boldsymbol{\mathcal{E}}^{sym}_{ep}\right) \leq \Re(\lambda_{min}(\boldsymbol{\mathcal{E}}_{ep})) \cdots \leq \lambda_{max}\left(\boldsymbol{\mathcal{E}}^{sym}_{ep}\right) \quad (137)$$

In other terms, the instability argument coincides with the loss of positive definiteness of the symmetric material operator,

$$\det \boldsymbol{\mathcal{E}}^{sym}_{ep} \stackrel{!}{=} 0 \rightarrow \lambda_{min}\left(\boldsymbol{\mathcal{E}}^{sym}_{ep}\right) = 0 \quad (138)$$

This argument leads to the critical value of plastic hardening for stability (see Maier and Hueckl, 1979):

$$H^{stabil}_p = \frac{1}{2}[\sqrt{(\boldsymbol{n} : \boldsymbol{\mathcal{E}}_o : \boldsymbol{n})(\boldsymbol{m} : \boldsymbol{\mathcal{E}}_o : \boldsymbol{m})} - \boldsymbol{n} : \boldsymbol{\mathcal{E}}_o : \boldsymbol{m}] \quad (139)$$

Therefore, the sufficient condition for stability may be turned around into an upper-bound argument of loss of stability in terms of the maximum plastic modulus $H^{stabil}_p \geq H^{limit}_p$.

B. Diffuse Failure/Loss of Uniqueness

According to Fig. 16, diffuse failure maintains continuity in the rate of displacement and the displacement gradient fields. It corresponds to a stationary stress state which defines a limit point on the response path of the material:

$$\dot{\boldsymbol{\sigma}} = \boldsymbol{0} \quad (140)$$

For incrementally linear materials with $\dot{\boldsymbol{\sigma}} = \boldsymbol{\mathcal{E}}_t : \dot{\boldsymbol{\epsilon}}$, this infers that:

$$\boldsymbol{\mathcal{E}}_t : \dot{\boldsymbol{\epsilon}} = \boldsymbol{0} \quad (141)$$

Thus, the condition for the loss of uniqueness is equivalent to a singular behavior of the tangent operator $\boldsymbol{\mathcal{E}}_t$,

$$\det \boldsymbol{\mathcal{E}}_t \stackrel{!}{=} 0 \rightarrow \lambda_{min}(\boldsymbol{\mathcal{E}}_t) = 0 \quad (142)$$

Recall that the elastoplastic tangent operator involves a rank-one update of the elastic material operator,

$$\boldsymbol{\mathcal{E}}_{ep} = \boldsymbol{\mathcal{E}} - \frac{1}{h_p}\bar{\boldsymbol{m}} \otimes \bar{\boldsymbol{n}} \quad (143)$$

whereby the update tensors were defined as follows:

$$\bar{\boldsymbol{m}} = \boldsymbol{\mathcal{E}} : \boldsymbol{m} \quad (144a)$$

$$\bar{\boldsymbol{n}} = \boldsymbol{n} : \boldsymbol{\mathcal{E}} \quad (144b)$$

Preconditioning of the elastoplastic tangent operator with the inverse elasticity tensor, the following relation appears.

$$\mathcal{E}^{-1} : \mathcal{E}^{ep} = \mathcal{I} - \mathcal{E}^{-1} : \frac{\bar{\boldsymbol{m}} \otimes \bar{\boldsymbol{n}}}{h_p} \tag{145}$$

Due to the rank-one update structure of the elastoplastic tangent operator, the critical eigenvalue λ_{min} of the generalized eigenvalue problem may be evaluated in closed form. This motivates the introduction of the scalar-valued measure of material integrity $d_{\mathcal{E}}$,

$$\lambda_{min}(\mathcal{E}^{-1} : \mathcal{E}^{ep}) = 1 - d_{\mathcal{E}} \;\text{ with }\; d_{\mathcal{E}} := \frac{\boldsymbol{n} : \mathcal{E} : \boldsymbol{m}}{H_p + \boldsymbol{n} : \mathcal{E} : \boldsymbol{m}}. \tag{146}$$

From the above definition, a necessary condition for loss of uniqueness leads to the critical hardening modulus $H_p^{limit} = 0$ associated with full loss of integrity:

$$1 - d_{\mathcal{E}} \overset{!}{=} 0 \tag{147}$$

The criterion presented above can only be understood as a necessary condition for the loss of uniqueness. For nonsymmetric elastoplastic tangent operators, which arise in nonassociated plasticity formulations, loss of stability in the form of $\det \mathcal{E}_{ep}^{sym} \overset{!}{=} 0$ provides a lower bound condition according to the Bromwich bounds. Consequently, loss of stability which is synonymous with a singularity of the symmetric tangent operator takes place before the limit point condition in Eq. (142) is reached.

C. Localized Failure/Loss of Ellipticity

The localization condition is based on the early works of Hadamard and Hill. In contrast to the diffuse failure mode described in the previous section, localized failure of weak discontinuities is characterized through a discontinuity in the rate of the displacement gradient, while the field of the displacement rate itself is still continuous. According to the Maxwell compatibility condition, the jump in the rate of the displacement gradient may be expressed in terms of a scalar-valued jump amplitude α, the unit jump vector $\boldsymbol{M}$, and the unit normal vector to the discontinuity surface $\boldsymbol{N}$ shown in Fig. 16:

$$[\![\nabla \dot{\boldsymbol{u}}]\!] = \alpha \boldsymbol{M} \otimes \boldsymbol{N} \rightarrow [\![\dot{\boldsymbol{\epsilon}}]\!] = \alpha [\boldsymbol{M} \otimes \boldsymbol{N}]^{sym} \tag{148}$$

Equilibrium along the discontinuity surface requires that the traction vectors are equal and opposite on both sides of the discontinuity:

$$[\![\dot{\boldsymbol{t}}]\!] = \dot{\boldsymbol{t}}^{+} - \dot{\boldsymbol{t}}^{-} = \boldsymbol{0} \tag{149}$$

According to Cauchy's theorem, we have

$$[\![\dot{\boldsymbol{t}}]\!] = \boldsymbol{N} \cdot [\![\dot{\boldsymbol{\sigma}}]\!] = \boldsymbol{N} \cdot [\![\mathcal{E}_t : \dot{\boldsymbol{\epsilon}}]\!] = \boldsymbol{0} \tag{150}$$

With the assumption of a linear comparison solid, $[\![\mathcal{E}_t]\!] = \mathcal{E}_t^{+} - \mathcal{E}_t^{-} = \boldsymbol{0}$, the localization condition may be expressed in the form of an eigenvalue problem:

$$\alpha\, \boldsymbol{Q}_t \cdot \boldsymbol{M} = \boldsymbol{0} \quad \text{with} \quad \boldsymbol{Q}_t = \boldsymbol{N} \cdot \mathcal{E}_t \cdot \boldsymbol{N} \tag{151}$$

whereby $\boldsymbol{Q}_t$ denotes the tangential localization tensor. The necessary condition for the onset of localization is thus characterized through the singularity of the localization tensor which happens to coincide with the acoustic tensor in wave propagation:

$$\det \boldsymbol{Q}_t \overset{!}{=} 0 \rightarrow \lambda_{min}(\boldsymbol{Q}_t) = 0 \tag{152}$$

The elastoplastic localization tensor may be expressed as a rank-one update of the elastic acoustic tensor,

$$\boldsymbol{Q}_{ep} = \boldsymbol{Q} - \frac{1}{h_p} \boldsymbol{e}_m \otimes \boldsymbol{e}_n \tag{153}$$

whereby the following abbreviations have been introduced for the update vectors $\boldsymbol{e}_m$ and $\boldsymbol{e}_n$:

$$\boldsymbol{e}_m = \boldsymbol{N} \cdot \mathcal{E} : \boldsymbol{m} \tag{154a}$$

$$\boldsymbol{e}_n = \boldsymbol{n} : \mathcal{E} \cdot \boldsymbol{N} \tag{154b}$$

Instead of probing the lowest eigenvalue of the localization tensor, Eq. (152), we resort to the generalized eigenvalue problem $\det[\boldsymbol{Q}^{-1} \cdot \boldsymbol{Q}^{ep}] \overset{!}{=} 0$ in order to detect whether the localization tensor turns singular (see Ottosen and Runesson, 1991). Preconditioning of the elastoplastic acoustic tensor with the inverse of the elastic acoustic tensor leads to:

$$\boldsymbol{Q}^{-1} \cdot \boldsymbol{Q}_{ep} = \boldsymbol{1} - \boldsymbol{Q}^{-1} \cdot \frac{\boldsymbol{e}_m \otimes \boldsymbol{e}_n}{h_p} \tag{155}$$

The closed form solution for the lowest eigenvalue λ_{min} introduces a scalar-valued measure of integrity with respect to localization in the form of:

$$\lambda_{min}\big(\boldsymbol{Q}^{-1} \cdot \boldsymbol{Q}_{ep}\big) = 1 - d_Q \;\text{ with }\; d_Q = \frac{\boldsymbol{e}_n \cdot \boldsymbol{Q}^{-1} \cdot \boldsymbol{e}_m}{H_p + \boldsymbol{n} : \mathcal{E} : \boldsymbol{m}} \tag{156}$$

It defines a necessary condition for localization as well as the critical hardening modulus H_p^{loc} indicating loss of ellipticity, when

$$1 - d_Q \overset{!}{=} 0 \rightarrow H_p^{loc} = \boldsymbol{e}_n \cdot \boldsymbol{Q}^{-1} \cdot \boldsymbol{e}_m - \boldsymbol{n} : \mathcal{E} : \boldsymbol{m} \tag{157}$$

For nonsymmetric elastoplastic tangent operators in nonassociated plasticity, loss of strong ellipticity in terms of a vanishing determinant of the symmetric part of the tangent acoustic operator, $\det \boldsymbol{Q}_{ep}^{sym} \leq \det \boldsymbol{Q}_{ep} \overset{!}{=} 0$, provides a lower bound condition according to the Bromwich theorem. Consequently, loss of strong ellipticity, which is synonymous with a singularity of the symmetric localization operator, may take place before the limit point condition

Eq. (142) is reached. But, most importantly, the critical hardening modulus for localization may still be positive (compare Rudnicki and Rice, 1975) before a limit point is reached because of the symmetric features of the second-order localization tensor when compared to the fourth-order material tensor.

D. Geometric Localization Condition

The analytical localization condition, Eq. (152), may be illustrated geometrically in the form of an envelope condition in analogy to the contact condition of Mohr. In the case of the two-invariant plasticity model, the localization condition plots as an ellipse in the $\sigma - \sigma_T$ Mohr coordinates. The localization envelope was developed by Benallal and Comi (1996) for elastoplastic material models.

In the following, the geometric localization format will be established reproducing some results from Kuhl *et al.* (2000) to illustrate the model problem of simple shear. To this end we briefly summarize the basic ideas of the geometric localization analysis for the nonassociated parabolic Drucker–Prager model introduced in the previous section. Eq. (157) indicates the onset of localization when

$$H_p^{loc} + \boldsymbol{n} : \mathcal{E} : \boldsymbol{m} = \boldsymbol{e}_n \cdot \boldsymbol{Q}^{-1} \cdot \boldsymbol{e}_m \tag{158}$$

Herein, $\mathcal{E}$ denotes the fourth-order elasticity tensor, while $\boldsymbol{Q}^{-1}$ is the inverse of the elastic acoustic tensor which can be determined analytically:

$$\mathcal{E} = \Lambda \mathbf{1} \otimes \mathbf{1} + 2G\mathcal{I} \tag{159a}$$

$$\boldsymbol{Q}^{-1} = \frac{1}{G}\mathbf{1} - \frac{1}{2G[1-\nu]}\boldsymbol{N} \otimes \boldsymbol{N} \tag{159b}$$

In the following, the localization condition, Eq. (158), is recast in the Mohr coordinates σ and σ_T:

$$\sigma := \boldsymbol{N} \cdot \boldsymbol{\sigma} \cdot \boldsymbol{N} \tag{160a}$$

$$\sigma_T^2 := [\boldsymbol{\sigma} \cdot \boldsymbol{N}] \cdot [\boldsymbol{\sigma} \cdot \boldsymbol{N}] - \sigma^2 \tag{160b}$$

Combination of these terms defines the localization ellipse in the Mohr coordinates,

$$\frac{[\sigma - \sigma_o]^2}{A^2} + \frac{\sigma_T^2}{B^2} = 1 \tag{161}$$

where σ_o locates the center, while A and B determine the half axes of the ellipse in the normal and tangential stress components, respectively. In the case of the parabolic Drucker–Prager two-invariant model, the normal to the yield surface $\boldsymbol{n}$ and the normal to the plastic potential $\boldsymbol{m}$ take the form:

$$\boldsymbol{n} = \frac{\partial F}{\partial \boldsymbol{\sigma}} = \boldsymbol{s} + \alpha_F \mathbf{1} \tag{162a}$$

$$\boldsymbol{m} = \frac{\partial Q}{\partial \boldsymbol{\sigma}} = \boldsymbol{s} + \alpha_Q \mathbf{1} \tag{162b}$$

The abbreviations introduced in Eq. (154) result in the following expressions:

$$\boldsymbol{e}_n = 2G\boldsymbol{s} \cdot \boldsymbol{N} + 3K\alpha_F \boldsymbol{N} \tag{163a}$$

$$\boldsymbol{e}_m = 2G\boldsymbol{s} \cdot \boldsymbol{N} + 3K\alpha_Q \boldsymbol{N} \tag{163b}$$

The localization ellipse is defined in terms of the parameters,

$$\begin{aligned} \sigma_o &= \frac{1}{3} I_1 - \frac{1+\nu}{2[1-2\nu]}[\alpha_F + \alpha_Q] \\ A^2 &= 2\frac{1-\nu}{1-2\nu} B^2 \\ B^2 &= \frac{1}{4G} H_p + J_2 + \frac{[1+\nu]^2}{8[1-2\nu][1-\nu]}[\alpha_F + \alpha_Q]^2 \\ &\quad + \frac{1+\nu}{1-\nu}\alpha_F \alpha_Q \end{aligned} \tag{165}$$

where α_F, α_Q denote the friction and dilatancy parameters of the parabolic Drucker–Prager model.

The geometrical properties of the localization ellipse are illustrated in Fig. 17. Note that the center and the shape of the ellipse are not influenced by the plastic hardening modulus H_p. The hardening modulus only influences the size of the ellipse. Figure 17 also depicts the major Mohr circle of stress,

$$[\sigma - \sigma_c]^2 + \sigma_T^2 = R^2 \tag{166}$$

which characterizes the actual stress state. Herein, σ_c and R denote the center and the radius of the circle, respectively, in terms of the principal stresses σ_1 and σ_3:

$$\sigma_c = \frac{\sigma_1 + \sigma_3}{2} \tag{167a}$$

$$R = \frac{\sigma_1 - \sigma_3}{2} \tag{167b}$$

The tangency condition results in a quadratic equation, which defines the analytical solutions for critical failure

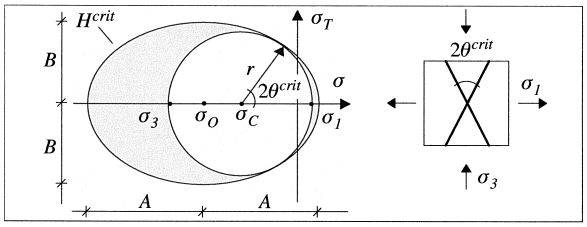


FIGURE 17 Mohr representation of localization ellipse and major stress circle.

angle θ^{crit} and the critical hardening modulus H_p^{loc}. For the nonassociated Drucker–Prager plasticity model, the critical failure angle may be expressed as follows:

$$\tan^2\theta^{crit} = \frac{R - [[1-2\nu][\sigma_c - I_1/3] + [1+\nu][\alpha_F + \alpha_Q]/2]}{R + [[1-2\nu][\sigma_c - I_1/3] + [1+\nu][\alpha_F + \alpha_Q]/2]} \tag{168}$$

The critical failure angle is strongly influenced by the friction coefficient α_F and the dilatancy parameter α_Q. Furthermore, the failure angle is also influenced by Poisson's ratio. The analytical solution for the critical localization modulus takes the following form:

$$H_p^{loc} = 4G\left[[1-2\nu]\left[\sigma_c - \frac{1}{3}I_1 + \frac{1+\nu}{2[1-2\nu]}[\alpha_F + \alpha_Q]\right]^2 + R^2 - J_2 - \frac{[1+\nu]^2}{8[1-2\nu][1-\nu]}[\alpha_F + \alpha_Q]^2 - \frac{[1+\nu]}{[1-\nu]}\alpha_F\alpha_Q\right] \tag{169}$$

which may be positive only in the case of nonassociative flow when $\alpha_F \neq \alpha_Q$.

E. Model Problem of Simple Shear

Under strain control the direct shear experiment exhibits different levels of dilatancy depending on the dyadic product in the plastic portion of the elastoplastic operator assuming that the elastic behavior is isotropic. The confinement of full strain control introduces failure modes that vary from a ductile shear failure mode under high confinement to brittle failure under zero confinement. An intriguing consequence of the Reynolds effect is the apparent hardening of the perfectly plastic material model which arises because of the constrained dilatancy.

In the numerical simulation of simple shear (see Fig. 18) the in-plane shear strain γ_{12} increases monotonically. The critical directions of localized failure are studied by means of the nonassociated parabolic Drucker–Prager plasticity formulation when $\alpha_Q = 0$, which corresponds to incompressible plastic flow. Therefore, the influence of lateral confinement is increased gradually by increasing the contrast ratio of compressive to tensile strength, $f_c' : f_t'$. This ratio directly affects the value of the friction coefficient α_F, whereas the parameter α_Q is kept zero according to a plastic potential of the von Mises type.

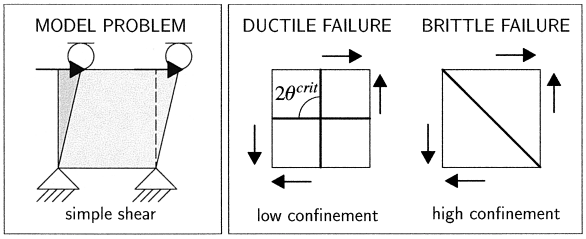


FIGURE 18 Model problem, simple shear.

TABLE I Critical Failure Angle of Parabolic Drucker–Prager Model, Simple Shear Problem

$f_c' : f_t'$	$\nu = 0.000$ (°)	$\nu = 0.125$ (°)	$\nu = 0.250$ (°)	$\nu = 0.375$ (°)	$\nu = 0.499$ (°)
1:1	45.00	45.00	45.00	45.00	45.00
3:1	35.26	33.99	32.69	31.36	30.01
5:1	29.45	27.24	24.90	22.38	19.64
8:1	22.20	18.26	13.37	5.39	0.00
12:1	11.78	0.00	0.00	0.00	0.00

Table I summarizes the critical failure angle θ^{crit} for different Poisson ratios and strength ratios $f_c' : f_t'$. The critical failure angles range from 0° to 45°, thus including failure modes which range from pure mode I decohesive failure to mode II slip failure. As expected, the critical failure angle decreases with an increase of lateral confinement which is caused by increasing the contrast of the compressive to tensile strength. Moreover, the influence of lateral confinement increases for larger values of Poisson's ratio. Figures 19 to 21 show the result of the corresponding geometric localization analysis for a Poisson's ratio of $\nu = 0.2$. The first figure is for a strength ratio of $f_c' : f_t' = 1{:}1$. This choice corresponds to a vanishing friction coefficient, $\alpha_F = 0.0$, representing the classical yield function of the von Mises type. For this analysis, two critical directions are found under $\theta^{crit} = 45°$ and $\theta^{crit} = 135°$ indicating mode II shear failure. This mode is typically

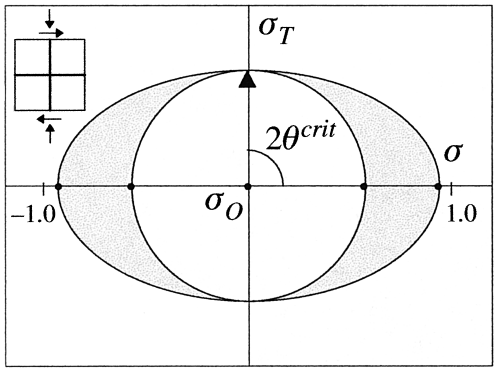


FIGURE 19 Geometric localization analysis, $f_c' : f_t' = 1{:}1$. [From Kuhl, E. *et al.* (2000). *Int. J. Solids Structures* **37,** 48–50. With permission.]

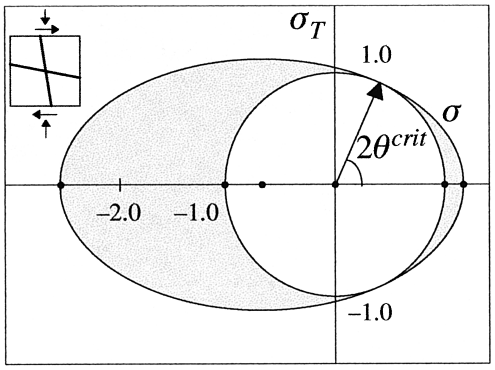


FIGURE 20 Geometric localization analysis, $f_c':f_t' = 3{:}1$. [From Kuhl, E. *et al.* (2000). *Int. J. Solids Structures* **37,** 48–50. With permission.]

observed in pressure-insensitive metals exhibiting Lüders bands.

Figure 20 depicts the result of localization analysis for the strength ratio $f_c':f_t' = 3{:}1$ representative for cast iron. The corresponding friction coefficient of $\alpha_F = 0.667$ induces relatively low confinement. Again, two critical directions are found, which have rotated slightly towards the direction of maximum principal strain at 45°. The corresponding critical failure angles of $\theta = 33.211°$ and $\theta = 146.79°$ indicate mixed shear-compression failure.

Finally, the compressive strength is assumed to be 12 times the tensile strength, $f_c':f_t' = 12{:}1$, which introduces high confinement. The friction coefficient takes a very large value of $\alpha_F = 3.667$, which is representative of cementitious materials such as concrete. For this class of materials with a very large contrast between compressive and tensile strength values, a brittle mode of failure is observed. Figure 21 illustrates mode I decohesive failure for which the critical direction corresponds to the direction of maximum principal stress, where $\theta^{crit} = 0°$.

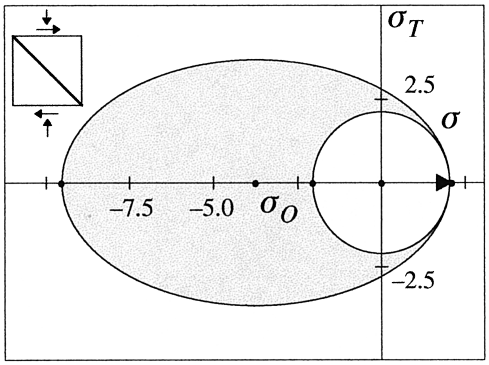


FIGURE 21 Geometric localization analysis, $f_c':f_t' = 12{:}1$. [From Kuhl, E. *et al.* (2000). *Int. J. Solids Structures* **37,** 48–50. With permission.]

V. ELASTIC DAMAGE MODELS

Elastic degradation formulations have received increasing attention since their inception by Lazar Markovich Kachanov (1914–1993). Nowadays, damage models are used regularly to describe the reduction of the elastic stiffness properties in quasibrittle materials, such as concrete, rocks, ceramics, etc. (see Lemaitre, 1992). For improved realism, additional features have been incorporated beyond the basic format of a scalar damage factor in order to account for different stiffness properties in tension and compression, to capture stiffness recovery due to microcrack closure, and to model the interaction of damage and plasticity processes.

As indicated in Section II.B.1, the basic secant formulation encompasses the simplest format of isotropic damage, which has emerged in the literature. Most scalar damage models start from the basic notion of the effective area concept by Kachanov,

$$A_{eff} = A - A_d = [1-d]A, \quad \text{where} \quad d = \frac{A_d}{A} \quad \text{and} \quad 0 \le d \le 1 \tag{170}$$

according to which the internal stresses are transferred by the intact material skeleton. In other words, distributed microdefects and stress concentrators reduce the effective load-bearing area as compared to the nominal area. This argument leads directly to the effective stress concept through equilibrium considerations, whereby:

$$\sigma A = \sigma_{eff} A_{eff} \quad \text{leads to} \quad \sigma = [1-d]\sigma_{eff} \tag{171}$$

The strain equivalence concept $\epsilon_{eff} = \epsilon$ of a parallel system of damaging elements leads to the classical scalar format of elastic damage,

$$\sigma_{eff} = E\epsilon_{eff} \quad \text{such that} \quad \sigma = [1-d]E\epsilon \tag{172}$$

The $[1-d]$ factor reduces the elastic stiffness properties when $d = 0 \to 1$. In summary, the elementary damage model leads to the secant stress–strain relationship,

$$\sigma = E_s\epsilon, \quad \text{where} \quad E_s = [1-d]E \quad \text{and} \quad d = 1 - \frac{E_s}{E} \tag{173}$$

For a more differentiated isotropic approach, the degradation distinguishes between volumetric and deviatoric damage. In the case of anisotropic degradation, a number of representations have been proposed, starting from vector to second-, fourth-, and eighth-order tensor formulations. Thus, the most popular concept is the second-order damage tensor $[d_{ij}] = \boldsymbol{d}$ and its equivalent integrity tensor,

$$\phi = \mathbf{1} - \mathbf{d} \quad \text{or} \quad \phi_{ij} = \delta_{ij} - d_{ij} \tag{174}$$

Whereas the secant formulation of elastic stiffness degradation is well established for isotropic as well as anisotropic damage, the constitutive formulation of the damage process requires a loading function, a damage rule, and some hardening/softening laws that describe the degradation process in terms of a reduced number of internal variables. In the spirit of the internal variable theory, they need to be defined in the space of conjugate thermodynamic forces, which turns out to be the strain energy density function when a scalar damage variable is used.

A. Basic Format of Elastic Scalar Damage

In the traditional $[1-d]$ model of scalar damage, the isotropic stiffness and compliance tensors are replaced by their secant values,

$$\mathcal{E}_s = [1-d]\mathcal{E}_o \quad \text{and} \quad \mathcal{C}_s = \frac{1}{1-d}\mathcal{C}_o \tag{175}$$

Differentiation of the secant relations leads to:

$$\dot{\mathcal{E}}_s = -\dot{d}\mathcal{E}_o \quad \text{and} \quad \dot{\mathcal{C}}_s = \frac{\dot{d}}{1-d}\mathcal{C}_s \tag{176}$$

The compliance term suggests changing variables and using the logarithmic scalar damage variable,

$$\dot{\ell} = \frac{\dot{d}}{1-d}, \quad \text{where} \quad \ell = \int \dot{\ell} = ln\left(\frac{1}{1-d}\right) \quad \text{and} \quad d = 1 - e^{-\ell} \tag{177}$$

This permits us to rewrite the secant tensors and the derivative of the compliance in the form of:

$$\mathcal{E}_s = e^{-\ell}\mathcal{E}_o \quad \text{and} \quad \mathcal{C}_s = e^{\ell}\mathcal{C}_o \quad \text{such that} \quad \dot{\mathcal{C}}_s = \dot{\ell}\mathcal{C}_s \tag{178}$$

In analogy to plastic associativity we need to examine the thermodynamic force conjugate to the damage variable in the dissipation inequality. The recoverable part of the elastic energy density is a function of the current secant stiffness or compliance properties; that is,

$$2W = \boldsymbol{\epsilon} : \mathcal{E}_s : \boldsymbol{\epsilon} = \boldsymbol{\sigma} : \mathcal{C}_s : \boldsymbol{\sigma} \tag{179}$$

Differentiation leads to the following energy exchange:

$$\dot{W} = \boldsymbol{\sigma} : \dot{\epsilon} - \frac{1}{2}\boldsymbol{\sigma} : \dot{\mathcal{C}}_s : \boldsymbol{\sigma} \tag{180}$$

where $\boldsymbol{\sigma} : \dot{\epsilon}$ is the external energy supply and $\dot{W}$ the increase of elastic strain energy density. Thus, the difference defines the dissipation rate, which must remain positive according to the second thermodynamic law:

$$\dot{\mathcal{D}} = \boldsymbol{\sigma} : \dot{\epsilon} - \dot{W} = \frac{1}{2}\boldsymbol{\sigma} : \dot{\mathcal{C}}_s : \boldsymbol{\sigma} \geq 0 \tag{181}$$

The thermodynamic force $\boldsymbol{\mathcal{Y}}$, which is taken to be conjugate to the basic change of the secant compliance, is

$$\dot{\mathcal{D}} = -\boldsymbol{\mathcal{Y}} :: \dot{\mathcal{C}}_s \quad \text{where} \quad -\boldsymbol{\mathcal{Y}} = \frac{1}{2}\boldsymbol{\sigma} \otimes \boldsymbol{\sigma} \tag{182}$$

In the scalar format of logarithmic damage, the thermodynamic conjugate force reduces to the elastic strain energy density function W, as:

$$\dot{\mathcal{D}} = -\boldsymbol{\mathcal{Y}} :: \mathcal{C}_s\dot{\ell} = \frac{1}{2}\boldsymbol{\sigma} : \mathcal{C}_s : \boldsymbol{\sigma}\dot{\ell} = W\dot{\ell} \tag{183}$$

For associativity, the damage function needs to be expressed in terms of the conjugate thermodynamic force, which delimits the undamaged response regime of the strain energy resistance function r_d in analogy to the yield condition of plasticity,

$$F_d = F(W, \ell) = W - r_d(\ell) = 0 \tag{184}$$

In other terms, the resistance function $r_d = W$ is the internal strain energy of a specific test environment following the classical argument of Beltrami, or its deviatoric component in the case of the distortional energy criterion of Huber.

The derivation follows the elastoplastic format developed by Carol *et al.* (2001) for elastic dagradation. Starting from the decomposition of strain rate into an elastic stress producing and a degrading strain component, $\dot{\boldsymbol{\epsilon}} = \dot{\boldsymbol{\epsilon}}_e + \dot{\boldsymbol{\epsilon}}_d$, illustrated in Fig. 22, the current secant stiffness is being used to differentiate between the elastic stress-producing deformations and the inelastic damage stress. This results in zero permanent deformations under load–unload cycles, though dissipation will take place according to the degree of nonlinearity during loading as opposed to linear unloading.

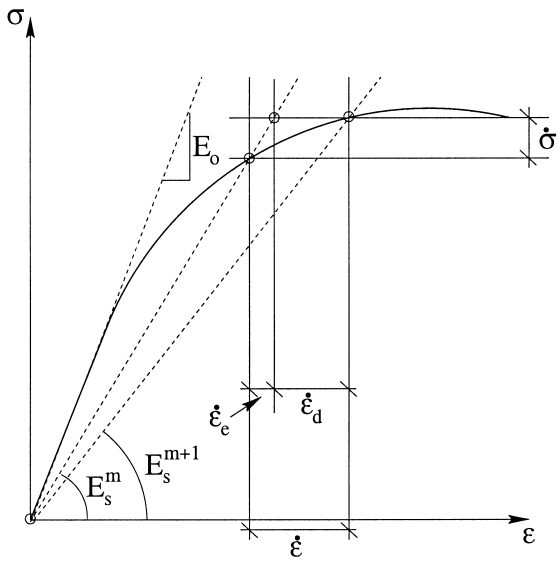


FIGURE 22 Response behavior of elastic degradation.

The set of rate equations which describes progressive degradation follows the individual steps of the flow theory of plasticity:

1. *Stress–strain rate relation*, $\boldsymbol{\sigma} = \boldsymbol{\mathcal{E}}_s : \boldsymbol{\epsilon}$ *and* $\dot{\boldsymbol{\sigma}} = \boldsymbol{\mathcal{E}}_s : [\dot{\boldsymbol{\epsilon}} - \dot{\boldsymbol{\epsilon}}_d]$, where the rate of damage strain defines the deviation from linearity with regard to the linear secant stiffness as indicated in Fig. 22.
2. *Damage rule*, $\dot{\boldsymbol{\epsilon}}_d = \dot{\lambda}\boldsymbol{m}_d$, where $\boldsymbol{m}_d = \boldsymbol{\mathcal{C}}_s : \boldsymbol{\sigma} = \boldsymbol{\epsilon}$ defines the directional properties of elastic degradation and where the damage multiplier $\dot{\lambda}$ defines the magnitude of the inelastic damage strain.
3. *Damage threshold condition*, $F_d = f(\boldsymbol{\mathcal{Y}}) - r_d(\lambda) = 0$, where $f(\boldsymbol{\mathcal{Y}})$ denotes the demand in terms of a scalar-valued representation of the conjugate force in the dissipation inequality, and where r_d defines the threshold resistance function.
4. *Consistency condition for persistent elastic damage*, $\dot{F}_d = \boldsymbol{n}_d \dot{\sigma} - H_d \dot{\lambda} = 0$, where $\boldsymbol{n}_d = \frac{\partial F_d}{\partial \boldsymbol{\sigma}}$ is the gradient of the damage function and where $H_d = -\frac{\partial F_d}{\partial \lambda}$ denotes the hardening/softening modulus of the resistance function.
5. *Damage multiplier*, $\dot{\lambda} = \frac{1}{h_d}\boldsymbol{n} : \boldsymbol{\mathcal{E}}_s : \dot{\boldsymbol{\epsilon}}$, where the strain-driven format of damage introduces the denominator $h_d = H_d + \boldsymbol{n}_d : \boldsymbol{\mathcal{E}}_s : \boldsymbol{m}_d$ which is subject to the same constraint, $h_d > 0$, as the analogous denominator term in elastoplasticity. This leads to the critical softening modulus for elastic damage $H_d^{crit} = -\boldsymbol{n}_d : \boldsymbol{\mathcal{E}}_s : \boldsymbol{m}_d$.
6. *Tangential scalar damage properties*, $\boldsymbol{\mathcal{E}}_{ed} = \boldsymbol{\mathcal{E}}_s - \frac{1}{h_d}\boldsymbol{\mathcal{E}}_s : \boldsymbol{m}_d \otimes \boldsymbol{n}_d : \boldsymbol{\mathcal{E}}_s$, where the main difference from elastoplasticity is the secant reference stiffness. The constitutive format has the same structure as elastoplasticity except that $\boldsymbol{\mathcal{E}}_s \neq const$.

In the isotropic case of the logarithmic scalar damage model, $\dot{\lambda} = \dot{\ell}$ and both the damage rule as well as the gradient of the damage function are associative as long as the damage function is expressed in terms of the conjugate force,

$$F_d = W - r_d(\ell) = 0 \quad \text{with} \quad \boldsymbol{n}_d = \frac{\partial F_d}{\partial \boldsymbol{\sigma}} = \boldsymbol{\mathcal{C}}_s : \boldsymbol{\sigma} = \boldsymbol{\epsilon} \tag{185}$$

The hardening/softening modulus H_d at constant stress is: defined as:

$$H_d = -\frac{\partial F_d}{\partial \lambda} = \frac{\partial r_d}{\partial \ell} \tag{186}$$

From these terms, the tangent stiffness of elastic degradation may be assembled in the standard manner which yields:

$$\boldsymbol{\mathcal{E}}_{ed} = e^{-\ell}\boldsymbol{\mathcal{E}}_s - \frac{1}{h_d}\boldsymbol{\sigma} \otimes \boldsymbol{\sigma} \tag{187}$$

where the denominator $h_d = \frac{\partial r_d}{\partial \ell} + \boldsymbol{\sigma} : \boldsymbol{\epsilon}$ must remain strictly positive under strain control. The symmetric format of the tangent stiffness operator is a consequence of the associative damage function, in contrast to damage functions such as the strain-based damage model of Mazars and his coworkers.

B. Simple Shear Response of Logarithmic Scalar Damage Model

In what follows we compare the results of the scalar-valued damage formulation with the corresponding elastoplastic results in Section III.D in order to assess their volumetric–deviatoric interaction. Analogous to plasticity, the elastic damage matrix may be partitioned into four submatrices which correspond to the normal and shear components of stress and strain. In the case of isotropic elasticity, there is no coupling between elastic normal and shear stresses and normal and shear strains in the initial stage before damage takes place.

Strain control of the simple shear test results in $\tau_{12} = G_0\gamma_{12}$ at the onset of damage. Substituting into the elastic damage operator leads to the rate equations:

$$\begin{Bmatrix} \dot{\sigma}_{11} \\ \dot{\sigma}_{22} \\ \dot{\sigma}_{33} \\ \cdots \\ \dot{\tau}_{12} \\ \dot{\tau}_{23} \\ \dot{\tau}_{13} \end{Bmatrix} = e^{-\ell} \left[\begin{array}{ccc:ccc} K_o + \frac{4}{3}G_o & K_o - \frac{2}{3}G_o & K_o - \frac{2}{3}G_o & 0 & 0 & 0 \\ K_o - \frac{2}{3}G_o & K_o + \frac{4}{3}G_o & K_o - \frac{2}{3}G_o & 0 & 0 & 0 \\ K_o - \frac{2}{3}G_o & K_o - \frac{2}{3}G_o & K_o + \frac{4}{3}G_o & 0 & 0 & 0 \\ \hdashline 0 & 0 & 0 & G_o - \frac{e^{\ell}}{h_d}\tau_{12}^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & G_o & 0 \\ 0 & 0 & 0 & 0 & 0 & G_o \end{array}\right] \begin{Bmatrix} 0 \\ 0 \\ 0 \\ \cdots \\ \dot{\gamma}_{12} \\ 0 \\ 0 \end{Bmatrix} \tag{188}$$

The coupling partitions are zero, showing no interaction between the normal stresses and the shear strains, when the material experiences damage in analogy to the relation in Eq. (106) for J_2-plasticity.

1. Singularity of Tangent Damage Operator

Possible singularities of $\boldsymbol{E}_{ed}$ may be studied again with the Schur theorem of zero subdeterminant. We note that the normal partition of the stiffness matrix $\boldsymbol{E}^{ed}_{nn}$ simply reflects the degradation of the intact elastic partition $\boldsymbol{E}^{0}_{nn}$ by the integrity factor $e^{-\ell} = 1 - d$ as $\ell \to \infty$. The second possibility, which could cause the tangent operator to become singular, is the shear partition, $\boldsymbol{E}^{ed}_{ss}$. The two diagonal terms E^{ed}_{55} and E^{ed}_{66} are affected by degradation of the elastic properties; they will induce a singularity only when $e^{-\ell} \to 0$. The diagonal term E^{ed}_{44}, however, can diminish to zero before the material is fully damaged, since:

$$E^{ed}_{44} = e^{-\ell} G_o - \frac{1}{h_d}\tau^2_{12} = 0 \tag{189}$$

when

$$h_d = \frac{\tau^2_{12}}{e^{-\ell} G_o} = \frac{\tau^2_{12}}{G_s} = \tau_{12}\gamma_{12} \tag{190}$$

The damage denominator $h_d = H_d + \boldsymbol{\sigma} : \boldsymbol{\epsilon}$ infers that the hardening modulus must be zero; that is,

$$H_d = \frac{\partial r}{\partial \ell} = 0 \quad \text{when} \quad h_d = \tau_{12}\gamma_{12} \tag{191}$$

for any value of damage, ℓ. Thus, for perfectly damaging behavior, when $H_d = 0$ we have $E^{ed}_{44} = 0$, and consequently the tangent operator turns singular, $\det \boldsymbol{E}_{ed} = 0$. This behavior is very much analogous to pressure-insensitive von Mises plasticity and holds independently of the load condition, when elastic damage is restricted to the deviatoric response behavior.

C. Damage Models under Simple Shear

To illustrate the performance of damage models let us consider again the strain-controlled load case of simple shear. To start with, we examine the relationship between scalar damage mechanics and the earlier development of one-invariant von Mises plasticity. Comparing the governing tangent operators in Eqs. (88) and (188) we recognize the same structure except for the integrity factor $e^{-\ell}$ in the case of scalar damage. Using the equivalent energy density value to calibrate the damage threshold $r_d = \frac{1}{2G}\tau^2_y$, Fig. 23a shows that the response behavior is identical to that of elastoplasticity shown for monotonic loading in Fig. 9. The basic difference between damage and plasticity shows up when we consider cyclic loading which takes place along the degraded secant stiffness, and which does not distinguish between tension and compression because of the underlying Beltrami condition of a maximum strain energy threshold. There is no confinement effect and the shear strength in this strain-controlled problem coincides with the tensile strength. Since the simple shear response may be viewed as an equi-biaxial tension-compression test with no direct-shear coupling, all damage stresses degrade exponentially according to the logarithmic damage evolution, in tension as well as in compression and shear.

In contrast, we consider the recent anisotropic damage formulation of Carol *et al.* (2001) to examine the influence of anisotropy on the volumetric–deviatoric interaction.

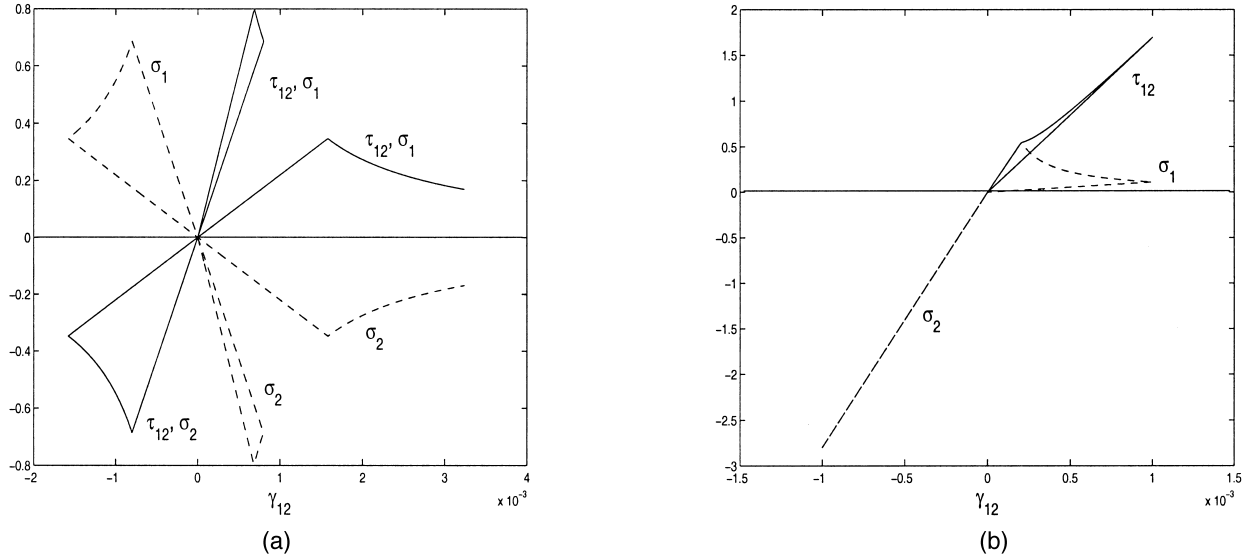


FIGURE 23 Simple shear response: (a) scalar damage model, (b) anisotropic damage model. [From Carol, I. *et al.* (2001). *Int. J. Solids Structures* **38,** 491–546. With permission.]

In this case, a second-order integrity tensor $\phi = \mathbf{1} - \boldsymbol{d}$ is adopted in extension of the pseudologarithmic scalar format above. Resorting to the so-called energy equivalence of Cordebois and Sidoroff, where neither effective stress nor effective strain coincides with its nominal counterpart, we develop a fourth-order damage effect tensor,

$$\boldsymbol{\sigma}_{eff} = \boldsymbol{\alpha} : \boldsymbol{\sigma} \quad \text{and} \quad \boldsymbol{\epsilon} = \boldsymbol{\alpha}^t : \boldsymbol{\epsilon}_{eff} \tag{192}$$

Assuming linear elastic behavior at the level of effective stress and effective strain, the resulting secant compliance retains symmetry:

$$\boldsymbol{\mathcal{C}}_s = \boldsymbol{\alpha}^t : \boldsymbol{\mathcal{C}} : \boldsymbol{\alpha} \tag{193}$$

In the principal axes of damage, the damage effect tensor reduces to diagonal form:

$$\boldsymbol{\alpha} = \left[\begin{array}{ccc|ccc} \phi_1 & 0 & 0 & & & \\ 0 & \phi_2 & 0 & & 0 & \\ 0 & 0 & \phi_3 & & & \\ \hline & & & \sqrt{\phi_1\phi_2} & & \\ & 0 & & & \sqrt{\phi_2\phi_3} & \\ & & & & & \sqrt{\phi_3\phi_1} \end{array}\right] \tag{194}$$

Consequently, the corresponding secant compliance of anisotropic damage has the format

$$\boldsymbol{C}_s = \frac{1}{E}\left[\begin{array}{ccc|ccc} \phi_1^2 & -\nu\phi_1\phi_2 & -\nu\phi_1\phi_3 & & & \\ -\nu\phi_2\phi_1 & \phi_2^2 & -\nu\phi_2\phi_3 & & 0 & \\ -\nu\phi_3\phi_1 & -\nu\phi_3\phi_2 & \phi_3^2 & & & \\ \hline & & & 2[1+\nu]\phi_1\phi_2 & & \\ & 0 & & & 2[1+\nu]\phi_2\phi_3 & \\ & & & & & 2[1+\nu]\phi_3\phi_1 \end{array}\right] \tag{195}$$

where the secant stiffness involves five independent material parameters as opposed to nine in orthotropic elasticity. For damage evolution we need to define a damage criterion. For associated damage the demand function $F_d = f(\boldsymbol{\mathcal{Y}}) - r_d(\ell) = 0$ is based on the conjugate force which, in this case, is the second-order energy tensor $\boldsymbol{\mathcal{Y}} = \boldsymbol{\sigma} \cdot \boldsymbol{\epsilon}$. In this case, we resort to the Rankine criterion of maximum principal energy where the scalar damage function is $F_d = \mathcal{Y}_{max} - r_d(\ell) = 0$. Adopting a no-damage switch in compression, the simple shear response results in the response behavior shown in Fig. 23b. In comparison to the scalar damage format in Fig. 23a, the simple shear response of the anisotropic damage model in Fig. 23b shows some very interesting features: (1) after reaching the damage threshold at $\tau_{12} = 0.8\ ksi$, the principal tensile stress diminishes again exponentially; and (2) the shear stress and the compressive principal stress increase without bounds because of volumetric–deviatoric interaction which introduces increasing confinement. Moreover, there is no damage in compression, and unloading occurs at the secant stiffness in tension, leaving no permanent deformation, while the compressive stiffness remains intact.

In contrast to the scalar damage model, we notice that the anisotropic damage model does exhibit volumetric–deviatoric coupling. Thus, the shear stress increases far beyond the initial shear strength threshold $\tau_y = 0.8\ ksi$ in the presence of out-of-plane confinement. At the same time, the nominal tensile stress is subject to softening due to progressive tensile damage according to the logarithmic degradation law.

In summary, isotropic scalar damage models do not exhibit shear dilatancy, while anisotropic models do reproduce the Reynolds dilatancy effect of frictional materials similarly to pressure-sensitive plasticity.

VI. CONCLUSIONS

This state-of-the-art report has focused on major developments in rate-independent elasticity and inelasticity. Many topics have not been covered for the sake of brevity—most notably, nonlinear hardening and softening formats in plasticity and damage (Lubliner, 1990); the general form of the internal variable description (Halphen and Nguyen, 1975); the microstructural features thereof (Kröner, 1963); the size and gradient effects (Bažant and Planas, 1997); and last but not least the computational aspects of inelastic analysis.

The exposition presented an apercu of linear and nonlinear elastic constitutive models as well as plastic and elastic damage models. Specifically, we have examined the interaction of normal and shear components which is directly related to the population in the constitutive gradients $\boldsymbol{n}$, $\boldsymbol{m}$ responsible for stress- and/or strain-induced anisotropy in the tangent operator.

ACKNOWLEDGMENTS

The author wishes to acknowledge the help of Dr. Eric Hansen and former coworkers Dr. Ellen Kuhl and Dr. Hong Kang for their assistance and for preparing some of the figures. This report was partially sponsored by the National Science Foundation grant CMS 9622940 to the University of Colorado, Boulder. Opinions expressed in this paper are those of the writer and do not reflect those of the sponsor.

APPENDIX 1: SOME RESULTS OF MATRIX AND TENSOR ANALYSIS

The mathematical tools behind material models are housed in linear algebra and tensor analysis, in particular. For this reason, we revisit a few well-established relationships which may provide additional insight beyond the mere mechanics of elementary tensor manipulation.

Stress and Strain Tensors

Stress and strain tensors are a set of objects $\boldsymbol{\sigma}, \boldsymbol{\epsilon} \in \Re^3$ which may be represented as (3×3) matrices whose coefficients are the coordinates of a set of orthonormal base vectors.

1. Scalar (inner) product of $\boldsymbol{\sigma}, \boldsymbol{\epsilon} \in \Re^3$ (double contraction):

This product operation generates a scalar value:

$$w = \boldsymbol{\sigma} : \boldsymbol{\epsilon} = \sigma_{ij}\epsilon_{ij} \tag{196}$$

The direction between two tensors is

$$\cos\theta = \frac{\boldsymbol{\sigma} : \boldsymbol{\epsilon}}{\sigma\epsilon} \tag{197}$$

where $0 \le \theta \le \pi/2$, and $\sigma = (\boldsymbol{\sigma} : \boldsymbol{\sigma})^{\frac{1}{2}}$ and $\epsilon = (\boldsymbol{\epsilon} : \boldsymbol{\epsilon})^{\frac{1}{2}}$ are the Euclidean lengths of the two vectors. *Note*: The internal strain energy is directly related to the scalar product of stress and strain.

2. Outer product of $\boldsymbol{\sigma}, \boldsymbol{\epsilon} \in \Re^3$ (single contraction): This product operation generates a second order tensor:

$$\boldsymbol{w} = \boldsymbol{\sigma} \cdot \boldsymbol{\epsilon} = \sigma_{ij}\epsilon_{jk} \tag{198}$$

Note: The thermodynamic force in the form of a tensor product of stress and strain is conjugate to the second order damage tensor.

3. Dyadic or tensor product of $\boldsymbol{\sigma}, \boldsymbol{\epsilon} \in \Re^3$ (no contraction): This product operation generates a fourth-order rank-one tensor:

$$\boldsymbol{W} = \boldsymbol{\sigma} \otimes \boldsymbol{\epsilon} = \sigma_{ij}\epsilon_{kl} \tag{199}$$

Note: The tensorial order of the dyadic product is the sum of the tensorial order of each factor; e.g., the dyadic product of two second-order tensors generates a fourth-order tensor, etc.

4. Inverse of rank-one modification of unit tensor $\boldsymbol{I}$:

$$\boldsymbol{B} = \boldsymbol{I} + \boldsymbol{a} \otimes \boldsymbol{b} \quad \text{then} \quad \boldsymbol{B}^{-1} = \boldsymbol{I} - \frac{1}{1 + \boldsymbol{a} \cdot \boldsymbol{b}} \boldsymbol{a} \otimes \boldsymbol{b} \tag{200}$$

as long as $\boldsymbol{a} \cdot \boldsymbol{b} \neq -1$.

5. Sherman–Morrison formula of inversion:

$$\boldsymbol{B} = \boldsymbol{A} + \boldsymbol{X} \otimes \boldsymbol{Y} \tag{201}$$

then

$$\boldsymbol{B}^{-1} = \boldsymbol{A}^{-1} - \frac{1}{1 + \boldsymbol{Y}\boldsymbol{A}^{-1}\boldsymbol{X}} \boldsymbol{A}^{-1}\boldsymbol{X} \otimes \boldsymbol{Y}\boldsymbol{A}^{-1} \tag{202}$$

6. Outer products do not commute:

$$\boldsymbol{A} \cdot \boldsymbol{B} \neq \boldsymbol{B} \cdot \boldsymbol{A}; \quad \text{however,} \quad \boldsymbol{B} \cdot \boldsymbol{A} = [\boldsymbol{A}^t \cdot \boldsymbol{B}^t]^t \tag{203}$$

This also holds for tensor products.

Determinant

The determinant of a second-order tensor is a single number which summarizes the tensorial property in the form of a multilinear functional.

1. Determinant of tensor products:

$$\det(\boldsymbol{A} \cdot \boldsymbol{B}) = \det \boldsymbol{A} \det \boldsymbol{B} \tag{204}$$

2. The Schur theorem states that the determinant of partitions in tensor $\boldsymbol{A}$ may be written in terms of the determinant product of partitions:

$$\begin{aligned} \det \boldsymbol{A} &= \det \boldsymbol{A}_{22} \det \boldsymbol{A}_{11}^{compl} \\ &= \det \boldsymbol{A}_{22} \det \left(\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}\right) \end{aligned} \tag{205}$$

Note: The Schur complement couples the partitions and leads to the bound $\det \boldsymbol{A} \le \det \boldsymbol{A}_{22} \det \boldsymbol{A}_{11}$ as long as $\boldsymbol{x} \cdot \boldsymbol{A}_{22} \cdot \boldsymbol{x} > 0$ is positive definite.

Eigenvalues and Eigenvectors

There exists a nonzero vector $\boldsymbol{x}$ such that the linear transformation $\boldsymbol{\sigma} \cdot \boldsymbol{x}$ is a multiple of $\boldsymbol{x}$:

$$\boldsymbol{\sigma} \cdot \boldsymbol{x} = \lambda \boldsymbol{x} \tag{206}$$

Note: The eigenvector $\boldsymbol{x}_i$ spans the triad of principal directions, and the eigenvalues λ_i define the three principal values of stress.

1. Characteristic polynomial: The eigenvalue problem is equivalent to stating that:

$$(\boldsymbol{\sigma} - \lambda \boldsymbol{I})\boldsymbol{x} = \boldsymbol{0} \tag{207}$$

For a nontrivial solution $\boldsymbol{x} \neq \boldsymbol{0}$, then $(\boldsymbol{\sigma} - \lambda_i \boldsymbol{I})$ must be singular. Consequently,

$$\det(\boldsymbol{\sigma} - \lambda \boldsymbol{I}) = 0 \tag{208}$$

generates the characteristic polynomial

$$p(\lambda) = \det(\boldsymbol{\sigma} - \lambda \boldsymbol{I}) \tag{209}$$

the roots of which are the eigenvalues $\lambda(\boldsymbol{\sigma})$. According to the fundamental theorem of algebra, a polynomial of degree 3 has exactly 3 roots, thus each matrix $\boldsymbol{\sigma} \in \Re^3$ has 3 eigenvalues. *Note*: All three eigenvalues are real as long as $\boldsymbol{\sigma} = \boldsymbol{\sigma}^t$ is symmetric, which is the case for nonpolar materials because of conjugate shear stresses $\sigma_{ij} = \sigma_{ji}$.

2. Cayley–Hamilton theorem: This theorem states that every square matrix satisfies its own characteristic equation. In other words, the scalar polynomial $p(\lambda) = \det(\lambda \boldsymbol{I} - \boldsymbol{\sigma})$ also holds for the stress polynomial $p(\boldsymbol{\sigma})$. One important application of the Cayley–Hamilton theorem is to express powers of the stress tensor $\boldsymbol{\sigma}^k$ as a linear combination of the irreducible bases $\boldsymbol{I}, \boldsymbol{\sigma}, \boldsymbol{\sigma}^2$ for $k \geq 2$.

3. Spectral properties of rank-one update of unit tensor: Spectral analysis of a square matrix generated by a rank-one update of the unit tensor of second-order

$$\boldsymbol{B} = \boldsymbol{I} + \boldsymbol{a} \otimes \boldsymbol{b} \tag{210}$$

reduces to the eigenvalue analysis,

$$[(\boldsymbol{b} \cdot \boldsymbol{a} - \lambda \boldsymbol{I})\alpha]\boldsymbol{a} = \boldsymbol{0} \tag{211}$$

The eigenvalues and eigenvectors of $\boldsymbol{B} = \mathbf{I} + \boldsymbol{a} \otimes \boldsymbol{b}$ are related to the eigenpairs of the update matrix; that is,

$$\lambda(\boldsymbol{B}) = 1 + \lambda \quad \text{and} \quad \boldsymbol{x}(\boldsymbol{B}) = \boldsymbol{x} \tag{212}$$

In the case of a single rank-one update of the unit matrix, we find $\lambda_1(\boldsymbol{B}) = 1 + \lambda$ and $\lambda_k(\boldsymbol{B}) = 1 \forall k = 2, 3, \ldots, n$ with the determinant $\det(\boldsymbol{B}) = \det(\boldsymbol{I} + \boldsymbol{a} \otimes \boldsymbol{b}) = 1 + \boldsymbol{a} \cdot \boldsymbol{b}$ and $\lambda = \boldsymbol{a} \cdot \boldsymbol{b}$.

APPENDIX 2: PLANE STRAIN CONSTRAINT OF ELASTOPLASTIC LIMIT POINT

On a final note, we observe that the kinematic constraint of zero out-of-plane deformations delays and may suppress altogether the formation of a limit point associated with peak strength in pressure-sensitive plasticity. We recall the shear response does not reach a limit point under plane strain. The reason for this puzzling observation is the kinematic constraint that restricts the formation of a limit point when:

$$\dot{\boldsymbol{\sigma}} = \boldsymbol{\mathcal{E}} : [\dot{\boldsymbol{\epsilon}} - \dot{\boldsymbol{\epsilon}}_p] = \boldsymbol{0} \quad \text{where} \quad \dot{\boldsymbol{\epsilon}}_p = \dot{\lambda} \boldsymbol{m} \tag{213}$$

In the plastic flow rule, $\dot{\lambda}$ denotes the plastic multiplier, which is strictly positive under sustained plastic flow, and $\boldsymbol{m} = \frac{\partial Q}{\partial \boldsymbol{\sigma}}$ is the direction of plastic flow in terms of the gradient of the plastic potential. Consequently, the only possibility to form a limit point at peak occurs if the strain rate equals the plastic strain rate, $\dot{\boldsymbol{\epsilon}} = \dot{\boldsymbol{\epsilon}}_p$, and the elastic strain rate remains zero, $\dot{\boldsymbol{\epsilon}}_e = \boldsymbol{0}$. The plane strain constraint infers that the out-of-plane strain rate must vanish; that is,

$$\dot{\epsilon}_{33} = \dot{\lambda} m_{33} = \dot{\lambda} \frac{\partial Q}{\partial \sigma_{33}} = 0 \tag{214}$$

In the three-invariant plasticity formulation above, this constraint implies that:

$$m_{33} = \frac{1}{\sqrt{3}} \left(Q_\xi - \frac{Q_\theta}{\tau_{12}} \right) = 0 \quad \text{or} \quad Q_\xi = \frac{Q_\theta}{\tau_{12}} \tag{215}$$

In short, pressure-sensitive plastic flow, $Q_\xi \neq 0$, does not permit formation of a limit point as shown in Figs. 9 and 14. And, even if this would be the case, the shear stress would never reach the peak value because loss of stability and localization take place much earlier in the ascending hardening regime.

SEE ALSO THE FOLLOWING ARTICLES

BONDING AND STRUCTURE IN SOLIDS • ELASTICITY • FRACTURE AND FATIGUE • MATERIALS CHEMISTRY • PLASTICITY

BIBLIOGRAPHY

Bažant, Z. P., and Planas, J. (1997). "Fracture and Size Effect in Concrete and Other Quasibrittle Materials," CRC Press, Boca Raton, FL.

Benallal, A., and Comi, C. (1996). "Localization analysis via a geometrical method," *Int. J. Solids Structures* **33,** 99–119.

Carol, I., Rizzi, E., and Willam, K. (2001). "On the formulation of anisotropic elastic degradation. Part I: Theory based on a pseudo-logarithmic damage tensor rate; Part II: Generalized pseudo-Rankine model for tensile damage," *Int. J. Solids Structures* **38,** 491–546.

Chen, W. F., and Han, D. J. (1988). "Plasticity for Structural Engineers," Springer-Verlag, New York.

Coleman, B. D., and Gurtin, M. E. (1967). "Thermodynamics with internal state variables," *J. Chem. Phys.* **47,** 597–613.

Drucker, D. C., and Prager, W. (1952). "Soil mechanics and plastic analysis of limit design," *Q. Appl. Math.* **10,** 157–175.

Halphen, B., and Nguyen, Q. S. (1975). "Sur les matériaux standards généralisés," *J. de Mécanique* **14**(1), 39–63.

Kang, H., and Willam, K. (1999). "Localization characteristics of a triaxial concrete model," *ASCE J. Eng. Mech.* **125,** 941–950.

Kröner, E. (1963). "Dislocation: a new concept in the continuum thery of platicity," *J. Math. Phys.* **42,** 27–37.

Kuhl, E., Ramm, E., and Willam, K. (2000). "Failure analysis of elasto-plastic materials at different levels of observation," *Int. J. Solids Structures* **37,** 48–50; 7259–7280.

Lemaitre, J. (1992). "A Course on Damage Mechanics," Springer-Verlag, Berlin.

Lubliner, J. (1990). "Plasticity Theory," Macmillan, New York.

Maier, G., and Hueckel, T. (1979). "Nonassociated and coupled flow rules of elastoplasticity for rock-like materials," *Int. J. Rock Mech. Min. Sci.* **16,** 77–92.

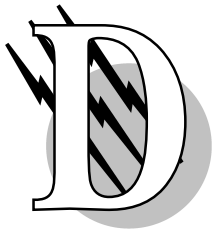
Ogden, R. W. (1984). "Non-Linear Elastic Deformations," Ellis Horwood, Chichester, U.K.

Ottosen, N. S., and Runesson, K. (1991). "Properties of discontinuous bifurcation solutions in elasto-plasticity," *Int. J. Solids Structures*, **27,** pp. 401–421.

Rudnicki, J. W., and Rice, J. R. (1975). "Conditions for the localization of deformation in pressure-sensitive dilatant materials," *J. Mech. Phys. Solids*, **23,** 371–394.

Valanis, K. C. (1975). "On the foundations of the endochronic theory of viscoplasticity," *Arch. Mech. Stos.* **27,** 857–868.

Willam, K. J., and Warnke, E. P. (1975). "Constitutive model for the triaxial behaviour of concrete" *Proc. Int. Assoc. Bridge Structural Engineers*, Report 19, Section III, Zürich, Switzerland, pp. 1–30.



Dams, Dikes, and Levees

Robert B. Jansen
Consulting Civil Engineer

I. Types of Dams
II. History of Dams
III. Appurtenant or Related Works
IV. Design of Dams
V. Construction of Dams
VI. Safety of Dams
VII. Weighing Benefits and Detriments of Dams

GLOSSARY

Dam A barrier constructed for storage of water for any of various purposes, including urban water supply, irrigation, flood control, power production, navigation, soil conservation, debris retention, and recreation. A multiple-purpose project may provide several of these benefits.

Dike A barrier, usually an embankment of earth or stone, for retaining a body of water or a stream. It may be an auxiliary embankment related to a main dam.

Levee An embankment (sometimes considered to be the same as a dike) designed to control a stream in a defined course or to prevent a body of water from intruding into areas to be protected. River walls and sea walls also serve such purposes.

DAMS HAVE BEEN keys to development in many lands, providing sustenance and protection for human survival. In the United States, systems of dams, locks, powerhouses, and boating facilities on the Ohio, Tennessee, and Mississippi Rivers are outstanding examples of projects serving multiple purposes. Of historical interest in other regions are storages or diversions for hydraulic mining, log ponds, saw mills, and canal transportation. Dams enable diversion of water for agriculture, as well as regulated discharges to natural streams for preservation of fish and wildlife. There are more than 45,000 large dams (over 15 meters high) in the world, of various types consisting of embankments and concrete and masonry structures. Public safety is a primary consideration in their design, construction, and operation.

By general definition, dams include dikes and levees. Dikes have been used for land reclamation, such developments being famous in Europe, particularly exemplified by the large-scale projects built in the Ijsselmeer (Zuider Zee) in the Netherlands. In ancient Egypt, seasonal diversions of excess flows of the Nile were made into extensive basins enclosed by dikes. In some regions, dikes used for dividing wet lowlands have been called levees. River levees are usually comparatively low earthfills, sometimes offset at a distance from the river to contain overbank flows in a floodplain. Some of the earliest river levees were those along the Nile in Egypt. Among the other important levee systems developed in major watersheds are those on the

Mississippi River, those in California on the American and Sacramento Rivers, and those that enclose islands in the San Joaquin delta.

I. TYPES OF DAMS

Dams may be classified by their materials and by their design to carry the imposed loads.

A. Embankments

Embankments may be earthfills or rockfills or combinations thereof. Their impervious elements may be interior cores of earthfill or walls composed of conventional or asphaltic concrete or of masonry (in older designs). In other cases, a waterproof membrane has been established on the upstream face of the embankment by constructing a concrete slab, which is a common design for rockfills but is seldom seen on newer earthfills. An example of a concrete-faced rockfill is the 187-meter-high (614-foot-high) Aguamilpa Dam in Mexico, completed in 1994. Other materials have been used for facings, including asphalt, timber planks, steel plates, and synthetic sheets. The local materials for the embankment itself may come from quarries and borrow areas or from excavations for structures and conduits.

Earthfills are the most common type of dam. An early earth dam was generally composed of the same material throughout, constituting a "homogeneous" section. As designs evolved, strength and seepage control were enhanced by cross sections with zones of various gradations. Most modern earthfill dams are zoned, careful attention being given to the transitions between zones. Even those embankments that are predominantly single zones are provided with drainage, such as by horizontal and inclined layers of coarse materials. A typical zoned earthfill dam has an interior core of compacted fine-grained soil of low permeability bolstered by a succession of outer pervious zones grading from sand to gravel, crushed rock, cobbles, or rockfill. An eminent earthfill is Tajikistan's Nurek Dam, which heads the list of the world's high dams.

An earthfill that is now less common is the hydraulic fill dam, which contains soils that have been conveyed to the site in a liquid mix, typically from a dredged deposit. This involved continuous ponding. At some projects on alluvial foundations, seepage underflow was controlled by dumping fine-grained soils into water-filled trenches. Semihydraulic fills were constructed of soils hauled and dumped at the site and then sluiced into place. Generally the sluicing of either type was done with the objective of depositing fine-grained earth in the center and coarse materials in the outer parts of the dam section. However, this was a relatively crude process that created marginally stable embankments. Hydraulic fills have lost favor to earthfills placed and compacted by heavy equipment, which give more reliable structural performance, particularly under seismic loading.

B. Gravity Dams

A gravity dam is a structure designed to withstand loads by its own weight and by its resistance to sliding and overturning on its foundation. Newer dams of this type are typically composed of unreinforced concrete monoliths with seals at the joints. Foremost among the world's gravity dams is the 285-meter-high (935-foot-high) Grande Dixence in Switzerland, completed in 1961. In past centuries, many gravity dams were constructed of stone masonry. New masonry dams, of both gravity and arch designs, are being built in India and China and in other lands where the cost of labor is low. Also of historic interest is the filled crib, the stability of which depends upon the weight of earth or rockfill enclosed in cells formed by a framework of timbers or concrete beams. Such dams, common a hundred years ago, were usually faced with timber planking.

C. Arch Dams

An arch dam is a concrete or masonry structure that carries load to a large extent by lateral thrust into its abutments and therefore can be thinner than a gravity dam. The modern arch is usually curved in both plan and section. Some of the early arch dams were conservatively proportioned so that much of their resistance was due to weight. A notable example of this arched gravity type is Hoover Dam. Some true arch dams of record-breaking heights are the Soviet-engineered Inguri (272 m/892 ft) and Chirkei (233 m/764 ft), Vaiont Dam in Italy (262 m/860 ft), Ertan Dam in China (240 m/787 ft), Mauvoisin Dam in Switzerland (250 m/820 ft), El Cajón Dam in Honduras (234 m/768 ft), and the Swiss Contra Dam (220 m/722 ft).

D. Buttress Dams

Buttress dams are concrete or masonry structures usually comprising inclined panels or arches supported by buttresses. In the Ambursen type, the upstream face is a sloping plane formed by flat slabs bearing directly, or on corbels, at the upstream edges of the buttresses. A special type is the massive-head dam, in which the face is formed by enlarging the upstream ends of the buttresses so that they join to eliminate the need for a spanning deck slab. Figure 1 shows the concrete section of Pueblo Dam, consisting of 23 massive-head blocks. This structure, completed in 1975, was bolstered in 1999 against potential

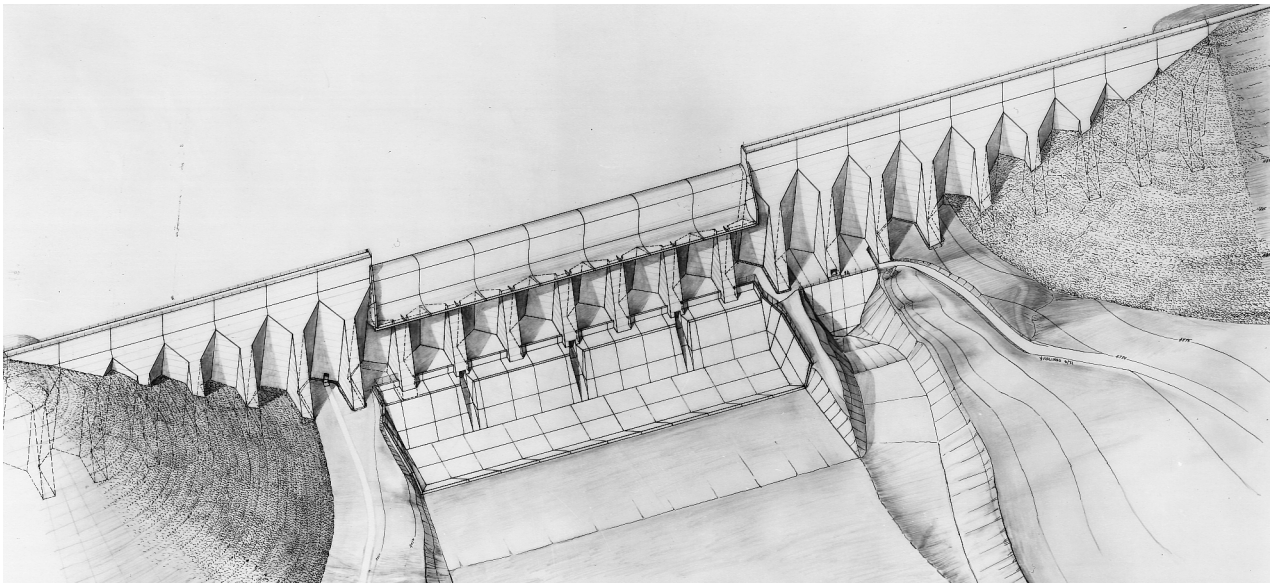


FIGURE 1 Massive-head, buttress-type concrete section of Pueblo Dam on the Arkansas River in Colorado, 533 m (1750 ft) long and 76 m (250 ft) high above the lowest foundation. The overflow spillway in the middle has a crest length of 168 m (550 ft). The dam, which was completed in 1975, extends as earth embankment at each end of the concrete section.

foundation sliding by adding a large roller-compacted concrete plug in the spillway plunge pool and a rock-bolted block at the toe of the dam. Discharges into the modified stilling basin were begun in the spring of 2000.

A multiple-arch dam has a series of arches joined at the buttresses. A unique extension of this concept is seen at Coolidge Dam in Arizona, composed of three large domes supported by massive buttresses. Engineering of multiple-arch dams was advanced notably in the first third of the 20th century at sites where they had particular economic advantages, such as at Florence Lake Dam in the Sierra Nevada of California (Fig. 2). Fewer have been built more recently; however, one of the most remarkable structures of this type, completed in 1968, is Daniel Johnson (Manicouagan No. 5) Dam in Québec, Canada (Fig. 3). It is 214 m (703 ft) high and 1314 m (4311 ft) long, comprising 13 arches, the largest spanning 161.5 m (530 ft).

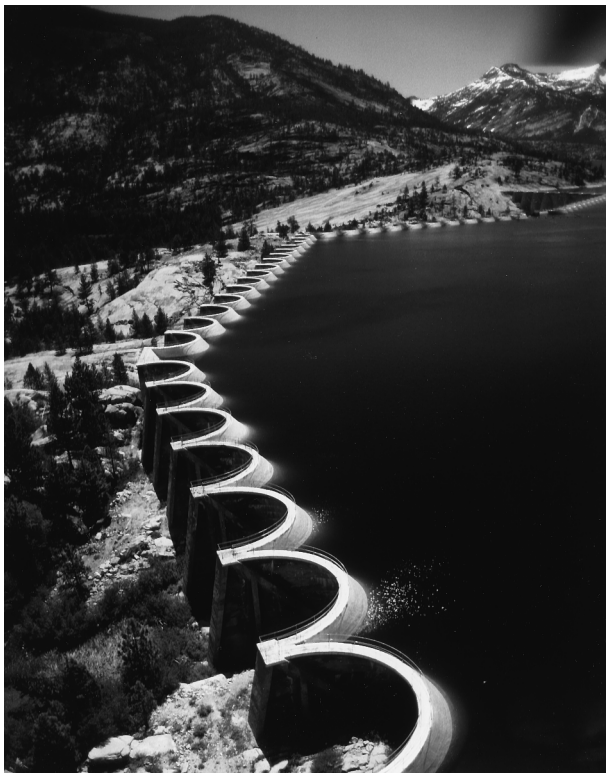


FIGURE 2 Florence Lake Dam in California, 46 m (150 ft) high and 962 m (3156 ft) long, with 58 arches. This dam was completed in 1926.

E. Roller-Compacted Concrete Dams

An innovation that has won worldwide acceptance in recent times is roller-compacted concrete (RCC, or rollcrete), which has been adopted for construction of new gravity and arch dams, as well as rehabilitation of conventional dams. A rollcrete mix typically may consist of sand, coarse aggregate such as gravel or crushed rock, portland cement plus pozzolan (commonly fly ash), and water. Only limited use has been made of admixtures for water reduction and set retardation, or for air entrainment in special

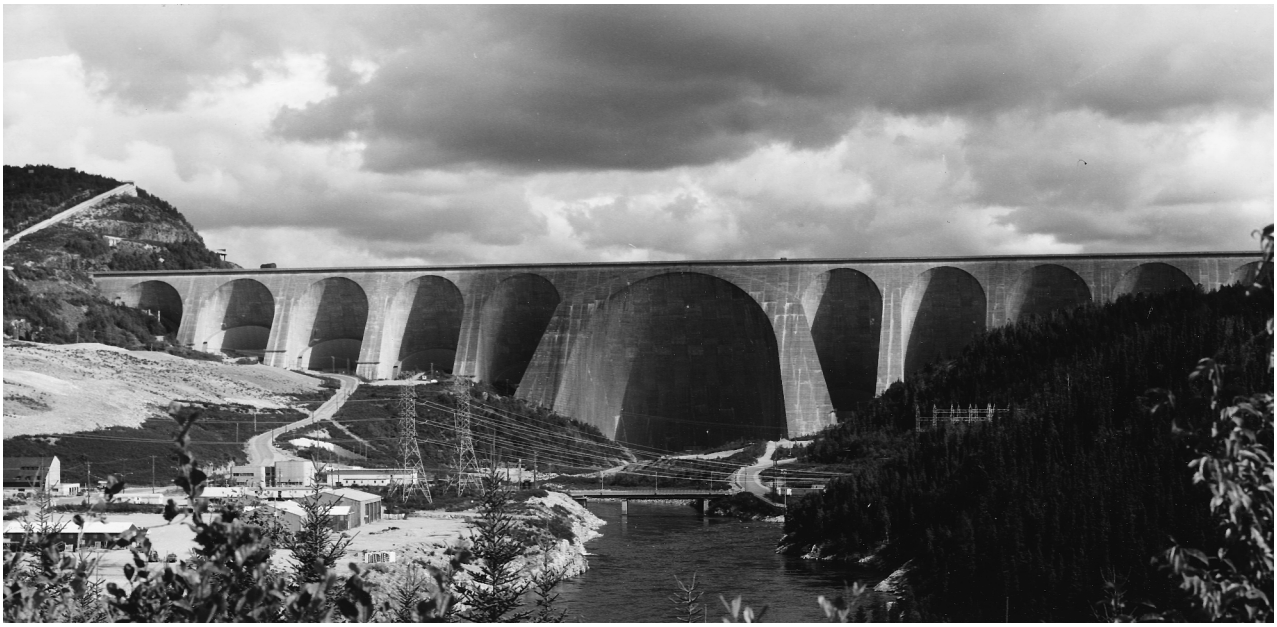


FIGURE 3 Daniel Johnson Dam in Québec, Canada, a multiple-arch structure 214 m (703 ft) high, with the highest arch spanning 161.5 m (530 ft).

applications. The mix is hauled, placed, and compacted by heavy equipment. RCC has the advantages of rapid construction, low cement content and consequent reduced heat generation, less exacting materials requirements, and usually lower cost than in conventional concrete construction. The first RCC dams were gravity structures. Their achieved heights increased rapidly in the 1990s, exemplified by the Japanese Miyagase Dam (155 m/509 ft) in 1996. Yet higher RCC dams are being planned, including Longtan Dam in China, with an ultimate height of 216 m (709 ft). As methods have advanced, arches of this type have become feasible. An outstanding example is Shapai Dam, 128 m (420 ft) high, completed in 1999 on an upper reach of the Chaopohe River in Sichuan Province, China. At that project, use was made of grout-enriched roller-compacted concrete (GERCC) adjacent to forms, a new technique that involves pumping cement grout onto the uncompacted RCC and applying a large internal vibrator to secure a mix of high quality. This and other innovations in the design and construction of RCC dams, including the Chinese-developed sloped-layer placement, have improved bonding and control of seepage, which were problems in the early dams of this type.

F. Cofferdams

Cofferdams are water barriers, usually temporary, for dewatering sites for construction of dams and other waterworks. Steel sheetpile cellular cofferdams typically are filled with sand, gravel, or broken rock. Because of the limited space that they occupy, they have particular advantages where river navigation must be continued during construction of the main facilities. Embankment cofferdams are used commonly where the river is diverted around the worksite. In many projects, the upstream cofferdam has been incorporated into the permanent embankment dam. Retention of the downstream cofferdam is less common, and in such cases may require modifications to ensure internal drainage of the final dam.

G. Saddle Dams

Saddle dam is the term given to a water barrier built in a topographic depression or gap on the rim of a reservoir. It is typically an auxiliary feature to increase the water storage capability. Most saddle dams are relatively small embankments, but some are of substantial size. A current example is the saddle dam of the Eastside Reservoir Project in Riverside County, CA, completed in 1999. It is an earth-core rockfill 130 ft (40 m) high and 2300 ft (701 m) long, with volume of 2,500,000 yd^3. The foundation is a rock (phyllite and schist) ridge that required grouting to depths of as much as 100 ft (30 m).

H. Diversion Dams

A diversion dam is a structure constructed across a stream to turn it into another channel or conduit. It may be a

temporary facility, as a cofferdam, for passage of flow during construction of principal works, or it may be permanent for diversion and conveyance of water for various continuing purposes.

I. Determination of Dam Type

The most suitable type of dam may depend upon many factors, including environment, topography, geology, weather, site accessibility, sources of materials, stream diversion requirements, and the potential for floods and earthquakes.

Earthfills are adaptable to most sites but may be less favored at narrow canyons where they might be subjected to large deformation and where river diversion and flood control would be difficult. They have advantages on alluvial or glacial foundations, in comparison with concrete dams, which are most suited for construction on rock. However, placement and compaction of earthfills may be hampered by excessive rainfall or frost. Soils used in embankments generally have lower shear strength (friction plus cohesion) than does typical rockfill, which is characterized by high frictional resistance to sliding.

Rockfills may be favored where rock sources are plentiful but earth materials are scarce. Dams of this type with earth impervious zones can be susceptible to differential settlement in narrow gorges. Rockfills typically have high shearing strength and therefore can be built with steep faces, with attendant benefits of economic volume and accommodation in topographically limited sites. Because neither rockfills nor earthfills are resistant to sustained overtopping, they usually require separate spillways that may be expensive.

Concrete dams have distinct advantages in their ability to stand overpour. They usually require sound rock foundations, which may preclude some laminated shales and sandstones. Nonuniform rock formations, including those with faults and shear zones, pose special challenges in concrete dam design. For arches, such conditions might be decisive in weighing their comparative merits, or even their feasibility. Gravity dams are well suited for construction on most rock sites of any width. Their accommodation of spillways and outlets may give them an advantage over embankments. Some low gravity dams have been built on soil or gravel foundations, but in most cases the potential for detrimental deformation or erosion would rule against such construction.

Where site conditions have not been amenable to incorporating a spillway in an arch, short tunnels have been economical alternatives for passing floods. Arches and buttress dams may have favorably economic proportions, even though the unit costs of their construction are relatively high. Buttress dams can be built at nearly any site that would accommodate a gravity dam. Their comparatively reduced concrete volumes have given them an advantage in regions where sources of materials were limited and costs of labor were not excessive. However, offsetting factors are the details of the slender structural members, which may necessitate high-strength concrete and steel reinforcement and sometimes precautions to counter the effects of weathering and ambient temperature variations. Because of their smaller bases, buttress and arch dams usually pose less concern about uplift by foundation water pressures than do gravity dams. However, on some foundations with low bearing capacity, buttresses of modest size have been built on concrete base slabs to spread the load. In such a case, uplift may require special attention.

Over the years, advances in technology have altered preferences for the various dam types. Rockfills have become more attractive as improvements have been made in quarrying, zoning, placement, and compaction, as well as in details of face slabs. Roller-compacted concrete has won wider acceptance as measures for seepage control have been developed and as use of RCC for armoring or bolstering embankments has gained favor.

II. HISTORY OF DAMS

Dams were among the earliest structures, having been vital in the development of water resources in Asia and the Middle East. Historically, dams have been associated closely with the rise and fall of civilizations. Many of the ancient water projects survived for centuries and then deteriorated. Without water, the societies that it sustained withered away. The causes of decline were many, including the changes in peoples from generation to generation. Some works were wasted by neglect. Others suffered from inherent problems, such as sedimentation of the reservoirs.

Table I is a list of some notable historical dams. The historian Herodotus wrote of a dam on the River Nile attributed to Menes, the king of Egypt, about 3000 BC. Some translations of the ancient records suggest that the dam was at Kosheish, about 25 miles upstream from Cairo, and was constructed of stone masonry. The remains of other masonry dams are found in Egypt, including the Sadd-el-Kafara Dam in the Wadi Garawi about 20 miles south of Cairo, the origin of which has been dated to between about 2900 BC and 2700 BC. Measurement of the remnants has suggested that this structure was as high as 14 m (46 ft) and consisted of a central zone of gravel, rubble, and fines, bolstered by rockfill sections faced with stepped cut limestone blocks. It evidently was destroyed by flood.

In Babylonia and Assyria, there were diversions from the Tigris and Euphrates Rivers and their tributaries as early as 2100 BC, accomplished by earth and stone dams.

TABLE I Some Notable Historical Dams

Year built[a]	Dam name	Country	Stream	Dam type	Height(m)
3000 BC ±	Kosheish	Egypt	Nile	Masonry	15 (?)
2700 BC ±	Sadd-el-Kafara	Egypt	Wadi el-Garawi	Earth/masonry	14
2100 BC ±	Adheim	Iraq	Adheim	Masonry	17
2000 BC ±	Marduk	Iraq	Tigris	Earth	12
750 BC ±	Sadd-el-Arim	Yemen	Wadi Dhana	Earth/masonry	14
591 BC ±	Shaopi	China	—	Earth	10
240 BC ±	—	China	Gukow	Stone-crib	30
150 BC ±	Alcantarilla	Spain	Guajaraz	Masonry/earth	20
110 AD ±	Proserpina	Spain	Pardillas	Masonry/earth	19
162 AD	Kaerumataike	Japan	Yodo	Earth	17
190 AD ±	Cornalbo	Spain	Albarregas	Earth	24
960 AD	Band-I-Amir	Iran	Kur	Masonry	9
1050 AD ±	Bhojpur	India	Kaliasot	Earth	27
1300 AD ±	Kebar	Iran	Kebar	Masonry arch	26
1500 AD	Mudduk Masur	India	—	Earth	33
1594 AD	Alicante (Tibi)	Spain	Monegre	Arched gravity	41
1671 AD	St. Ferreol	France	Laudot	Earth/masonry	36
1791 AD	Valdeinfierno	Spain	Luchena	Arched gravity	36
1800 AD ±	Mir Alam	India	From Esee River	Multiple arch	12
1854 AD	Zola	France	—	Masonry arch	42
1866 AD	Gouffre d'Enfer	France	Furan	Arched gravity	60
1897 AD	Periyar	India	Periyar	Concrete gravity	54
1904 AD	Cheesman	U.S.	South Platte	Arched gravity	72
1905 AD	New Croton	U.S.	Croton	Masonry gravity	90.5
1910 AD	Buffalo Bill	U.S.	Shoshone	Concrete arch	99
1915 AD	Arrowrock	U.S.	Boise	Arched gravity	107
1931 AD	Salt Springs	U.S.	N. Fk. Mokelumne	Rockfill	100
1932 AD	Owyhee	U.S.	Owyhee	Arched gravity	127
1936 AD	Hoover (Boulder)	U.S.	Colorado	Arched gravity	221

[a]The ± sign indicates approximate date of construction.

In the second and first millennia BC, diversion dams were built in southern Arabia, the largest evidently being the 14-meter-high (46-ft-high) Sadd-el-Arim Dam (also referred to as Sadd Ma'rib or Sudd-al-Arim) near Ma'rib, capital of Sheba (Saba), in the region now known as Yemen. It has been described as an earth embankment with an upstream face protected by a layer of stones. Its construction has been dated to about 750 BC. This is believed to be the structure in the Wadi Dhana (also identified as Wadi Denne or Wadi Sadd or Wadi Saba), where remnants of canal intakes and a spillway have been found.

Other notable dams were built in Sri Lanka (formerly Ceylon), India, Baluchistan (now part of Pakistan), and China in the millennia BC. A Chinese dam of special interest is the 10-meter-high (33-ft-high) earthfill forming Shaopi Reservoir, which was completed in 591 BC and is still operating for supply of irrigation water in Anhui Province.

Important water projects were developed in Spain by the Romans. The Alcantarilla Dam near Toledo, believed to have been constructed in the second century BC, was composed of masonry and concrete buttressed by earthfill. Later Spanish waterworks built by the Romans include Proserpina and Cornalbo Dams, near Mérida. In these and many other dams in the Roman empire, effective use was made of stone masonry. Records dating to about 560 AD tell of a curved dam on a tributary of the Khabur River at the Turkish-Syrian border near Daras. Some historians regard it as the first known arch dam. It disappeared long ago.

Significant advances were made in the engineering of dams in the Middle Ages. An important, long-surviving arch dam of the Mongol period in Persia was built on the Kebar River near Qum, southwest of Teheran in about 1300 AD. About 26 m (85 ft) high and 55 m (180 ft) long, it was composed of cemented rubble masonry with mortared stone block facing. An arch-gravity dam of similar

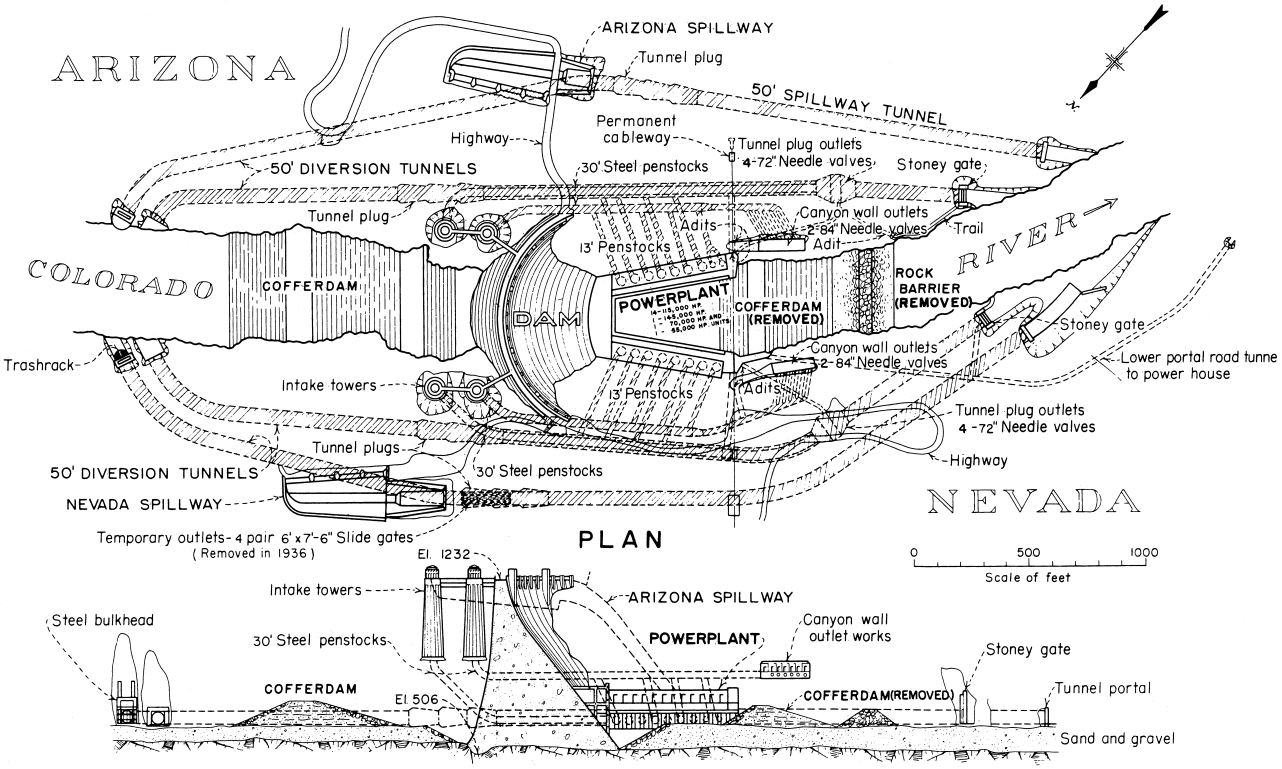


FIGURE 4 Hoover Dam on the Colorado River.

composition was constructed in 1384 AD near Almansa in Spain. After successive enlargements, the latest in 1921 AD, the height of the Almansa Dam reached 25 m (82 ft). This was surpassed by another curved masonry dam in Spain, the Alicante (Tibi) Dam, completed in 1594 AD. For nearly three centuries, this was the highest dam in the world, towering 41 m (135 ft) above its base. Its height was increased to 46 m (151 ft) in 1943.

The earliest true multiple-arch dam in the records is Mir Alam Dam, constructed in about 1800 AD near Hyderabad, India. This mortared masonry structure, about 12 m (40 ft) high, had 21 vertical arches. In the 19th century, significant progress was made in the engineering of dams. Methods of analysis of concrete and masonry structures were advanced in England and France. Engineers in the United States built notable dams, including Old Croton Dam in 1842 for New York City; Mill River Dam in 1862 for New Haven, CN; Lake Cochituate Dam in 1863 for Boston; and Druid Lake Dam in 1871 for Baltimore. Many earth dams, as high as 38 m (125 ft), were constructed for irrigation projects in the western states. Dams made of stone-filled log cribs were developed for storage of water for hydraulic mining in the California gold fields. Larger reservoirs were created by dumped rockfills with timber plank facing.

An American precedent was established in the construction in 1904 of Colorado's Cheesman Dam, 72 m (236 ft) high and arched even though it was a full gravity section. New Croton Dam, a gravity structure 90.5 m (297 ft) high completed in 1905, was one of the last large American dams made of masonry. It was the highest dam in the world, but was surpassed in 1910 by Buffalo Bill (Shoshone) Dam in Wyoming at 99 m (325 ft). This is a true arch, as is Pathfinder Dam, another Wyoming dam, completed in 1909.

Many advances were made in the engineering of concrete dams in the 20th century as design and construction methods evolved rapidly. Hoover (Boulder) Dam, completed in 1936 on the Colorado River, is an arched gravity concrete structure 221 m (726 ft) high (Fig. 4). On the same river is Glen Canyon Dam, an arch dam completed in 1966. It is 216 m (710 ft) high. Dworshak Dam, a straight concrete gravity structure completed in 1973 on the North Fork of the Clearwater River in Idaho, is 219 m (717 ft) high (Fig. 5). Shasta Dam (1945) in California is of the arched gravity type, 183 m (600 ft) high (Fig. 6). Mossyrock Dam (1968) in the State of Washington is an arch 185 m (607 ft) high (Fig. 7). Other monumental achievements in concrete construction have been Vaiont Dam (1961) in Italy, and three Swiss Dams: Mauvoisin

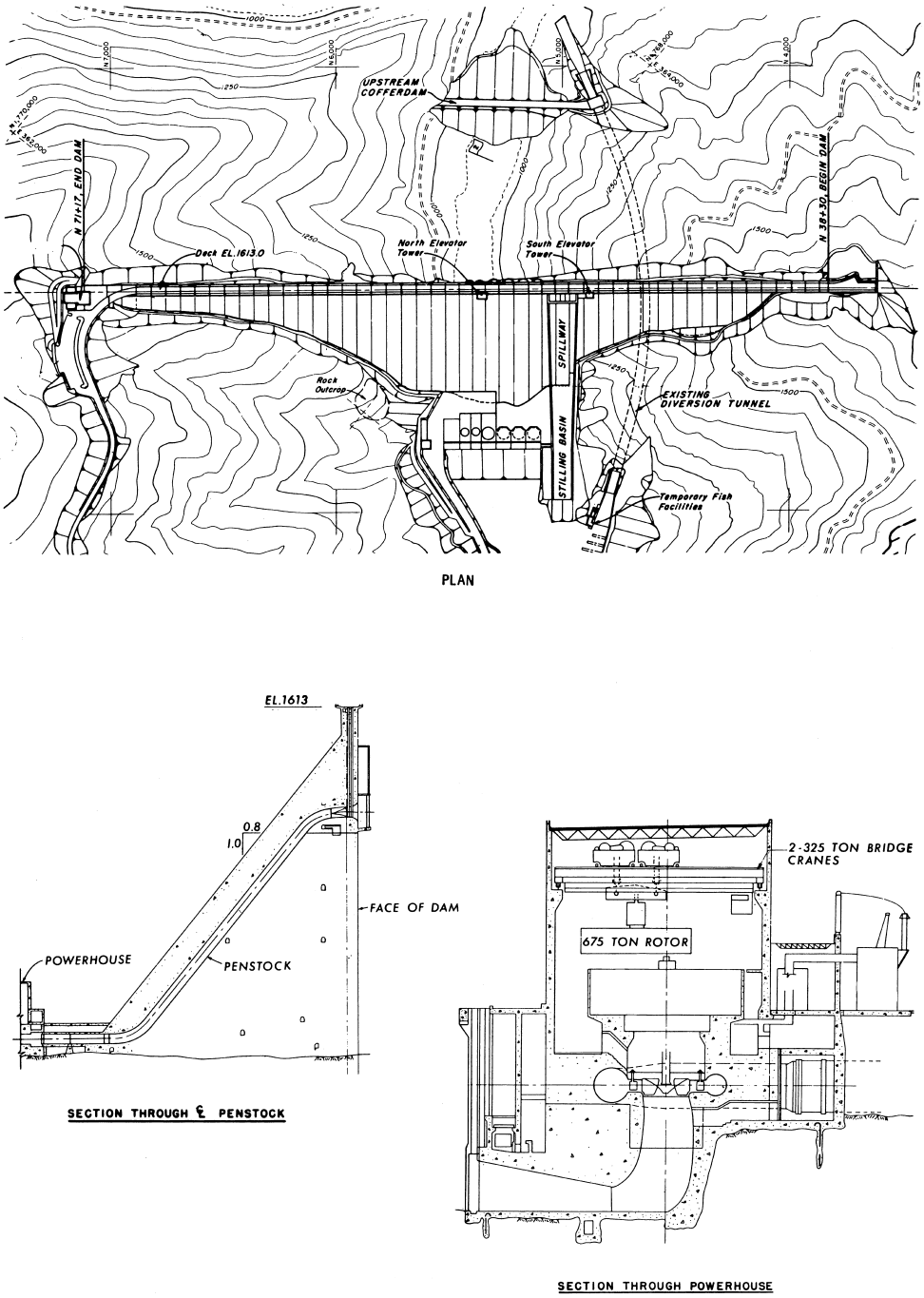


FIGURE 5 Dworshak Dam on the North Fork of the Clearwater River in Idaho.

(completed originally in 1957 and raised in 1988), Grande Dixence (1961), and Contra (1965) (see Fig. 8). These were followed by Inguri, Sayano-Shushensk, and Chirkei, all shown in Table II.

Major advances have also been made in the engineering of embankment dams. Outstanding examples are California's Oroville Dam (1968), an earthfill 230 m (755 ft) high (Fig. 9); Mica Dam (1973), an earthfill-rockfill 242 m (794 ft) high in western Canada; La Esmeralda (Chivor) (1976), an earthfill-rockfill 237 m (777 ft) high in Colombia (Fig. 10); Manuel M. Torres (Chicoasén) (1980), an earthfill-rockfill 261 m (856 ft) high in

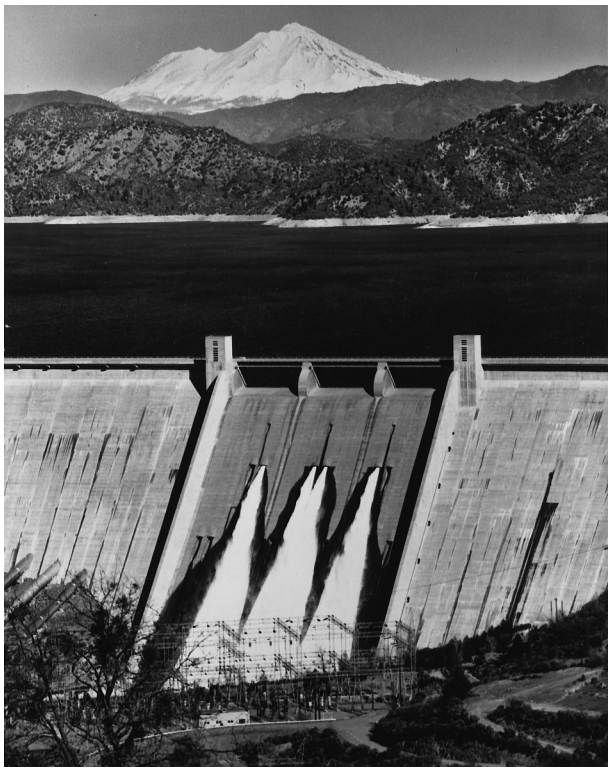


FIGURE 6 Shasta Dam on the Sacramento River in California.

Mexico; Alberto Lleras Camargo (Guavio) (1989), an earthfill-rockfill 243 m (797 ft) high in Colombia; Tehri (under construction), an earthfill-rockfill 261 m (856 ft) high in India; and Nurek (1980), an earthfill 300 m (984 ft) high in Tajikistan. A notable example of a rockfill with a concrete face is the 160-meter-high Foz do Areia Dam, completed in 1980 in Brazil. Figure 11 is a section of this dam, showing the geometry of its zones and the varied thicknesses of the compacted rockfill layers.

As the 21st century dawned, building of large dams in China was continuing at a vigorous pace, including Ertan (1999), an arch 240 m (787 ft) high, as well as Hongjiadu, a concrete-faced rockfill 182 m (597 ft) high, and Three Gorges, a 181-m (594-ft) gravity dam, both under construction. The Chinese have adopted and advanced new concepts in the design and construction of roller-compacted concrete dams and concrete-faced rockfills.

Sharing of knowledge about dams throughout the world has been facilitated by the work of the International Commission on Large Dams (ICOLD), founded in 1928. Its Congress in Beijing in 2000 listed 83 member countries. This cooperation has enabled significant advances in dam engineering with consequent enhancement of national economies and public safety.

III. APPURTENANT OR RELATED WORKS

A. Spillways

Spillways are facilities for conveying excess streamflows past a dam. In important projects on major rivers, spillways are sized to discharge extreme floods. At a concrete gravity dam, the spillway is typically made part of the main structure, providing an overflow that is either uncontrolled or regulated by gates (Fig. 12). At an embankment dam, the spillway is usually a channel (chute) or conduit in a separate location in or on an abutment or at a farther site on the reservoir rim (Fig. 13). The types of spillways include overflow with an ogee-crest shape and shaft and tunnel with a glory hole (morning glory), funnel-shaped entrance. Another type that has been used under special conditions is the side channel, in which the crest is approximately parallel to a discharge channel consisting of a trough leading to a tunnel or chute. Hoover Dam has two spillways of this kind (Fig. 4). At some arch dams, the spillway design entails free fall from the crest to a plunge pool created by a smaller dam immediately downstream. At other types of dams, where the streambed would be subject to erosion by high-velocity flow, the potentially destructive energy of the discharging water is dissipated in a stilling basin. Terminal structures also include chutes and aprons with baffles, and flip (deflector) buckets, sometimes referred to as ski-jump spillways, which loft a trajectory so it will impinge at a safe distance.

At many spillways, gates and other devices have been used to control discharges. In the past, flashboards often were placed on the crest to raise the reservoir level. A few of these are still seen at older projects. Some of them were simple, temporary barriers fabricated of planks or timber panels designed to fall and be washed away at predetermined flood levels. Others were, in effect, permanent automatic gates that lowered in response to rising water pressure. A typical flashboard of this type is hinged and counterweighted so that it rotates and lies down to conform to the fixed crest shape during flood passage and then rises to resume its retention of the normal water surcharge. In modern controlled spillways, conventional designs include hinged steel gates of the radial (taintor) type lifted by electric hoists on decks overhead. To safeguard against power outages during storms, provision often is made for standby generators using diesel or gasoline fuel.

B. Outlets

Outlets are facilities for water discharges to serve normal project requirements, as well as to augment flood control capacity in some cases. They are usually classified

TABLE II Some High Dams of the World[a]

Dam name	Country	Stream	Dam type	Height (m)	Year completed
Nurek	Tajikistan	Vakhsh	Earth	300	1980
Grande Dixence	Switzerland	Dixence	Gravity	285	1961
Inguri	Georgia	Inguri	Arch	272	1980
Yusufeli[b]	Turkey	Coruh	Earth/rock	270	Underway
Vaiont[c]	Italy	Vaiont	Arch	262	1961
Manuel M. Torres (Chicoasén)	Mexico	Grijalva	Earth/rock	261	1980
Tehri	India	Bhagirathi	Earth/rock	261	Underway
Mauvoisin	Switzerland	Drance de Bagnes	Arch	250	1957/1988
Deriner	Turkey	Coruh	Arch	247	Underway
Alberto Lleras C. (Guavio)	Colombia	Guavio	Earth/rock	243	1989
Mica	Canada	Columbia	Earth/rock	242	1973
Sayano-Shushensk	Russia	Yenisey	Gravity/arch	242	1987
Ertan	China	Yalongjiang	Arch	240	1999
La Esmeralda (Chivor)	Colombia	Batá	Earth/rock	237	1976
El Cajón	Honduras	Humuya	Arch	234	1985
Chirkei	Russia	Sulak	Arch	233	1978
Oroville	U.S.	Feather	Earth	230	1968
Bhakra	India	Sutlej	Gravity	226	1963
Luzzone	Switzerland	Brenno di Luzzone	Arch	225	1963/1998
Hoover	U.S.	Colorado	Arch/gravity	221	1936
Contra	Switzerland	Verzasca	Arch	220	1965
Mratinje	Montenegro	Piva	Arch	220	1976
Dworshak	U.S.	N. Fork Clearwater	Gravity	219	1973
Glen Canyon	U.S.	Colorado	Arch	216	1966
Toktogul	Kyrgyzstan	Naryn	Gravity	215	1978
Daniel Johnson	Canada	Manicouagan	Multiple arch	214	1968
Keban	Turkey	Firat (Euphrates)	Earth/rock/gravity	208	1974
Zimapan	Mexico	Moctezuma	Arch	207	1994
Karun III	Iran	Karun	Arch	205	Underway
Lakhwar	India	Yamuna	Earth	204	Underway
Dez	Iran	Dez	Arch	203	1962
Almendra	Spain	Tormes	Arch	202	1970
Berke	Turkey	Ceyhan	Arch	201	Underway
Khudoni	Georgia	Inguri	Arch	201	Underway
Karun I	Iran	Karun	Arch	200	1976
Kölnbrein	Austria	Malta	Arch	200	1977
San Roque	Philippines	Agno	Earth/rock	200	Underway
Itaipú	Brazil/Paraguay	Paraná	Earth/rock/gravity	196	1983
Altinkaya	Turkey	Kizil Irmak	Earth/rock	195	1988
New Bullards Bar	U.S.	North Yuba	Arch	194	1970
Seven Oaks	U.S.	Santa Ana	Earth/rock	193	1999
New Melones	U.S.	Stanislaus	Earth/rock	191	1979
Miel I	Colombia	Miel	RCC	188	Underway
Aguamilpa	Mexico	Santiago	Rock	187	1994
Kurobe	Japan	Kurobe	Arch	186	1964
Swift	U.S.	Lewis	Earth	186	1958
Zillergründl	Austria	Ziller	Arch	186	1987
Mossyrock	U.S.	Cowlitz	Arch	185	1968
Oymapinar	Turkey	Manavgat	Arch	185	1984
Atatürk	Turkey	Firat (Euphrates)	Earth/rock	184	1990
W.A.C. Bennett	Canada	Peace	Earth	183	1967
Shasta	U.S.	Sacramento	Arch/gravity	183	1945

[a] Dams 183 m (600 ft) or higher.
[b] Yusufeli Dam design was complete in 1999; financing and construction contracts were set in 2000.
[c] Slide of a canyon wall in 1963 made Vaiont Reservoir inoperable.

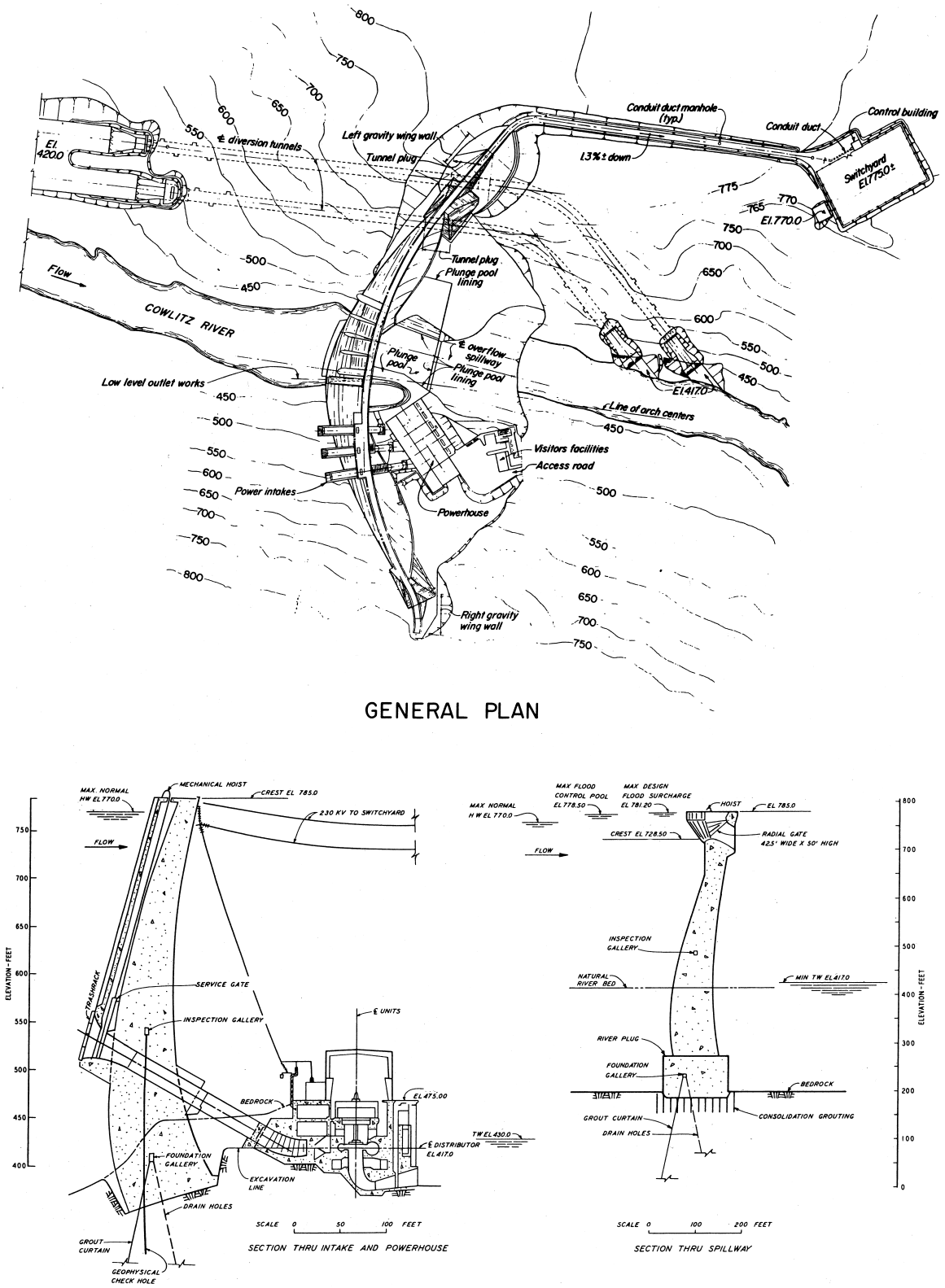


FIGURE 7 Mossyrock Dam on the Cowlitz River in Washington.

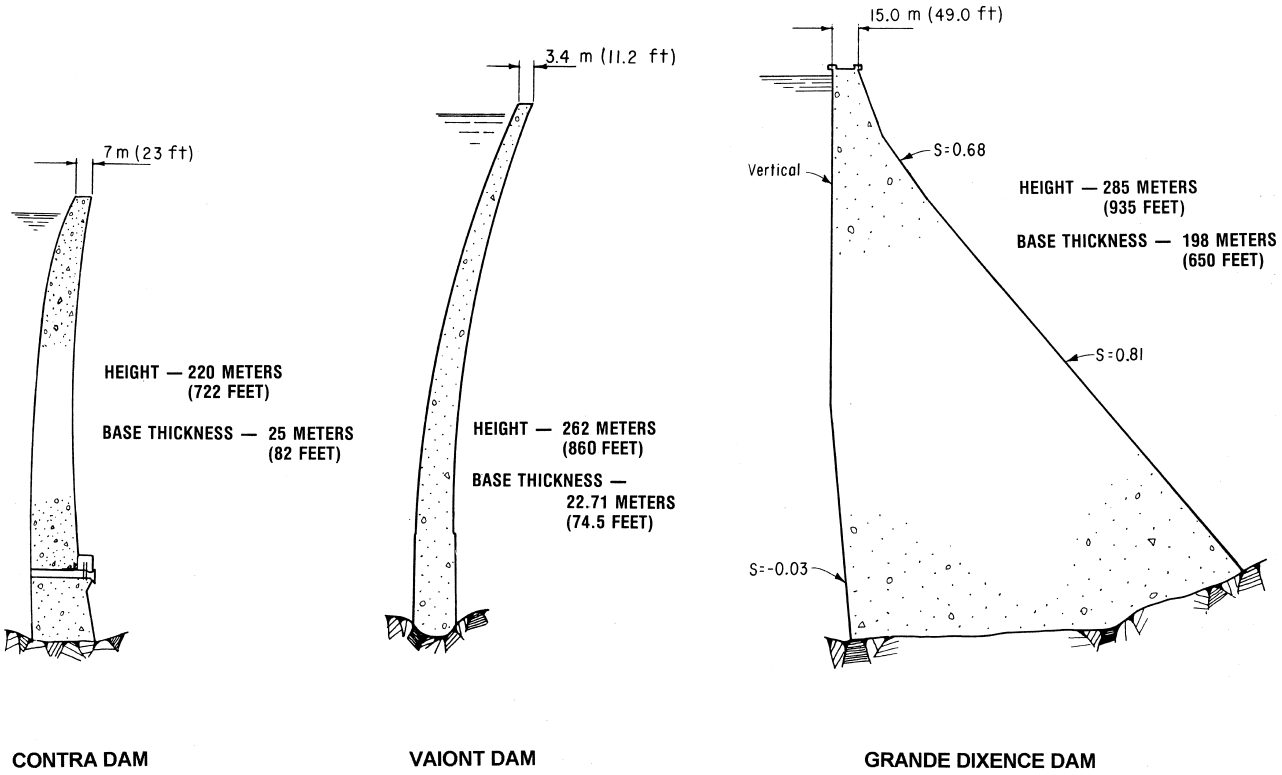


FIGURE 8 Sections of high concrete dams, showing the contrast in volume of arch and gravity structures.

by their purposes, including irrigation, municipal or industrial water supply, power, flood, and river release. River outlets at concrete dams are commonly called *sluices*. Outlets may be of high value in releasing stored water to lower the reservoir level in case of emergency. Where dissipation of the energy of the discharges is required, it is conventionally accomplished by devices and structures similar in concept to those used for this purpose at spillways. The flows through outlets are regulated by gates or valves, which may be of a variety of types. Depending upon the purposes to be served, the outlets may discharge to the open air or they may convey the flows into penstocks for power installations or into open or closed conduits for other water supply needs. To satisfy water quality requirements at some projects, the outlet works have been designed with multiple-level intakes that enable withdrawals at optimum temperatures and chemical content. At Oroville Dam, the inclined intake structure for the penstocks leading to the underground Edward Hyatt Powerplant was provided with water-level control shutters so that selection could be made of cool water for salmon and steelhead in the Feather River or of warmer water for rice cultivation in the region, as the season demands. (At some other projects, downstream temperature requirements have been met by warming ponds.) At Shasta Dam, also in California, a multilevel temperature control facility was added in 1997, at a cost of $80 million, to enable selective withdrawal for protection of endangered salmon species.

C. Other Works

Facilities that may be closely related to dams are powerhouses, navigation locks, and fish ladders. Where such works adjoin or are integral parts of the water barrier, they must meet the same structural demands as the dam itself.

IV. DESIGN OF DAMS

Methods of design of dams and other water barriers have progressed substantially in recent decades, drawing fully from the interdisciplinary work of engineers, geologists, hydrologists, and seismologists, and from research that has expanded the horizons in this field. Increasingly effective use has been made of physical and mathematical models in hydraulic and structural analysis. Observations of the performances of dams during earthquakes and the examination of histories of accidents and failures have shown the way to safer dams.

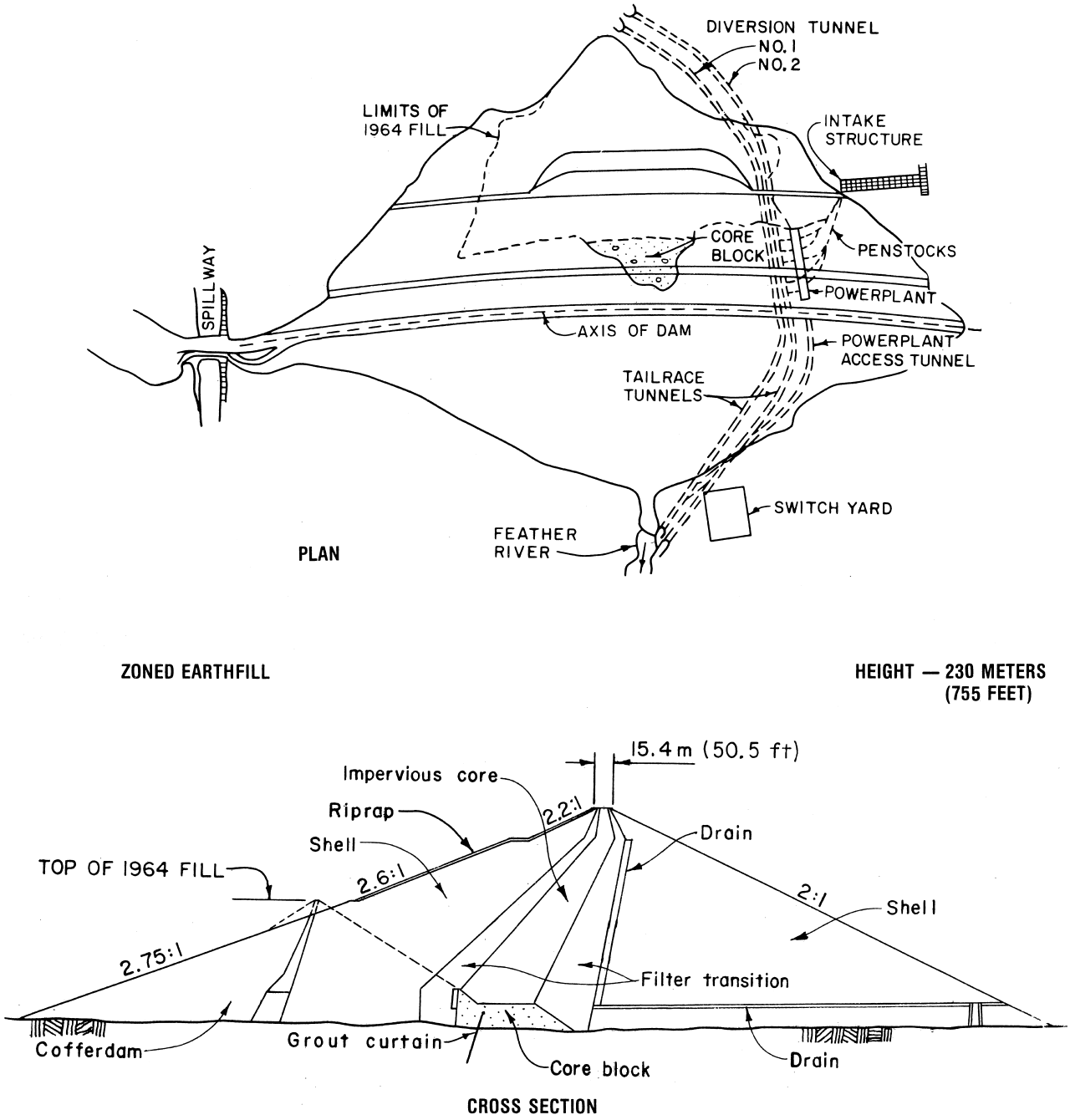


FIGURE 9 Oroville Dam on the Feather River in California.

A. Earthfills

Design of earthfills has evolved as the science of soil mechanics has advanced and become an integral part of the broad field of geotechnical engineering. One of the important fundamentals is the relationship between the optimum degree of compaction and the moisture of the soil being placed. Rudimentary methods of slope stability analysis, while still useful, have been joined with more sophisticated procedures for assessing the response to loading, particularly to the effects of seismic vibration and consequent variation in internal water pressures. The value of the new techniques has been measured by instrumental monitoring of real structural behavior. Control of seepage through earth zones is ensured by various design elements, which may include foundation cutoffs, impervious earth

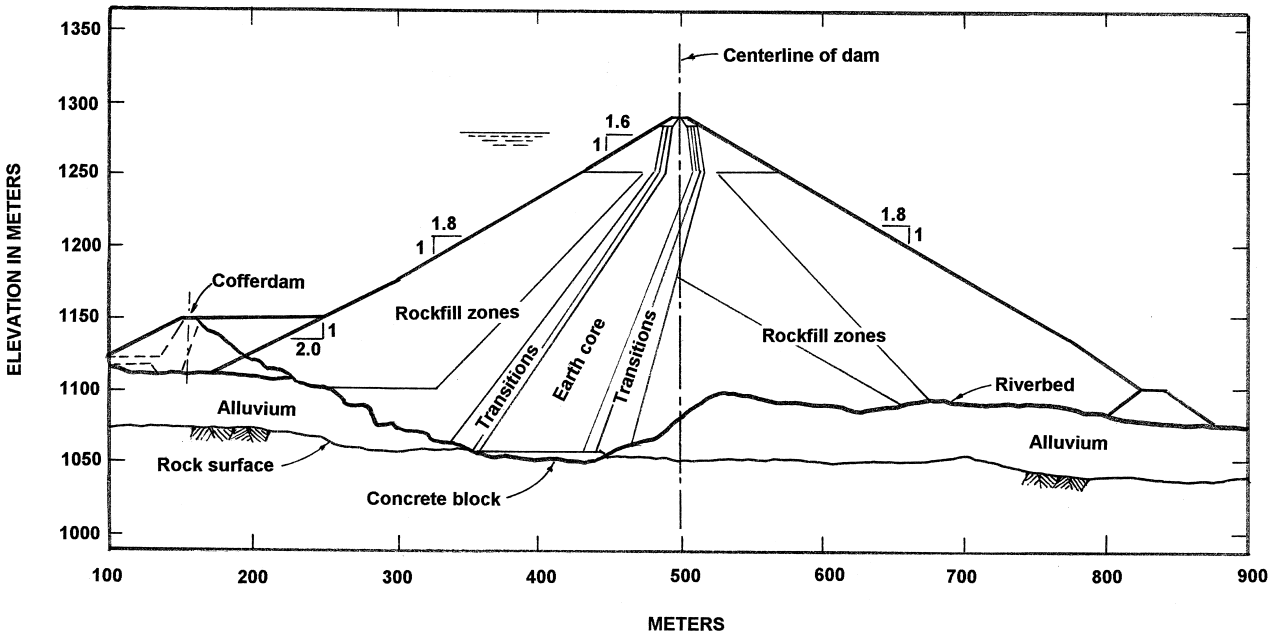


FIGURE 10 La Esmeralda (Chivor) Dam on the Batá River in Colombia.

blankets on the reservoir floor immediately upstream from the dam, internal filter and transition zones, drains, or pressure relief wells downstream from the embankment.

B. Rockfills

Design of rockfills also has benefited from these advances, although the proportioning of this type of embankment is still comparatively simple, drawing amply from successful precedent. A fundamental consideration in the design of rockfill embankments is the gradation of the constituent materials, which influences stress–strain characteristics and permeability. The actual condition of the *in situ* fill is determined not only by the sizes and durability of the particles, but also by the thickness of the placed layers and the degree of compaction, a basic objective being to

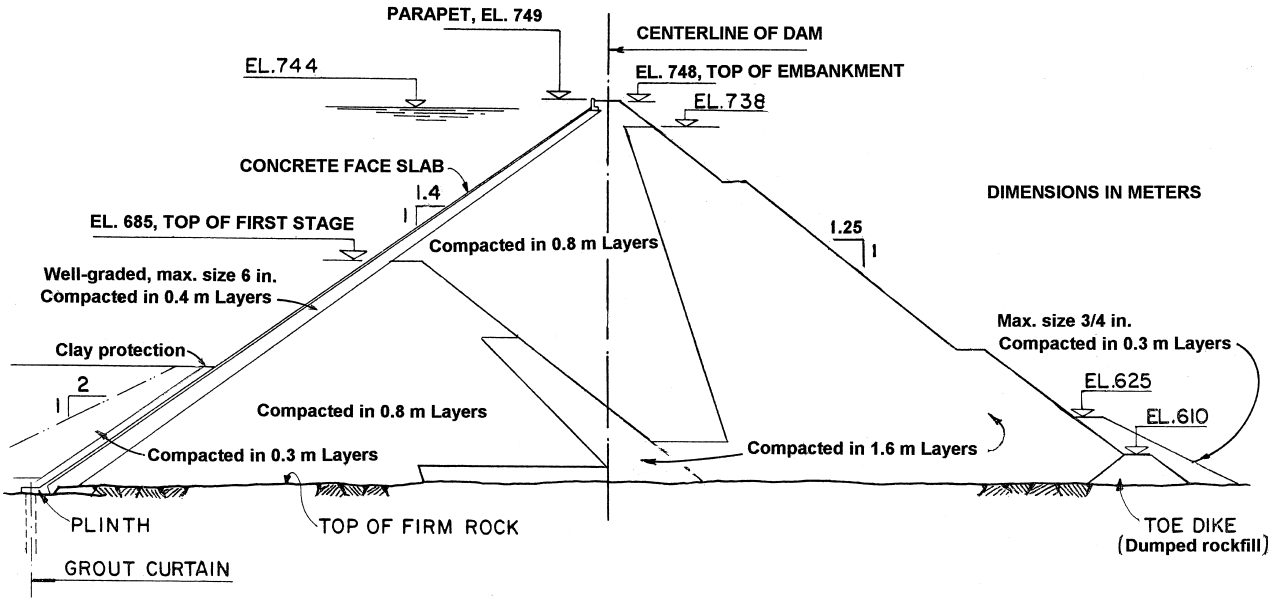


FIGURE 11 Foz do Areia Dam, a rockfill on the Iguaçu River in Brazil.



FIGURE 12 Piedra del Aguila Dam on the Limay River in Argentina, a concrete gravity structure 174 m (571 ft) high. The project was completed in 1992. The spillway is in the right of the view.

ensure that settlement will be minimal so that the impervious element will not be broken.

The design of concrete-faced rockfill dams has been greatly improved in recent times and this type has won acceptance at many new projects. Primary advances have been made in the required degree of compaction, in the details of the plinth and the face slab, and in the gradations of the rockfill zones, especially under the face slab and at its edges, where finer rock sizes are specified. Careful attention is given to the reinforcement and anchoring of the plinth and to treatment of the foundation under and immediately downstream from it. To limit deformation at the perimeter joint of the face slab, typical specifications call for the underlying fill to be fine rock filter that is specially compacted. Rockfill stability is analyzed by procedures that have been applied to earthfills, assessing the potential for sliding on assumed internal planar or curved surfaces, using friction factors determined by experimentation or experience with comparable materials at other projects.



FIGURE 13 Alicura Dam on the Limay River in Argentina, an earthfill 130 m (427 ft) high, completed in 1984. The dam is on the left, the powerhouse is in the center, and the spillway is on the right.

C. Gravity Dams

A typical gravity dam comprises spillway and nonoverflow blocks, some of them including outlet works and intake facilities for control of water releases. Low-level outlets for river discharges are located typically in the spillway section. Intakes for penstocks leading to a powerhouse are commonly incorporated into a nonoverflow section. A long-standing objective has been to avoid tension in the unreinforced concrete mass. To establish an adequate base width to ensure stability, the downstream faces of the blocks are usually inclined no steeper than about 55 degrees with the horizontal, which is expressed as a 0.7:1 slope. The actual geometry in each project is determined by analysis of the forces that might be imposed, including the loads from extreme floods and earthquake, as well as from possible silt and ice forces on the upstream face. The overflow crest in the spillway section is conventionally shaped to conform to the trajectory of a discharging stream. The mass concrete of the common dam of this type does not have steel reinforcement, but it is provided in parts subject to tensile stresses, as in walls and piers and around galleries and interior valve and gate chambers. A gravity dam typically is founded on rock, which may be grouted to reduce its permeability. The foundation may also be provided with drains to control uplift pressures. Both of these treatments are common in modern dams.

D. Arch Dams

Design of arch dams has evolved as techniques for determining their optimal geometry have been refined. Structural behavior is analyzed by complex computer programs and verified by instrumental measurement of actual response to loading. These advances have enabled a significant change from the early simple designs to present-day, thin, high-strength concrete sections with varied curvature to optimize stress distributions. For the analysis of arches, as well as other kinds of dams, widespread use is being made of the finite element method, which entails mathematically modeling the structure as a composite of individual elements joined at nodes to form a mesh. Nodal deflections and element stresses are computed by solving a series of simultaneous equations.

E. Buttress Dams

The stability of a buttress dam depends primarily on the slope of the upstream face, which typically has been set between about 45 and 50 degrees with the horizontal. The

buttresses of flat-slab and multiple-arch dams must be designed to carry the loads transmitted by the slabs or arches as well as to resist lateral forces, including importantly the impact from seismic load. They may be solid, single walls, or double-walled structures. Whatever the case, both the upstream watertight members and the buttresses are provided with steel reinforcement. For lateral stability, the buttresses of many dams of this type have counterforts or stiffeners incorporated as flanges or pilasters. Some of them have also been braced by struts in the spans between buttresses.

V. CONSTRUCTION OF DAMS

A. Stream Diversion

Although some dams are built offstream, as in pumped storage projects and for other special purposes, most are located where construction requires isolation of the work areas from natural streamflows. This is commonly accomplished by cofferdams that enable temporary redirection of discharges through a tunnel in an abutment or over a restricted part of the streambed. In the latter alternative, the construction may be done in stages so the configuration of the cofferdams changes as the work progresses. In the construction of concrete dams, diversion may be facilitated by passing the river over low blocks or through sluices.

B. Earthfill Dams

Construction of earthfill dams has evolved with the development of high-capacity hauling and compacting equipment. The fill is compacted in layers, commonly using sheepsfoot rollers for finer grained soils, rubber-tired rollers for impervious materials well-graded from gravel to sand to silt and clay, and vibratory rollers for pervious zones. Because of the inherent susceptibility of some soils to erosion, strict attention must be given especially to the gradation and compaction of material at the interfaces between zones and between the embankment and its foundation and adjoining structures. At the boundaries, protection is ensured by measures that prevent the migration of fine particles, including intermediate filter zones composed of compatibly graded coarser materials, concrete membranes, and grouting of the underlying rock formation. Because the natural conditions at each project are unique and may not be known adequately until construction is well under way, the final details of an earthfill may have to be determined as the work at the damsite proceeds. Designers and allied specialists, including geologists, therefore must continue to be involved.

Oroville Dam, the highest dam in the United States, was completed in 1968 and is a landmark structure that introduced important innovations and set the stage for construction of very high embankments. It has a concrete core block 39 m (128 ft) high with a volume of 193,000 m^3 (252,000 yd^3) which sealed a gorge and provided a base for construction of the dam's earth core. The block served as a key element in the first year of construction by retaining fill in a zone immediately upstream, affording flood protection while the works for diversion of the Feather River were under way. The outer zones (shells) of the dam, as seen in Fig. 9, are composed of gravels from dredger tailings left by gold mining in the area. A 668-ton, self-propelled wheel excavator removed these materials from the pits at a rate of 5700 tons per hour. Its rotating wheel, 30 ft in diameter, had eight large buckets. The materials were transported to hoppers at loading stations in the borrow area by a system of conveyor belts and by 100-ton, bottom-dump trucks loaded by draglines. Forty-car railroad trains were loaded and traveled 19 km (12 miles) to the damsite, where they were unloaded two cars at a time by a car dumper with swivel couplings that enabled rotating the cars and dumping without separating the cars from the rest of the train. Special laboratory facilities were constructed for testing the large gravel and cobble sizes. This involved the largest testing machine of its type in the United States which could accommodate a specimen of earth and rock 36 inches in diameter and 7.5 ft high.

C. Rockfill Dams

Close coordination of design and construction engineers is also required throughout the development of a rockfill project. The ultimate foundation treatment and the optimum use of the available materials may have to be decided as the site conditions are exposed. The foundations for rockfills are usually sound rock, but at some projects dense glacial or alluvial deposits have served as a base for pervious embankment zones. The properties of the rock sources influence construction procedures. Since about 1960, the increasingly common use of vibratory rollers for compaction of rockfill in relatively thin layers has greatly improved the quality of this type of dam. Before that, the established practice was to dump and sluice the rockfill in high lifts without mechanical compaction, which resulted in embankments that were susceptible to excessive deformation and consequent cracking of the impervious elements (concrete face slab or earth core). In current projects, the common transport of rockfill to the damsite from the quarry or other source is by heavy dump trucks or bottom-dump wagons. Final placement and leveling of the rockfill layers are generally by bulldozer for the large sizes

and often by grader for the finer material. The amount of water applied to the fill to enhance the compactive effort varies with the quality of the rockfill, the hardest and least friable rock requiring a minimum of sprinkling.

Illustrative of current U.S. practice in construction of large embankment dams is the Eastside Reservoir Project of the Metropolitan Water District of Southern California, completed in 1999. It comprises three earth-core rockfill dams with total volume of 84 million m^3 (110 million yd^3), 70% of which is rockfill. The project involved the largest available heavy equipment and machinery and set new records for handling earth and rock materials. Loading of the rock required five hydraulic shovels, each with 17 m^3 (22 yd^3) capacity, plus large front-end loaders. Dump trucks weighing up to 350 tons when full hauled the rock to the embankment, where it was spread by 100-ton bulldozers. The rockfill was compacted by smooth-drum vibratory rollers. The materials for the earth cores of the dams were excavated and moved by belt loaders at the source and hauled to the fill by 120-yd^3 belly-dump trailers. The three dams required a total of about 10 million yd^3 of processed rock for filters and drains. The larger of two rock-crushing plants for this purpose had a maximum design capacity of 3200 tons per hour.

D. Concrete Dams

Construction of concrete dams starts with site preparation that entails foundation excavation to suitable rock, placement of concrete for sealing and shaping the excavated surfaces, deep grouting to establish a seepage curtain under the structure, sometimes shallow grouting to consolidate the rock, and drilling of holes for drainage. For construction of the dam itself, careful attention is given to the aggregates of the concrete mix. Mass concrete especially requires control of aggregate temperatures to help in limiting cracking. Cooling the placed blocks of gravity dams has been done by passing cold water through embedded piping and by installation of surface insulation.

VI. SAFETY OF DAMS

A. Elements of Safe Dams

Populations in many countries live in close proximity to dams and depend upon their benefits and their safety for survival. These vital structures require a high level of defensive engineering to anticipate and cope with potentially adverse conditions. They must incorporate multiple safeguards to control the hazards that may be inherent at each site. There are reliable ways to design and maintain safe dams, drawing from advanced technology and from the experiences at reservoirs that have failed. The characteristics of the types of dams and their foundations provide guidance. Gravity dams founded on sound rock are intrinsically stable, as are arches with unyielding abutments. Clean, hard, well-compacted rockfill is highly resistant to deformation and internal erosion. Earthfills properly zoned for seepage control and protected at structural and foundation boundaries can endure under a full range of loadings. Buttress dams composed of high-quality, well-reinforced concrete, constructed on stable rock and braced laterally, have served for many years.

B. Instrumentation and Surveillance

Public confidence in the safety of dams is sustained by long records of trouble-free performance. Risks have been limited by continuously improved engineering and by strict surveillance programs, including overview by governmental agencies. These efforts are focused primarily on control of those usually small amounts of water that might escape from the reservoir through the dam or through its foundation. Safeguards must also be provided against extreme floods and seismic events, although the probabilities of their occurrence are generally very low.

Dams must be watched to ensure early detection of abnormal conditions or deviations in their performance. This surveillance relies heavily upon regular visual inspection by experienced observers, but also includes analysis of data from instrumentation. In all cases, the types and numbers of installed instruments should be determined so that problems can be pinpointed and the volume of data to be processed is not so excessive that there may be delays in giving an alert.

An earthfill dam needs to be monitored to detect deformation, cracking, leakage, and erosion. Variations in conditions must be watched to understand their causes and the possible adverse indications. Particular attention must be directed to foundation and structural interfaces where seepage may be conducive to erosion. This requires that regular measurements be made of the water passing through the dam and its foundation and that individual paths of leakage be traced. Monitoring is done commonly at weirs at or near the downstream toe of the embankment, giving attention also to any indications of eroded materials. Internal instruments for measuring water pressures (piezometers), earth pressures, deformation, and temperatures are also useful in assessing the dam's behavior under the changing conditions that are imposed. Diagnostic devices include borehole cameras, hydrophones, and current meters for examining the inside of the embankment. Emphasis is placed on quick analysis of surveillance readings to highlight any deviations from established norms. This has been facilitated by electronic data entry, telemetering,

and computer processing. However, most problems at embankments have been found by inspectors on the ground.

Water pressures at the bases of concrete dams and in the foundations are measured by uplift cells (pore pressure cells) and by piezometers in holes drilled in the rock. Structural movement is monitored by surveys, as well as by joint meters, and pendula (plumblines) in interior shafts. Other instrumentation may include strain meters, stress meters, thermometers, and weirs or other devices for measuring leakage.

In seismically active regions, accelerometers may be installed on dams and in their environs to measure patterns of response to tremors. For embankments, significant data may also be obtained by gauging variations in internal water pressures in earthfill zones during earthquakes.

C. Notable Events

Only a few dams have suffered much damage from major earthquakes, even where there were many reservoirs in the zones of high intensity. A notable exception was Sheffield Dam, an earthfill dam 7.6 m (25 ft) high that failed in the Santa Barbara, CA, quake in June 1925, with a Richter magnitude of 6.3 and an epicenter about 11 km (7 mi) from the damsite. Another example was Lower Van Norman (San Fernando), an earthfill dam 43 m (142 ft) high that was damaged, but survived without reservoir loss, in the 1971 San Fernando earthquake (magnitude 6.6). In both of these cases, the problems were caused by liquefaction of soils, a phenomenon that has been studied extensively in more recent times, with the result that the necessary safeguards have been developed and are widely accepted.

Some of the most significant problems at concrete dams have been related to conditions in the natural sites rather than to the body of the dam itself. Notable cases include St. Francis, Malpasset, and Vaiont Dams.

St. Francis (Fig. 14) was an arched concrete gravity structure 62 m (205 ft) high in San Francisquito Canyon about 72 km (45 mi) north of Los Angeles, CA. It collapsed just before midnight on March 12, 1928. Large blocks of the dam were washed thousands of feet down the canyon. A single monolith was left standing. The basic weaknesses that led to the failure were in the rock foundation, consisting of schist susceptible to sliding on one abutment

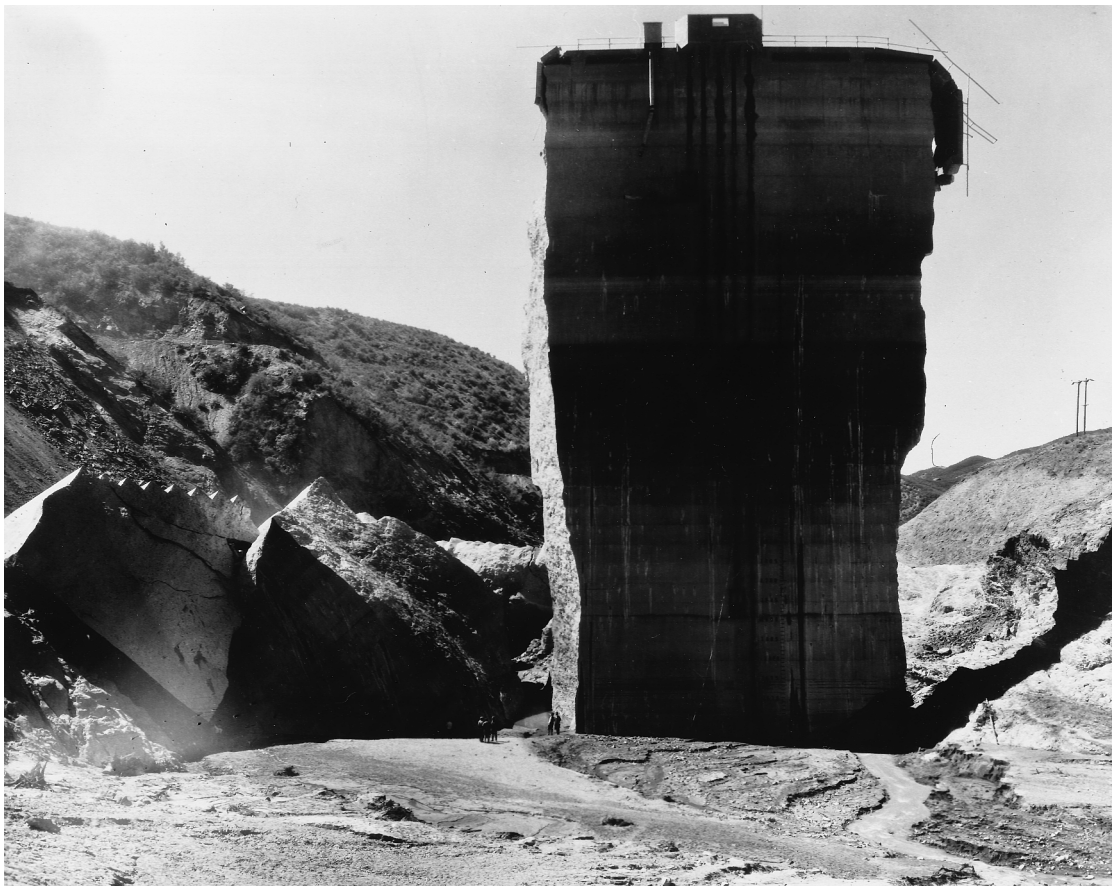


FIGURE 14 The remains of St. Francis Dam after failure in 1928.

and water-softened conglomerate on the other (west) abutment, the two formations joining at a geologic fault.

Malpasset Dam was a concrete arch 61 m (200 ft) high on Le Reyan River in southern France. It broke on December 2, 1959. The break occurred within the foundation, as an abutment slid along a weak seam in the rock and the arch rotated about its other end. After the failure, most of the arch on the sliding side had disappeared along with much of the underlying rock.

Vaiont Dam is a concrete arch 262 m (860 ft) high on Vaiont River near Belluno, Italy. On October 9, 1963, there was a massive slide of a canyon wall into the reservoir, causing overtopping of the dam to a height of about 100 m (330 ft) above the crest. Tremors from the slide were recorded on seismographs in a large part of western and central Europe. Despite the tremendous overload, the arch remained intact, attesting to the inherent strength of this kind of dam. However, the flood wave downstream destroyed the town of Longarone at the junction of the Vaiont and Piave Rivers, where 2600 people died. The reservoir was made inoperable.

Rogun Dam in Tajikistan was expected to be the world's highest dam (335 m). Its site is upstream from Nurek Reservoir on the Vakhsh River. The Rogun design was similar to that of Nurek Dam, with earth core and gravel shells and rockfill facing zones. Construction was progressing when, upon independence from the Soviet Union in 1991, political turmoil and civil war in Tajikistan led to suspension of the project. A large flood in 1993 severely damaged the incomplete works. The future of Rogun Dam is uncertain.

VII. WEIGHING BENEFITS AND DETRIMENTS OF DAMS

The benefits of dams are so interwoven in the fabric of society that some of them may be unrecognized. Others may be known but taken for granted, as are the light switch and the water faucet. Dams are essential to meeting the world's rapidly growing needs for food and energy. The primary benefits of drinking water, irrigation of crops, groundwater recharge, river flow maintenance (including dilution to improve water quality), power generation, flood control, navigation, fire protection, sewage transport, recreation (swimming, boating, fishing, and river rafting), wildlife enhancement, and protection from salinity intrusion engender secondary and tertiary effects including growth of industry, commerce, employment, and local and federal tax bases supporting schools; medical services; and other community needs. In arid regions, reservoirs have improved temperature ranges, increased humidity, and enhanced the environment for migratory birds and other creatures. The advantages of water power as a clean, perpetual resource are evident in comparison with nuclear and fossil-fuel generation. The contributions of hydroelectric powerplants to pollution control can be demonstrated by calculating the volumes of toxic particulates and gases that would be discharged to the atmosphere by alternative coal-fired powerplants of equal capacity. Energy generation from the renewable water source is of global importance in reducing emissions of harmful ozone-depleting chemicals to the atmosphere.

Objective analysis recognizes that dams have both positive and negative effects on fish. The challenge is to mitigate the harm while preserving the advantages. Ideal solutions are elusive. Benefits include the creation of lake fish habitat and the maintenance of downriver flows to sustain migrating fish. Streams without dams also present hazards to fish, including dangers from debris carried by uncontrolled floods and from low water leaving them and their eggs vulnerable to predators. Extensive study has been given to the hardships encountered by salmon in their long journeys, including commercial fishing, the harsh ocean environment, and the degradation of spawning habitat in watersheds by pollution and erosion. Only a very small percentage of the migrants survive the round trip to return to their birthplaces. The decline in the numbers of salmon, as well as the reduced numbers of other species of wildlife, may be attributable to many factors. Herring, an important source of food for salmon, are no longer numerous. Possible causes are disease, predators, shifts in ocean conditions, and increases in industrial pollution. On the Columbia River, where scientists have conducted long-term, exhaustive studies of salmon losses, a major factor has been found to be predation by the many thousands of terns nesting on islands of alluvium dredged from the navigation channels.

The interruption of fish migration by dams is an age-old problem, which engineers have tried to solve by various means, including fish hatcheries, fish ladders, screens, locks, elevators, bypasses, barging, trucking, and pumping. These have been effective at some projects, but there is a continuing need for improved designs. Hatcheries have provided benefits, but the fish produced do not generally survive as well as wild species. There may be substantial delays in passage of fish at dams, whether going upstream or downstream, during which exposure to disease and predators is prolonged. Migrants traveling downstream may be injured severely in passing through spillways and powerplant turbines. At some projects, deflectors have been installed on the faces of spillways to prevent deep plunging of the overflow so that supersaturation of nitrogen and other gases and consequent injury to fish are reduced. Other beneficial measures include scheduling of spillway discharges during downstream migrant passage

and placing screens on penstock intakes, as well as transporting fish around the dams. Some designs provide for interception of fish upstream from reservoirs and directing them through bypass channels to holding facilities to await transport. In the system of the Snake and Columbia Rivers in Washington, this has involved barging over hundreds of miles. Fish detoured around dams may not only escape physical injury but are exposed to less risk from nitrogen supersaturation. The possible effect of transport on the spawning instinct is under continuing study.

Some activists advocate removal of dams on certain rivers to allow fish passage. This option can simultaneously eliminate existing benefits and detriments. Consideration must be given to the infrastructures in the regions dependent upon the dams and their reservoirs. Experience has shown the destruction of a dam to be difficult and costly and not without its own negative effects, including reservoir sediment release. However, in a few unusual cases, demolition has been the selected alternative. In 1999, under a ruling by the Federal Energy Regulatory Commission, the Edwards Dam, a small 162-year-old timber and concrete structure on the Kennebec River in Maine, was breached to allow return of fish from the Atlantic Ocean. In the year 2000, plans were under way to remove dams on the Elwha River in Washington and on forks of Battle Creek in California. Elwha Dam and Glines Canyon Dam on the Elwha, completed in 1913 and 1927, respectively, were not provided with facilities for salmon protection and thus had a major impact on the economy of the local Native American tribe. The cost of removing the dams and restoring the river to its natural condition is projected to exceed $100 million. The several dams to be removed in the Battle Creek watershed are low diversion structures. The restoration plan also includes installation of fish ladders and screens at certain dams that remain, as well as increased stream releases to improve the environment for fish and wildlife. The cost of this plan is expected to be about $50 million. Today, in the development and operation of water projects, full attention must be given to achieving an optimum balance of environmental and economic factors. This demands that dams and other works meet strict standards for preservation of the river and enhancement of conditions in the region.

SEE ALSO THE FOLLOWING ARTICLES

EARTHQUAKE ENGINEERING • GEOENVIRONMENTAL ENGINEERING • HYDROELECTRIC POWER STATIONS • HYDROGEOLOGY • STREAMFLOW • WATER RESOURCE SYSTEMS

BIBLIOGRAPHY

Jansen, R. B. (1983). "Dams and Public Safety," U.S. Govt. Printing Office, Denver, Co.

Jansen, R. B., ed. (1988). "Advanced Dam Engineering for Design, Construction, and Rehabilitation," Van Nostrand-Reinhold, New York.

Jansen, R. B. (1988). "Dam safety in America," *Hydro Rev.*, 7(3), pp. 10–20.

Kollgaard, E. B., and Chadwick, W. L., eds. (1988). "Development of Dam Engineering in the United States," Pergamon Press, New York.

Leonards, G. A., ed. (1987). "Dam Failures," Elsevier, Amsterdam.

National Research Council (1983). "Safety of Existing Dams: Evaluation and Improvement," National Academy Press, Washington, D.C.

National Research Council (1985). "Safety of Dams: Flood and Earthquake Criteria," National Academy Press, Washington, D.C.



Electric and Hybrid-Electric Vehicles

John S. Reuyl
NEVCOR, Inc.

I. Representative Vehicles
II. Basic Engineering Principles
III. Important Policy Initiatives
IV. Electric Vehicle Design Options
V. Energy and Environmental Issues
VI. Analytical Tools

GLOSSARY

BEVs Battery-electric vehicles use a battery to store electricity, an electric motor to provide acceleration and braking torque, and a motor controller to provide moment-to-moment control of torque levels. When compared to conventional vehicles (CVs), BEVs have limited range capability and lengthy recharge times.

HEVs Hybrid-electric vehicles typically have two or more sources of energy and/or power. They use fuel (e.g., gasoline, diesel) for extended range capability and convenient refueling. They combine an engine with an electric motor(s) and electrical energy storage (e.g., battery, ultracapacitor) in a manner that improves fuel efficiency when compared to a CV. Some HEVs also have the option to recharge from the electricity grid.

ZEVs Zero-emissions vehicles were mandated in 1990 by the California Air Resources Board (CARB) in order to meet air quality standards. Until 2000, only BEVs were certified for ZEV credit. More recently, some HEVs have been certified by the CARB for partial ZEV credit.

Specific energy The amount of energy that can be stored in a battery (e.g., watt-hours per kilogram (Whr/kg)). Related to the maximum battery range of a vehicle.

Specific power The amount of power that can be delivered by a battery [e.g., watts per kilogram (W/kg)]. Related to the maximum high-speed acceleration capability of a vehicle.

ELECTRIC AND HYBRID-ELECTRIC propulsion technology is revolutionizing the design of motor vehicles. Battery-electric vehicles (BEVs) compete successfully in certain niche markets such as delivery vehicles, fork-lift trucks, and airport baggage handlers. By comparison, hybrid-electric vehicle (HEV) technologies are increasingly competitive in a wide range of applications, from personal automobiles and SUVs to large trucks and

buses. Unlike BEVs, HEVs typically have an engine in conjunction with electric propulsion components. By integrating these components into a synergistic system, energy efficiency is significantly improved and emissions are reduced; yet HEVs retain the high performance, long range, and convenient refueling that consumers take for granted in conventional vehicles. Some HEVs have the provision of plugging-in for overnight recharging; these vehicles can travel on around-town trips and local commutes on battery power alone, while fuel (for the on-board engine) supplements battery energy for the longer trips. As fuel-cell technology matures, HEVs with fuel cells may displace HEVs with engines, further improving energy efficiency and reducing emissions.

I. REPRESENTATIVE VEHICLES

In 1990 only a limited number of (BEVs) were commercially available, and these vehicles were designed for niche markets, such as electric golf carts, electric fork-lift trucks, and in the UK, electric milk floats.

That year the California Air Resources Board (CARB) undertook a controversial policy initiative, mandating that in 1998, 2% of all offerings in California by the major auto companies must be zero-emissions vehicles (ZEVs), growing to 10% by the year 2003 and beyond. The CARB certified at the time that BEVs would qualify for ZEV credit toward fulfilling a manufacturer's quota. The CARB also ruled at the time that HEVs would not qualify for ZEV credit.

In 1996, the GMEV1 was introduced and between 1997 and 2000, several more BEVs were offered by other major auto manufacturers. The GM EV1 was a purpose-built, from-the-ground-up new design. It emphasized light-weight construction and aerodynamic styling. Other car-makers offered BEVs that were adapted from popular conventional-vehicle platforms such as the Toyota RAV4 EV, Honda EV Plus, Ford Ranger EV pickup, and Daimler-Chrysler EPIC minivan. GM also offered its GM S-10 Electric pickup and Nissan the Altra EV. The specifications of these vehicles are summarized in Table I.

However, in spite of the diversity of models, these BEVs shared three attributes that have limited their marketability: limited range, lengthy recharge time, and high cost. As of the end of the year 2000, roughly 4000 of these vehicles have been sold or leased.

Meanwhile, the federal government had been funding joint development with the auto industry of advanced vehicle propulsion systems that could double the average fuel economy to 55 mpg.

In 1993, the decision was made to form a Partnership for a New Generation of Vehicles (PNGV) between the federal government and GM, Ford, and Chrysler to develop a six-passenger sedan that could triple fuel economy to 80 mpg. The ongoing program for development of advanced (55 mpg) vehicle propulsion systems was folded into PNGV, and the new vehicle platform development to begin under PNGV would further improve vehicle efficiency to the 80-mpg goal.

These two policy initiatives—one by CARB focused upon improving air quality and one by the federal government focused (primarily) upon improving fuel economy—have been key in guiding the development of a broad family of clean, fuel-efficient vehicles. During the years since 1990, the ZEV mandate itself has been modified, in part by the limitations of battery technology and in part by remarkable advances in lower tailpipe emissions of new conventional vehicles. Among the most important changes were (a) the delay until the year 2003 of numerical quotas and (b) the broadening of technologies that would qualify for ZEV credit.

Among these technologies are the broad spectrum of vehicles called hybrid-electric vehicles (HEVs). HEVs are in their essence vehicles with more than one form of propulsion. Most typically, HEVs incorporate forms of electric

TABLE I Representative BEVs

Name	Weight (lb)	Motor (kW)	Battery (V)	Battery range (mi)
GM EV1	3086	100	312, Lead-acid	55–95
GM EV1	2908	100	343, NiMH	75–130
GM S-10 Electric Pickup	4350	85	312, Lead-acid	45–55
GM S-10 Electric Pickup	4200	85	343, NiMH	65–80
Nissan Altra EV	3749	62	400, Li Ion	120
Ford Ranger EV Pickup	4936	67	312, Lead-acid	50
Ford Ranger EV Pickup	4196	67	312, NiMH	65–80
Toyota RAV4-EV	3440	50	288, NiMH	126
Honda EV Plus	3599	49	288, NiMH	105–125
Chrysler EPIC Minivan	N/A	75	336, NiMH	80–90

TABLE II Representative HEVs

Name	Weigh, lb	Engine	Motor (kW)	Battery	Battery range	Mileage (mpg city/hwy)
Toyota Prius	2,765	1.5 liter, gasoline, 4-cylinder	30 kW	274 V, NiMH	Short distance at low speed	52/45
Honda Insight	1,887	1.0 liter, gasoline, 3-cylinder	10 kW	144 V, 1 kWh NiMH	Short distance at low speed	61/68
Humvee	12,200	1.9 liter diesel, 55 kW	2 motors, 200 kW each	288 V, 24 kWh Pb-acid	20–30 mi	NA

propulsion and electrical energy storage, as well as use of fuels. Engines using conventional fuels (e.g., gasoline, diesel) are incorporated in a number of prototype vehicles, and fuel cells are promoted for their superior efficiency and lack of objectionable tailpipe emissions.

In the year 2000, two HEVs were offered for sale in the U.S., the Honda Insight and the Toyota Prius, and by the end of 2000, according to the Electric Vehicle Association of the Americas (EVAA), these two vehicles together comprised over 9000 sales or leases. In Table II, three HEVs already in the market are summarized: Honda Insight, Toyota Prius, and the military HEV version of the High Mobility Multipurpose Wheeled Vehicle (HMMWV, or "Humvee"). These HEVs are representative of the diversity of the current offerings, and their properties will be described in greater detail in this article.

Other HEVs already announced for production include GMC Sierra and Silverado pickup trucks, Dodge ESX3, Dodge Durango TTR SUV (Sport Utility Vehicle), Toyota RAV4, Honda Civic, Honda Spocket, Nissan Tina, Ford Escape SUV, and Mitsubishi SUW (Sport Utility Wagon) Advance.

II. BASIC ENGINEERING PRINCIPLES

A. Diverse Driving Cycles Require Rapid Changes in Vehicle Speed

Vehicles of all types are intended to fulfill fundamental user needs for mobility over a wide variety of driving cycles. The typical urban bus drive cycle is a repetitive sequence of steps—acceleration away from a stop, then a brief period of cruising, followed by a period of hard braking (deceleration) to the next stop. Commuter bus schedules typically include periods of the urban bus drive cycle interspersed with more lengthy periods of highway cruising. In the course of a day, buses may travel several hundred miles.

Delivery vehicles will undertake a somewhat similar stop-and-go schedule with frequent delivery stops and daily travel from 20 to 100 mi or more.

Drive cycles of personal vehicles (e.g., auto, light trucks, SUVs) are much more varied; some vehicles travel most days only short distances, while a few will travel each day several hundred miles. Figure 1 illustrates the distribution of daily travel for personal autos in the U.S.; about 40% of all personal autos on the road on a typical day will travel that day a total of only 15 mi or less, while at the other extreme about 10% will travel 70 miles or more.

Each of these drive cycles reflects itself in periods of acceleration, deceleration, cruising and stop-and-go idling, and these varied demands occur often in rapid succession. The consequences for wheel torque range from brief periods of high acceleration torque, periods of relatively constant low torque during highway cruising, periods of high braking (deceleration) torque, and periods of zero torque when the vehicle is stopped but the engine is still idling.

B. Thermal Engines Provide a Poor Torque/Speed Match for Typical Driving Cycles

These different drive cycles constitute, in general, a poor match for conventional engines. Figure 2 depicts the relationship between maximum (full-throttle) torque and engine speed for a representative internal combustion engine (ICE). Note that maximum torque for this engine occurs at about 4000 rpm.

The theoretical peak efficiency with which engines can convert the energy in the fuel to torque occurs in a small zone in the torque/speed map called by engine designers the engine's "sweet spot." Outside this sweet spot, efficiencies fall. While the four-cycle ICE tends to maintain reasonably high efficiencies away from the sweet spot, other engines (e.g., turbines) show a very sharply defined sweet spot, with efficiencies falling rapidly as the engine is operated at part load.

Today's modern four-speed automatic transmissions attempt to match the rapidly changing requirements for road torque with the engine's torque capability, selecting the gear that enables the engine to operate at the most efficient speed for any torque requirement.

However, this match is seldom optimum, and in instances of braking and engine idling, there is no match at all.

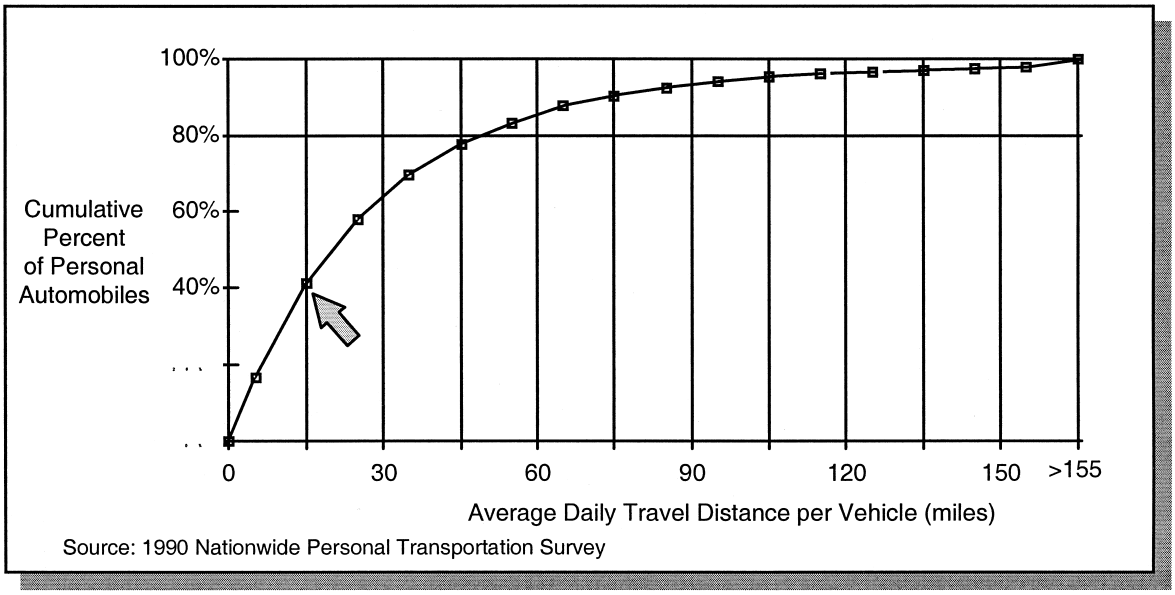


FIGURE 1 Distribution of personal automobile use.

As a result, in a conventional vehicle the actual efficiency of converting fuel to useful torque on any given driving cycle will vary from occasional peak efficiency to zero efficiency, and average efficiencies may be half or less of the peak efficiency.

C. Electric Motors/Controllers Are Exceptionally Well-Suited for Vehicle Powertrains

1. Electric Motors

The rapidly changing demands for road torque can be met with exceptional efficiency by electric motors. High torque can be provided over a wide range of motor speeds, from stall to more than 10,000 rpm. Efficiency over a wide range of speeds is typically >90%.

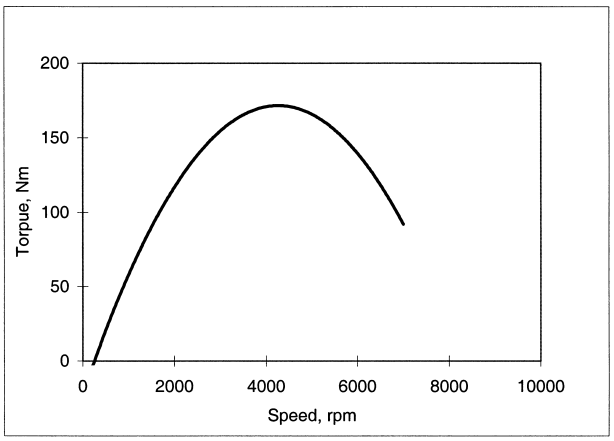


FIGURE 2 Typical torque/speed map for an ICE.

As a result, multispeed transmissions are not needed for many applications; a single speed gear reduction into a standard differential gear is sufficient to match engine speed to wheel speed. When the vehicle is stopped, no power is provided to the motor, unlike conventional vehicles, whose engines continue to run, albeit at idle speeds.

In some applications, individual electric motors are fitted into each wheel, thereby providing to each wheel individual torque control for both acceleration and braking. In other applications, the wheel motors can be mounted to the vehicle body with torque shafts to the wheel. In this configuration, the motor is part of the sprung weight of the vehicle, which typically improves ride and handling.

A number of motor technologies are under development for vehicle propulsion, including direct current (DC), three-phase alternating current (AC) induction, permanent magnet (PM), and variable reluctance motors (VRM). Each has a unique set of properties, including cost, thereby suiting them to particular applications. However, they all tend to share the basic advantages when compared to conventional engines: torque curve well-matched to the demands of motor vehicles, high efficiency over a wide range of motor speeds and torque demands, high power-weight ratio, long life, low maintenance requirements, and low cost.

2. Motor Controllers

While efficient electric motors have been used in specialized vehicle applications for decades, the development of solid-state electronic controllers has been the key to making electric propulsion feasible for many new applications. Such controllers provide precise control on an

instantaneous basis with exceptionally high efficiency (typically >95%). Electronic power controllers have dropped steadily in cost as development has advanced.

In the typical electric vehicle, the controller provides the interface between the motor and battery pack, and the driver provides input signals to the controller with the accelerator and brake pedals. Energy is stored in the battery pack at a relatively high voltage (typically 100–400 V, depending upon the application), and voltage is applied to the motor by the controller at a level sufficient to achieve the desired instantaneous torque. The power out of the batteries will be equal to the power into the motor plus controller losses. The motor torque is proportional to motor current and independent of motor speed. Thus, for maximum acceleration torque, the controller will provide a constant maximum current to the motor. However, as the vehicle speed (and motor speed) increase, the motor voltage will rise, and the motor power will rise according to Eq. (1):

$$P_{(\text{motor})} = E_{(\text{motor})} I_{(\text{motor})} \tag{1}$$

At some point, the motor voltage will rise to the level of the battery pack itself, and at this point the maximum power limits of the battery pack will be met. The vehicle can continue to accelerate to higher speeds, but the acceleration will now be at constant power, not constant torque. Hence, as vehicle (and motor) speed continue to increase, motor current (and torque) will decrease. These conditions are illustrated in Fig. 3.

3. Regenerative Braking

Electric motors can also act as generators, thereby providing regenerative braking. This braking energy can be stored in the vehicle's electrical energy storage system, thereby improving overall vehicle energy efficiency. Regenerative braking also extends the life of conventional brakes and can extend the life of related equipment. For example, braking energy in conventional buses can result in premature failure of the tire bead, due to excessive heat that is transferred from the brake drums to the wheel. Regenerative braking can reduce this heat buildup, thereby extending tire life as well as brake life.

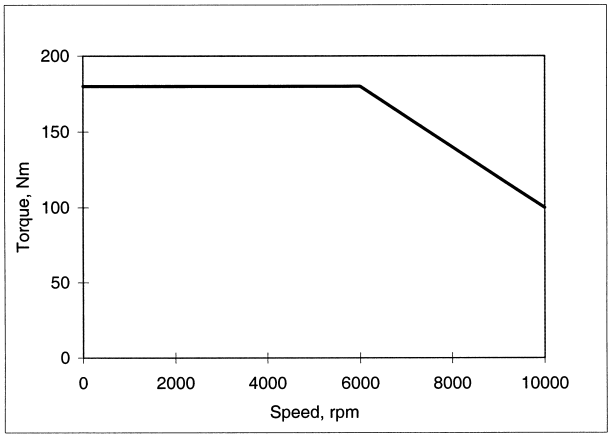


FIGURE 3 Typical torque/speed map for an electric motor.

4. Charging from the Grid

The vehicle motor and controller can also be integral parts of a charging system, enabling the vehicle to be charged from the electricity grid. By utilizing the motor inductance and the power electronics of the controller, the costs for an on-board charger can be low enough to permit simple plug-in recharging from any available receptacle. This capability can be especially suited to the overnight charging of HEVs and BEVs.

Note from Fig. 4 that overnight charging of personal vehicles from any 110 V receptacle could provide up to 50 mi of travel in a typical BEV. Recharging from a residential 220 V circuit would provide the same energy in 2.5 hr. The electricity infrastructure with 110 V receptacles is widely available and relatively convenient.

However, it becomes apparent from Fig. 4 that recharging rapidly for long trips is not as rapid as refueling a conventional vehicle, even with dedicated "rapid recharging" stations. In a conventional vehicle, refueling takes about 5 min, and a full tank of fuel will typically provide 400 mi of travel. Note that at "Level 3" recharging, a 10-min recharge of a typical electric vehicle would provide roughly 40 mi of travel. In addition, the infrastructure required to support a network of such stations is not in place.

D. Energy Storage Options

1. Fuels

When compared to batteries, fuels offer unmatched specific energy. In the 68-mpg Honda Insight, a 100-mi trip would take about 1.5 gal (roughly 10 lbs) of gasoline. By comparison, the GM EV1 with the 1147-lb NiMH advanced battery pack provides only the same (75–110) mi of travel.

2. Batteries

Sustained efforts have been made, especially since 1990, to develop advanced batteries that would provide BEVs with longer range than the baseline lead-acid battery. However, these efforts have had only limited success, and the costs of these new technologies have been greater than the lead-acid technology.

Batteries offer a viable means with which to store electrical energy for short periods of time and to provide

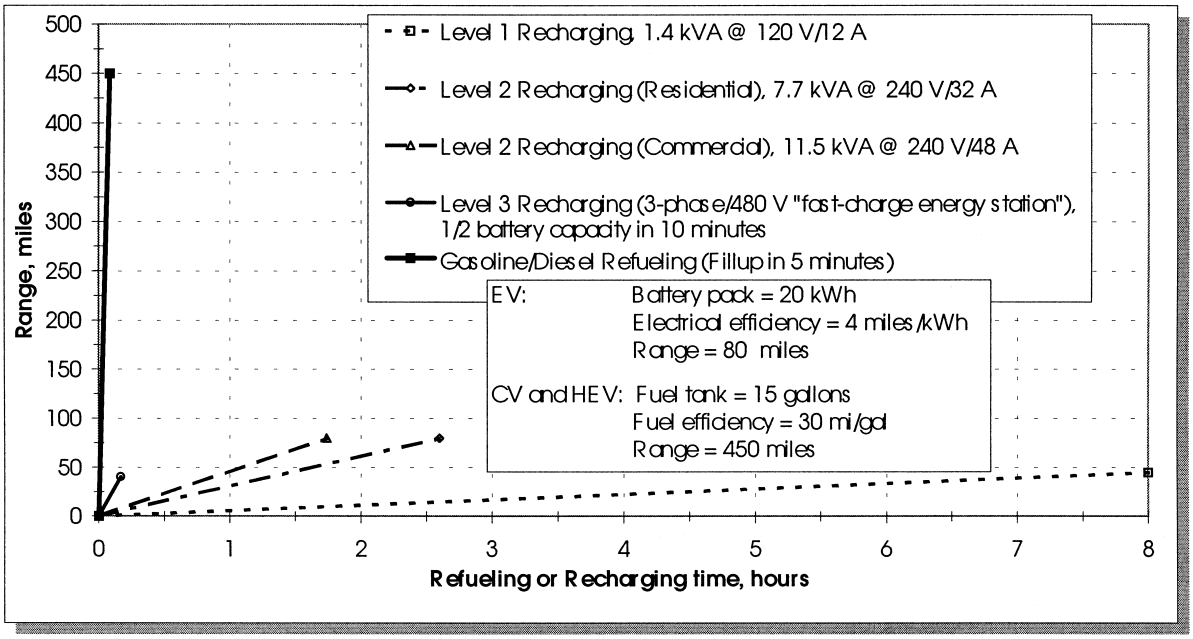


FIGURE 4 Comparison of range and refueling/recharging times for CVs, EVs, and HEVs.

modest battery-only range for HEVs. Batteries must store sufficient energy (Whr/kg) while at the same time deliver that energy at sufficiently high power levels (W/kg).

The specific energy of a battery is determined initially by the battery chemistry. The electrochemical couple (e.g., lead-acid) provides the baseline capability. NiMH batteries have roughly double the specific energy of lead-acid batteries; Lithium-based batteries are typically four times higher in specific energy than the lead-acid battery.

The specific power of a battery (W/kg) determines how rapidly the battery can deliver the stored energy into a load. The specific power of a battery is determined largely by the design of the battery. For example, increasing the effective surface area upon which electrochemical reactions can take place will increase the rate at which current can be discharged. Similarly, lowering the resistance of current-carrying conductors in the battery will reduce internal losses and increase the discharge rate.

The same issues must be addressed in order to determine the rate at which the battery can be charged.

Both the charge and discharge rate become important as the size of the battery is reduced. For example, a 500-kg battery in a typical BEV (like the GM EV1) that must provide a peak power of 100 kW must have a specific power capability of 200 W/kg. By comparison, a smaller 100-kg pack would need a specific power capability of 1000 W/kg in order to provide the same 100 kW of power.

The specific energy and specific power of any given battery typically relate to one another as presented in the ragone plot in Fig. 5. This figure plots specific power (W/kg) on the *Y* axis and specific energy (Wh/kg) on the *X* axis. Note that both axes are logarithmic scales.

Figure 5 depicts the relationship for two representative lead-acid batteries. Battery A is typical of lead-acid batteries used in BEV applications. When power is delivered at a rate of 20 W/kg, the battery provides 40 Wh/kg of usable energy. However, when power is delivered at 200 W/kg, the specific energy of the battery has dropped to 10 Wh/kg. The vehicle's range at any given power level is determined, in part, by this relationship. Thus, the vehicle's range at a power level of 200 W/kg would be only one-fourth of

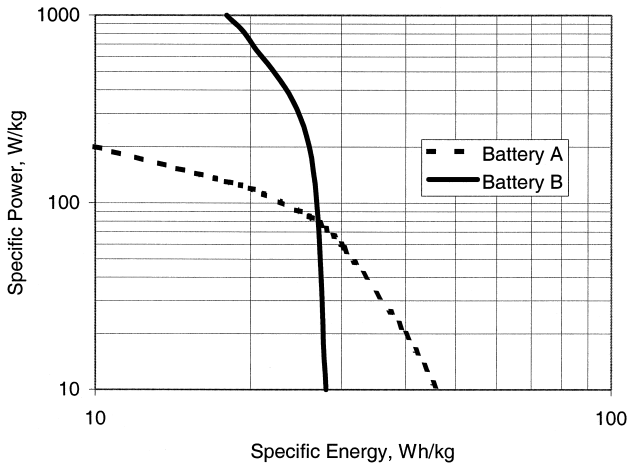


FIGURE 5 Ragone plot of two representative lead-acid batteries.

the range at 20 W/kg (assuming no change in the vehicle's energy efficiency).

Assume that the vehicle in the example above has a 500-kg Battery A, and further assume that it requires 10 kW to maintain 60 mph on the level (i.e., 20 W/kg). From Fig. 5, the battery pack's total energy can be calculated by multiplying its specific energy (40 Wh/kg) times its total weight (500 kg) to yield its total energy (20 kWh). Finally, assume that the vehicle's energy efficiency at this speed is 5 mi/kWh. The vehicle's range under these conditions can be calculated to be the total battery pack energy (20 kWh) times the vehicle efficiency (5 mi/kWh) or 100 mi.

Note from Fig. 5 that if the average power to be drawn from the pack is greater than 80 W/kg, then Battery B would provide the greatest specific energy. For example, if the average power is 200 W/kg, Battery B would provide approximately 24 Wh/kg, whereas Battery A would provide only 10 Wh/kg.

Such high average power levels could be typical for an HEV application. Assume the same vehicle above except that the Battery B pack size is 100 kg instead of 500 kg. The power used to maintain 60 mph is assumed to be the same (10 kW), and the specific power is now determined to be 100 W/kg. The pack's total usable energy at 60 mph from Fig. 5 is 25 Wh/kg times 100 kg or 2.5 kWh, and the range at an efficiency of 5 mi/kWh would be 13.5 mi.

3. Advanced Storage Devices

To compensate for the limitations in specific power of electrochemical batteries, capacitors and ultracapacitors have been developed that offer dramatically higher specific power, but at a substantially lower specific energy. However, the integration of these devices with a battery pack into an integrated energy storage system could, in theory, provide the moment-by-moment high power for acceleration or absorb the high power produced during regenerative braking, thereby cushioning the battery from these high power pulses.

In vehicle designs where little or no battery range is required, such high-power devices could be used alone without the electrochemical battery. In such HEV designs, the engine would be buffered only from the most rapid torque transients, and only limited deceleration (braking) energy would be recaptured.

Electromechanical flywheels have been tested in prototype vehicles as a substitute for battery packs. In theory, flywheels could offer very long cycle-life, and they could provide the high specific power needed for acceleration and braking pulses. However, in addition to cost barriers, flywheels pose new technical challenges in mobile applications that have yet to be mastered.

E. Many Applications, Including Personal Vehicles, Are Well-Suited to Battery-Electric Propulsion

A great many vehicle duty cycles can be met with battery-electric propulsion. Examples include postal vehicles, light-duty delivery vans, and fork-lift trucks.

Even more surprising, the personal automobile will travel on a typical day relatively short distances. Note from Fig. 1 that in the U.S. on a typical day, 40% of all personal autos on the road on that day will travel a total of 15 mi or less. Such distances are well within the capability of today's BEV.

The difficulty is that almost all of those same vehicles will also travel on some days very long distances, such as weekend trips, vacations, or even on unexpected emergency trips during an otherwise typical day. The battery-only vehicle, with its limited range and lengthy recharge time, makes travel on these longer trips less practical and more difficult to plan and execute. To the typical driver, used to a "full-tank" feeling of security and convenient refueling at any service station, BEVs are neither as convenient nor as secure. Even a BEV with a fully charged 100-mi battery is like starting the trip with the gas gauge on 3/4 empty, and there's no place to "refuel" until the vehicle returns at the end of the trip.

The HEV, on the other hand, could offer practical electric propulsion for both the short and long trips—an "unlimited range EV." Shorter-around town trips and local commutes could be driven on battery power alone. Overnight charging from any available 110 V receptacle could provide up to 50 mi of travel the next day. For many vehicles, the engine may not be needed for days at a time. The engine would function somewhat like a second-spare tire; there if you need it, but rarely used.

For long trips, fuels offer the superior energy storage medium, and the refueling infrastructure already in place assures convenient, quick refueling. Such an HEV would provide the "full-tank" feeling of security that drivers of conventional vehicles take for granted. Even on longer trips, the initial miles could be on battery power, and the engine could supplement the battery power for those miles beyond battery-only range.

III. IMPORTANT POLICY INITIATIVES

Two initiatives during the 1990s have played major roles in guiding the development of BEVs and HEVs. These are

1. The mandate in 1990 by the California Air Resources Board for ZEVs by 1998
2. The public/private Partnership for a New Generation Vehicle (PNGV) between the U.S. Department of

Energy and GM, Ford, and Chrysler to develop an 80-mpg six-passenger automobile

A. ZEVs Were Mandated by California Regulators in 1990

In 1990 the California Air Resources Board launched the ZEV mandate, requiring that 2% of all motor vehicles sold in California in 1998 be ZEVs, growing to 10% by the year 2003. At the time, only BEVs would qualify for ZEV credit, and HEVs were specifically excluded from qualifying.

The program has resulted in the development of BEV prototypes, but there also have been sustained challenges to the program. In addition, independent studies in the mid- and late-1990s concluded that acceptable batteries were not yet available.

As a consequence, the program has gone through changes during the biennial reviews. The numerical quotas for 1998–2003 were dropped in favor of memoranda of agreement (MOA) with the affected auto makers. BEVs have been offered for sale or lease, but the number of vehicles has been far less than hoped for by the regulators.

B. PNGV Was Launched in 1993 to Develop 80-mpg Six-Passenger HEVs

Like BEVs, HEVs languished in development until the CARB focused. political attention upon the issue of air quality. However, CARB explicitly excluded HEVs as qualifying for ZEV credit for the first 10 years of the ZEV program, in spite of protests to this decision.

However, during this same time period, the federal government was focusing its attention on developing advanced vehicle propulsion systems that could double fuel efficiency to 55 mpg, and HEV technology was key to achieving this goal. It was a national priority to reduce the nation's dependence upon imported petroleum; the nation's economy was closely linked to unfettered transportation, and a healthy transportation sector was dependent upon a reliable supply of fuel at a stable price.

In 1993, the public/private PNGV between the U.S. government and GM, Ford, and Chrysler was launched. Its major goal was the development of a six-passenger automobile that could achieve 80 mpg. Low exhaust emissions were also a goal, but the targets for ozone-precursor emissions were not as tight as the CARB ZEV regulations.

An important part of the PNGV program was the joint effort already underway to develop advanced vehicle propulsion systems. This ongoing program was folded into PNGV, and the new broader effort would also develop new, lighter, more aerodynamic platforms, in hopes of boosting the target vehicle's fuel efficiency to 80 mpg.

C. In 2000 Certain HEVs Were Certified for ZEV Credit

In 2000 the HEV technology was recognized as worthy, in certain circumstances, of ZEV credit. The CARB program also acknowledged that exhaust emissions from vehicles with engines had been reduced to levels that were not thought possible when the program was adopted in 1990.

This change in policy reflected the fact that HEVs offered low-emissions benefits similar to BEVs and HEVs appear to be much more marketable.

The ZEV regulations are likely to undergo further changes as new HEVs are produced and submitted for certification. The time is approaching when BEVs and HEVs with ultraclean combustion engines or fuel cells may successfully remove new-vehicle exhaust emissions from the air-quality equation.

IV. ELECTRIC VEHICLE DESIGN OPTIONS

A. BEV

The basic BEV incorporates an electric motor connected to a differential gear by a single-speed gear box, the motor controller, battery and associated accessories including the battery management system.

BEVs have been developed for a variety of applications, from specialty vehicles (fork-lift trucks, airport baggage handlers) to autos, buses, and trucks. The technology is most competitive in applications that place high value on quiet operation, low operations, and maintenance costs and zero emissions at the vehicle.

BEVs can provide excellent performance (acceleration, handling), and the operations and maintenance costs (other than battery maintenance and replacement) are very competitive with conventional vehicles. As an example, the GM EV1 has been widely praised by drivers for its impressive acceleration and quiet operation.

However, it is also necessary that the vehicle meet all of the expectations of the particular duty cycle. This requirement has been the major obstacle to the broad acceptance of BEVs as personal automobiles; the limited range and lengthy recharge time of BEVs have limited the appeal of electric autos. As a result, battery research has focused upon batteries with higher specific energy (for longer range). However, these batteries have been even more expensive than the baseline lead-acid technology, thereby contributing to an even higher cost penalty.

B. Battery–Battery HEV

Hybrids have been tested wherein two different battery technologies are combined in order to better meet the

disparate demands for high power and long range. While technically an HEV, these vehicles have attempted to offset to some extent the limitations of BEVs as discussed in the preceding section.

C. Series HEV

The simplest series HEV could incorporate a small engine/generator into what otherwise would be a BEV. Such an engine/generator could supplement the energy from the battery, thereby extending the range beyond the battery-only range. Trade-offs in sizing individual systems could be made, depending upon each particular application. If daily trips were always likely to be 50 miles or less and longer trips were likely to be infrequent, a battery pack sized for the 50-mi trips and a small engine/generator might be optimum. On the other hand, if daily trips were likely to be 20 miles or less, but longer trips were frequent, a smaller battery pack and larger engine/generator may be a better match.

The PEI HUMVEE is a series HEV with two motors, one for each axle. The vehicle is based upon the HMMWV platform and has a 20- to 30-mi range on battery power alone; a diesel-powered engine-generator supplements the battery energy for longer trips. It has been developed for military applications, but like the conventional HUMVEE, the HEV version is likely to be offered in a civilian version.

D. Parallel HEV

The parallel HEV typically makes use of the engine on all trips, and the electric propulsion system is integrated so as to achieve maximum fuel efficiency. The Honda Insight is typical of one configuration. An electric motor is mounted so that its torque can be added to that of the ICE, and the performance of the vehicle is similar to that of a vehicle with a much larger, but less fuel-efficient, engine. The electric motor provides an assist to the ICE, boosting acceleration for short passing maneuvers and providing regenerative braking. In addition, the engine can be shut down during deceleration and at stops and restarted by the electric motor when the need for engine torque returns. These features combine to achieve impressive fuel efficiency (see Table II).

Another parallel configuration places an engine and transmission on one axle and an electric motor on the other axle. Dodge incorporates this scheme for the upcoming HEV version of the Durango SUV. A computer provides the blending of torque from both drive systems. Under some driving conditions, it would be possible to drive in all-electric mode, but the battery-storage system would need to have sufficient capacity for this option.

E. Series/Parallel HEV

The Toyota Prius permits operation in both series and parallel modes. The Prius is among those HEVs that incorporate a planetary gear set to which the drive shaft, electric motor, engine, and generator are all connected. During certain modes, the vehicle can operate by using only an electric motor for propulsion and braking, and the engine remains shut down. During other modes, the planetary gear set functions as a continuously variable transmission, enabling the engine to provide power directly to the driveshaft while at the same time maintaining engine speed at ideal operating speed for the desired amount of torque.

The operation of a representative HEV using a planetary gear set as an infinitely variable transmission is shown in Fig. 6. In this particular configuration, the engine is connected to the shaft of the planetary carrier gear, the electric motor is connected to the shaft of the ring gear, and the starter/generator is connected to the shaft of the sun gear. The ring gear is also connected to the drive axle through a reduction gear and differential gear.

The three vertical lines in Fig. 6 represent the shafts in the planetary gear set with shaft speeds equal to zero at the bottom of the figure and shaft speeds increasing as one moves up vertically. Once the size of the sun and ring gear have been chosen, the planetary gear set with those gears is a fixed-ratio gear set, and the speeds of each shaft are fixed relative to one another. This relationship is shown in Fig. 6 for four representative conditions as heavy straight lines. Given the speeds of any two shafts, the speed of the third shaft is uniquely determined by the straight line that passes through all three points. For example, given the motor rpm at point 7 on line D and the engine rpm at point 8, the generator rpm must be point 9.

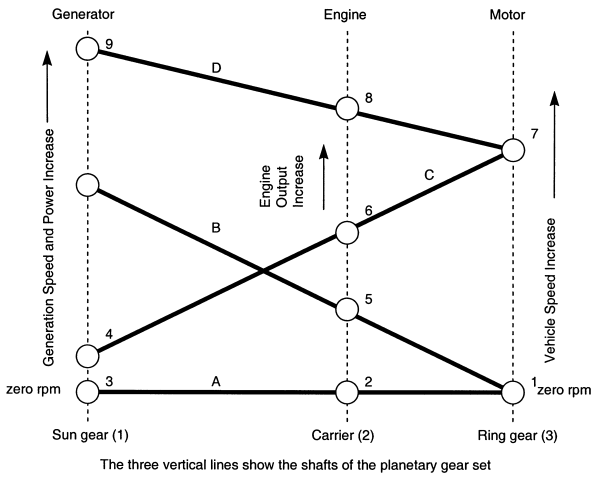


FIGURE 6 HEV using a planetary gear set as an infinitely variable transmission.

Torque and power must always balance such that

$$T_1 + T_2 + T_3 = 0 \quad (2)$$

$$P_1 + P_2 + P_3 = T_1\omega_1 + T_2\omega_2 + T_3\omega_3 = 0 \quad (3)$$

The discussion that follows will focus upon five representative operating modes:

1. When the Vehicle Is at a Standstill

The horizontal line marked A that includes points 1–3 represents the stopped condition; the engine is shut down, the motor is stopped, and therefore the generator is stopped. Torques and powers on all three shafts [see Eqs. (2) and (3)] are zero.

2. Initial Acceleration and Low Speed Travel

The electric motor provides propulsion torque for the vehicle at low speeds. Assume that the vehicle accelerates from a stop. In Fig. 6, the effect would be to pivot line A at point 3, allowing point 1 to move up the motor rpm axis. The electric motor on the ring gear provides torque directly to the ring gear shaft, not through the planetary set. The vehicle moves forward without any torque or power transfer through the planetary gear set. Losses in the motor are largely resistive (I^2R) and are therefore proportional to T^2, and losses in the motor controller are roughly proportional to power (EI).

A clutch permits the engine to remain stopped. Alternatively, the generator can reverse direction, with operation below position 3 in Fig. 6; the engine then could remain at position 2 until it is started.

When more torque is required or when a threshold vehicle speed has been reached, the generator acts as a starter to crank the engine. After the engine starts up, torque from the engine can be added via the planets carrier shaft to the output (ring gear) shaft, with a balancing torque from the generator. Line B represents the condition when the vehicle is still stopped, i.e., point 1 on the ring gear is at 0 rpm. Were the vehicle moving forward at low speed, point 1 on Line B would be positioned a short distance up the ring gear rpm axis.

Engine losses would be determined in the customary manner. With any given engine and generator torque into the planet carrier, losses in the planetary gears must be added to losses in the generator and motor to determine total losses.

3. Cruising

Power required to maintain constant highway speed is relatively low. The motor rpm is at point 7 on the C line, and engine speed is at point 6 (relatively low speed and low torque), requiring generator rpm to be at point 4 (very low speed and low torque). The engine provides torque via the planet carrier to the ring gear (and to the road), and this torque is balanced by the generator torque on the sun gear.

At any generator torque level, the generator speed (sun gear) can controlled over a wide range. Thus, for any given generator torque, the speed of the generator can be controlled to vary so that the engine (on the planet carrier) is always operating at the most efficient speed to produce its required torque. The resulting power from the generator is routed to the electric motor (ring gear), and the battery is in a float condition. In theory, the generator could provide the necessary balance torque at very low speeds (line C in Fig. 6), thereby allowing the engine to operate at low rpm.

4. Hill-Climbing or Acceleration

Under increased torque demands (e.g., for hill climbing), the engine provides increased torque, balanced by the torque of the generator. The generator rpm is controlled so that the engine rpm is optimal for the given engine torque; the operating mode changes smoothly from line C to line D in Fig. 6. Under full throttle, battery energy is used to further increase electric motor torque. Losses in the battery (charge and discharge) need to be added to motor and generator and gear losses to determine total losses.

5. Deceleration (Regenerative Braking System)

During braking, the electric motor on the ring gear is used as a generator, and the energy is used to recharge the battery.

Since no acceleration torque is required from the engine, the engine can be shut down and no power is transmitted through the transmission.

F. HEVs Can Integrate Energy-Efficient Electric Propulsion with Fuel-Efficient Engines

Initially electricity was the dominant "fuel" for vehicles early in the twentieth century. Electric vehicles were favored for their quiet, clean operation.

Then, with the invention of the electric starter, vehicles with engines became more practical, and petroleum-based fuels became dominant.

Now, with HEVs, a combination of electric propulsion and fuel as an energy source may be optimum.

The four-cycle spark-ignited internal combustion engine, coupled with advanced multispeed transmissions, has been refined over the decades to meet the demands placed upon modern motor vehicles. It is natural that this engine is now being adapted to the new HEV applications.

However, other engine concepts that were not well suited to conventional propulsion are now being reconsidered for HEV applications. By reducing or eliminating the need for rapid throttle response, HEVs remove one of the most challenging barriers to engine design. For example, when turbine engines can be designed to operate at essentially their ideal operating conditions at rated power, their efficiencies can be considerably higher than the four-cycle engine operating under its optimum conditions. Diesel engines also show promise both in inherent efficiency and in recent developments in emissions control systems. Each of the following engine types has attributes that may suit it well to particular HEV applications.

1. Otto (four-cycle) spark-ignition internal combustion engine (ICE)
2. Miller cycle (high-efficiency ICE)
3. Two-cycle spark-ignition ICE
4. Diesel compression-ignition
5. Brayton (turbine)
6. Stirling (continuous external combustion)

G. Fuel Cells Promise to Improve Fuel Efficiency Beyond That Possible with Engines

Fuel cells are considered a longer term replacement for the thermal engine in HEVs. Considerable effort is now underway for the development of fuel cells for both stationary and mobile applications. Fuel cells can operate at high efficiencies with no ozone-precursor exhaust emissions. However, fuel cells raise a number of technical issues regarding the source of the hydrogen fuel, and these issues must be addressed in any comprehensive systems assessment of energy and environmental impacts. Fuel cells also face cost barriers, and it is likely that stationary applications will become commercially viable first.

V. ENERGY AND ENVIRONMENTAL ISSUES

A. Energy Sources

1. Petroleum

Petroleum-based fuels have been uniquely suited as fuels for transportation, particularly for on- and off-road vehicles. The amount of energy stored in a pound of fuel is far greater than that in any battery technology. The fuels infrastructure that has been developed over the decades means that, in the developed nations, fuels are available wherever they are needed.

However, huge sums are spent annually to protect the petroleum supply infrastructure. Developed nations are increasingly dependent upon imported fuels, and energy security and economic stability are closely linked.

Therefore, it has been national policy to reduce the need for imported petroleum and thereby improve energy security, increase economic growth, and reduce international political risks.

2. Electricity

BEVs have been proposed as a way of displacing fuel use with grid electricity. The majority of electricity is produced from nonpetroleum sources, and the majority of these sources are domestic.

HEVs that could plug-in at night and that had as little as 25-mi battery range could reduce petroleum use by 50% (see Fig. 7). Short around-town trips may require no use of fuels whatsoever. On longer trips, the initial miles traveled each day could be on battery power, and fuels could extend the range as necessary. For this reason, HEVs with a given battery range actually reduce fuel use by a larger amount than an equivalent BEV.

The first HEVs from the major auto makers (the Honda Insight and the Toyota Prius) do not have the provision for recharging from the grid. Instead, they are designed to maximize fuel efficiency. However, it is relatively simple to incorporate a grid connection, and users may well wish to choose such a charge capability as an option.

3. Displacing petroleum with Renewable Electricity

When BEVs and HEVs use grid electricity instead of fuels, the effect in terms of public policy is to reduce petroleum fuel use. Just as one can reduce fuel use through improved fuel efficiency, BEV and HEV technology can reduce fuel use by using grid electricity instead of fuels.

It is also public policy to move increasingly from fossil-based energy supplies to renewable energy supplies, and transportation can be a part of the transition. Once vehicles can be charged using grid electricity, that electricity can be from renewable sources.

It is straightforward enough to envision renewable-energy systems such as wind generators recharging vehicles at night when loads are low.

It is also possible to see daytime solar-power generators as providing a transportation fuel. For example, every kWhr of solar photovoltaic electricity generated during the day could delay the generation of a like amount of hydroelectric electricity until later that night. In this manner, solar photovoltaic electricity could displace petroleum use in vehicles that can plug in at night.

Such a scheme is much more energy efficient on a per-acre basis than, for example, renewable biomass fuels; the conversion efficiency of solar radiation to PV electricity

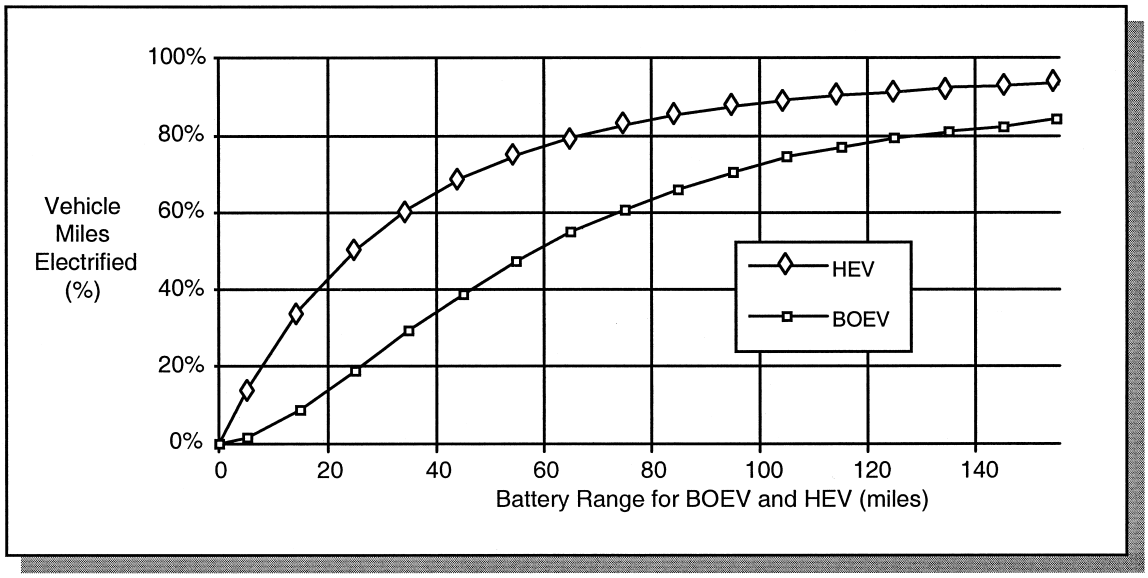


FIGURE 7 Personal automobile miles electrified.

is typically 10%, while the conversion efficiency of solar radiation to biomass (through photosynthesis) is more like 0.1%.

B. Carbon Dioxide (CO_2) Emissions

Global climate change is increasingly viewed as linked, in part, to the increasing global concentrations of carbon dioxide (CO_2) in the atmosphere. Carbon dioxide CO_2 is a natural by-product of the combustion of carbon-based fuels, including transportation fuels in engines. CO_2 is one of the "greenhouse" gases, transparent to incoming solar radiation but relatively opaque to infrared radiation from the earth back into space. As a result, increasing CO_2 levels act to block increasing amounts of heat energy from being radiated back to space, while incoming solar radiation levels are unchanged. The result is a gradual rise in the earth's temperature.

Limits to CO_2 emissions can be achieved most directly by limits to the use of carbon-based fuels. Thus, the improvements in fuel efficiency of new HEV designs result in a corresponding one-to-one reduction in CO_2 emissions.

Improvements in energy efficiency of BEVs and HEVs can also reduce CO_2 emissions by enabling non carbon energy sources, such as renewable energy, to displace use of carbon-based fuels.

C. Ozone-Precursor (VOCs, NO_x) Emissions

1. Role of VOCs and NO_x

Troposphere (low-level) ozone is formed when volatile organic compounds (VOCs) and nitrogen oxides (NO_x) react in the presence of heat and sunlight.

According to the National Research Council (NRC), "Of the six major air pollutants for which National Ambient Air Quality Standards (NAAQS) have been designated under the Clean Air Act, the most pervasive problem continues to be [troposphere (i.e., near-ground)] ozone, the most prevalent photochemical oxidant and an important component of 'smog'." As the NRC makes clear, NO_x control is increasingly understood to be the key to the control of urban and regional troposphere ozone.

Volatile organic compounds [also known as reactive organic gases (ROG)] occur naturally in nature (biogenic sources, mainly vegetation) as well as from a variety of anthropogenic sources (i.e., actions of human society such as automobile exhaust). NO_x, on the other hand, results almost entirely from anthropogenic sources (high-temperature combustion including automobile engines and electricity generating plants). "The Committee on Tropospheric Ozone Formation and Measurement was established by the National Research Council to evaluate scientific information relevant to precursors and tropospheric formation of ozone .... The Committee was specifically charged to address emissions of volatile organic compounds (anthropogenic and biogenic) and oxides of nitrogen ...."

On the central finding of VOC versus NO_x control, the NRC report states that "NO_x control is necessary for effective reduction of ozone in many areas of the United States .... in many urban cores and their environs, even if anthropogenic VOC emissions are totally eliminated, a high background concentration of reactive VOCs will remain .... In the presence of anthropogenic NO_x and under favorable meteorological conditions, these background biogenic VOCs can contribute to summertime ozone

concentrations exceeding the NAAQS [National Ambient Air Quality Standard] concentration of 120 ppb."

2. Exhaust Emissions

Exhaust emissions of NO_x from modern vehicles can be controlled to exceptionally low levels. For example, the 1998 Mercury Grand Marquis (a full-size six-passenger sedan with a large V-8 engine) was tested by the California Air Resources Board to have NO_x emissions of 0.004 g/mi. Such levels were well below the ULEV standard of 0.2 g/mi, the most stringent NO_x standard then in place through the 2003 model year. In fact, the 1998 Mercury Grand Marquis NO_x emissions are below even the SULEV standard of 0.02 g/mi, a standard proposed in 1998 to take effect beginning with the 2004 model year. Even then, only 5% of the offerings of any manufacturer need meet this standard.

As discussed in the next section, such exhaust emissions of NO_x are so low that the "fuel-cycle" NO_x emissions that occur upsteam of the vehicle during fuel production and distribution dominate total emissions (exhaust plus fuel cycle).

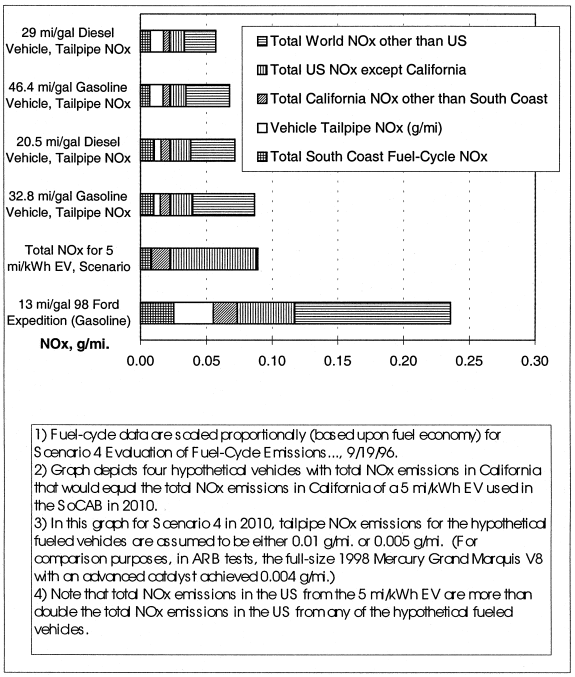


FIGURE 8 Total NO_x emissions of vehicles operated in the South-Coast Air Basin (SoCAB).

3. Fuel-Cycle Emissions

Fuel-cycle emissions occur during the production of fuels including drilling, pumping, refining and eventual distribution of fuels to the corner service station. These fuel-cycle emissions can be apportioned to the vehicles that use them on a per-mile basis. Thus, fuel-cycle emissions are directly proportional to the vehicle's fuel economy—for example, doubling fuel economy would reduce fuel-cycle emissions by 50%.

Under a study funded by the California Air Resources Board, the emissions associated with the production of a variety of fuels, including gasoline, diesel, and electricity, were assessed. In all cases, the fuels were assumed to be used in the Southern California Air Basin (SoCAB), a region that includes the greater Los Angeles area. The fuels were assumed to be used in the year 2010. In the scenario 4 data, it was assumed that all equipment met 2010 standards, and several additional efficiency and environmental improvements were assumed, such as lower distribution losses and lower fuel spillage.

These emissions data from the CARB study were used during a study funded by the National Renewable Energy Laboratory (NREL) to determine the fuel-cycle emissions for six representative vehicles. The results are shown in Fig. 8.

At the bottom of the figure, the Ford Expedition, a large 6000-lb SUV is shown to have total NO_x emissions of almost 0.24 g/mi. Note, however, that only about 0.12 g/mi occur within the U.S.; the other half occur outside the U.S. (e.g., tankers at sea, overseas drilling). Total NO_x in the SoCAB is comprised of exhaust emissions and SoCAB fuel-cycle emissions and totals almost 0.06 g/mi.

The vehicle above the Ford Exhibition is a 5 mi/kWh BEV. Such a vehicle is typically a small two- to four-passenger auto. Although it has zero exhaust emissions, the per-mile emissions of NO_x Based upon the electricity fuel-cycle result in total NO_x emissions of about 0.09 g/mi. Note that only about 0.02 g/mi occur in California; the rest occur elsewhere in the U.S.

The top four entries represent vehicles that would have the same NO_x emissions in California as the BEV (about 0.02 g/mi). The top two entries have assumed tailpipe NO_x emissions of 0.01 g/mi. The top (first) vehicle uses diesel fuel and the second uses gasoline.

The third and fourth vehicles assume that tailpipe NO_x emissions are only 0.005 g/mi. (recall that the 1998 Mercury Grand Marquis has NO_x emissions of 0.004 g/mi). Fuel efficiency of these two vehicles can be poorer than their counterparts above and still equal the total NO_x emissions in California of the 5 mi/kWh BEV.

Note that all four of the theoretical vehicles have total NO_x emission in the U.S. that are less than half the U.S. NO_x emission of the BEV.

Another comparison of importance is the relative size or utility of the hypothetical vehicles when compared to the two- to four-passenger BEV. Conventional vehicles

with fuel economy of 46 mi/gal would be similar to the BEV. However, conventional vehicles with fuel economy of 32 mi/gal would be somewhat larger that the comparison BEV.

The diesel vehicles would be even larger; for example, diesel trucks and SUVs typically fall in the 20–30 mi/gal range.

As noted earlier, HEVs can provide significant improvements in fuel economy (and tailpipe emissions) when compared to their conventional counterparts. Thus, these vehicles could be even larger than their counterparts and still achieve equal or lower NO_x emissions when compared to the two- to four-passenger BEV. For example, the two-passenger Honda Insight has fuel economy of 70.

D. HEVs Are Also Proposed as Power Generators for the Electricity Grid

Electric Power Research Institute (EPRI) is promoting plug-in HEVs not just as loads for recharging at night but as generators of power for the grid during peak hours, a technology first patented in 1980. HEVs that are plugged in for recharging could, instead, be providing power back into the grid from their batteries and engine. Such schemes have already been demonstrated to be technically feasible, and the ability to control such distributed generation resources is the focus of active study by the utility industry.

Such an application changes the vehicle from an expense item into a capital investment with a potential return on that investment. The use of HEVs for daytime peak power could be most appealing economically for both the HEV owner and the utility. Most utilities in the U.S. experience daytime peaks in loads, and nighttime baseload levels are much lower. For example, the disparity is so significant that one utility, Southern California Edison, offered an experimental time-of-day pricing structure to BEV owners; in exchange for low nighttime rates of $0.04 per kWh, the utility would charge $0.325 per kWh for any power used during the daytime.

VI. ANALYTICAL TOOLS

ADVISOR is a computer model developed at the National Renewable Energy Laboratory (NREL) in conjunction with a wide user group.

The code is available for free download from the Internet at http://www.ctts.nrel.gov/analysis/download.html. It requires the use of the software package MATLAB and SIMULINK available from The Math Works at http://www.mathworks.com/index.shtml.

An on-line demonstration of the capabilities of the ADVISOR program can be viewed at http://www.ctts.nrel.gov/analysis/demo/.

SEE ALSO THE FOLLOWING ARTICLES

AUTOMOBILE TECHNOLOGY • BATTERIES • ELECTRIC MACHINES, DESIGN • ENERGY FLOWS IN ECOLOGY AND IN THE ECONOMY • ENERGY RESOURCES AND RESERVES • GREENHOUSE WARMING RESEARCH • INTERNAL COMBUSTION ENGINES • OZONE MEASUREMENTS AND TRENDS • POLLUTION, AIR • TRANSPORTATION APPLICATIONS FOR FUEL CELLS

BIBLIOGRAPHY

California Air Resources Board. (1990). "Proposed regulations for low-emission vehicles and clean fuels," Staff Report.

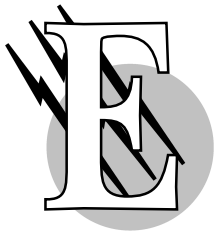
Electric Power Research Institute. (1999). "Electricity technology roadmap C1-112677-V1."

National Research Council. (1991). "Rethinking the Ozone Problem in Urban and Regional Air Pollution," Committee on Tropospheric Ozone Formation and Measurement, National Academy Press, Washington, DC.

Reuyl, J. S., and Schuurmans P. J. (1996). "Policy Implications of Hybrid-Electric Vehicles," under contract to NREL, can be downloaded from http://www.ott.doe.gov/pdfs/nevcor.pdf.

Reuyl, J. S. U.S. Patent No. 4, 182, 960, "Integrated residential and automotive energy system," January 8, 1980.

Unnasch, S., *et al.* (1996). "Evaluation of fuel-cycle emissions on a reactivity basis," Final Report FR-96-114, Acurex Corporation.



Earthquake Engineering

Robert V. Whitman
Massachusetts Institute of Technology

I. Scope of Earthquake Engineering
II. Basic Concepts Concerning Structural Response
III. General Principles of Design Against Earthquakes
IV. Analysis of Buildings
V. Lifeline Components and Systems
VI. Identification and Evaluation of Earthquake Hazards

GLOSSARY

Ductility A property of a structure that allows it to continue to support its load after the structure has yielded or cracked.
Geotechnical engineering Engineering that deals with construction on, within, and with soil and rock (e.g., foundations and earth dams).
Intensity The strength of ground shaking at a site. One measure of intensity is the peak ground acceleration.
Liquefaction A loss of soil strength or stiffness. It occurs when cyclic straining causes large pressures to build up in the water phase of a soil.
Magnitude A measure of the strain energy released by an earthquake.
Nonstructural The parts of a building system that do not support loads (e.g., partitions, lighting fixtures).
Seismology A branch of earth science dealing with earthquakes.

EARTHQUAKE ENGINEERING is a multidisciplinary field that assesses threats to people and their natural and constructed environment that occur as a result of earthquakes, and that develops and puts into action plans and designs that will reduce the consequences of earthquakes. Earthquake engineering overlaps with earth sciences on one hand and public policy issues on the other. Membership in the professional society for earthquake engineering (Earthquake Engineering Research Institute, or EERI) is primarily drawn from the structural and geotechnical engineering fields, but it also includes architects, building code officials, public safety officials, transportation officials, regulators of utilities, city planners, land use planners, medical personnel, economists, political scientists, and other professionals and scientists.

I. SCOPE OF EARTHQUAKE ENGINEERING

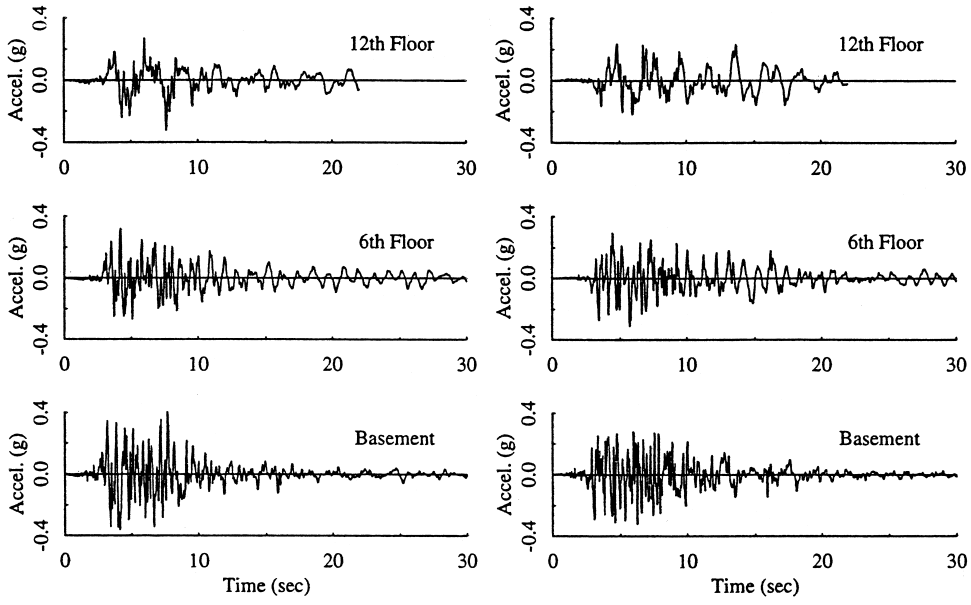
A. Damage Caused by Earthquakes

Large, damaging earthquakes are the result of a sudden slip along a fault. If the fault break reaches the surface of the earth and passes directly beneath a building, bridge, highway, or pipeline, for example, or through a tunnel, the damage to these structures can be enormous. However, usually the greatest damage from earthquakes is caused by the resulting release of strain energy stored in the adjacent rock. The release causes stress waves to spread out through the earth from the epicenter, which causes the ground at any location to shake back and forth as the waves pass. Special instruments—strong-motion seismographs—have been developed to record the very strong motions that are of primary interest for earthquake engineering. Figure 1 shows horizontal acceleration measured at the base of a building during a strong ground motion. The motion changes erratically with time, as a consequence of many different frequencies present in the motion. Although the vertical component of ground motions can be important in some engineering problems, the horizontal components are usually of the greatest consequence. The number of such records has been increasing dramatically since about 1970, and they have become essential for understanding the ground motions that are caused by earthquakes.

Ground motions such as those shown in Fig. 1 will likely cause severe damage and even collapse of poorly constructed buildings, resulting in serious injuries and fatalities. After major earthquakes, especially those occurring in developing countries, newspapers feature pictures of collapsed houses made of adobe bricks, rounded field stones, or even brick masonry. However, such damage can also happen to modern-looking structures that have been poorly constructed. Even when buildings have been designed with some seismic resistance and constructed properly, there may be expensive-to-repair damage—especially to nonstructural features such as exterior walls, partitions, ceilings, mechanical equipment, contents, etc.—accompanied by serious injuries. Business or occupancy interruption can add significantly to the cost of all such damage.

The same potential problems and challenges apply to the many other parts of our built infrastructure. It is especially important that *essential facilities*—hospitals, fire stations, and police stations—be able to function in the hours and days following a damaging earthquake. Damage that impairs the functioning of *lifelines*—highways and other transportation systems, as well as electrical, water supply, and waste disposal systems—can seriously disrupt the functioning of a community.

Strong ground shaking resulting from earthquakes can also induce other geological hazards, which in turn may cause damage to buildings and other structures.



(a) N-S Component (b) E-W Component

FIGURE 1 Horizontal accelerations measured in two directions—(a) vertical and (b) horizontal—at various levels in a 12-story building. Accelerations are normalized to the acceleration of gravity, *g*. [From Uang, C. M. *et al.* (1997). "Earthquake Spectra," Earthquake Engineering Research Institute, Oakland, CA.]

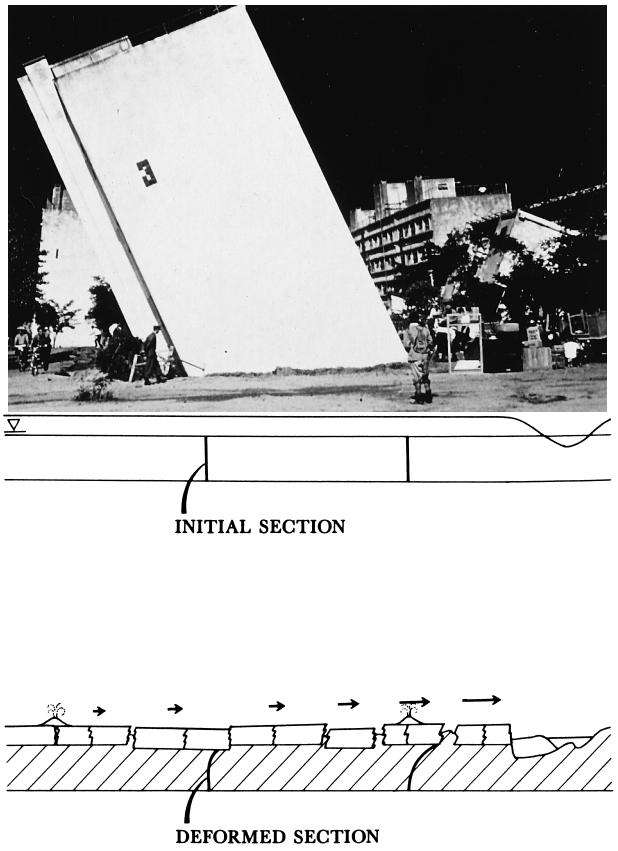


FIGURE 2 Consequences of liquefaction. (a) Settlement and tipping of an apartment building in Niigata, Japan, as a consequence of the 1964 earthquake. [Courtesy of George Housner.] (b) Schematic diagram of lateral spreading.

Landslides, rock falls, and avalanches may be triggered by earthquakes and in extreme cases have been known to bury entire towns or to dam valleys and thus create new lakes. When loose, saturated soils are shaken strongly, they may momentarily lose stiffness and strength and thus lose the capacity to support the weight of structures (Fig. 2a). This phenomenon, known as *liquefaction,* may also cause lateral earth movements of many feet, even in nearly horizontal ground (Fig. 2b), and landslides in steeper ground. Liquefaction effects can be severe along the edges of oceans, lakes, and rivers and may be especially damaging to port facilities and to highways, railroads, and pipelines.

Extreme damage to buildings and other facilities such as gas pipelines, fuel storage sites, and factories may be the cause of secondary hazards such as fires and releases of hazardous and toxic substances. Fires following earthquakes have often been a major cause of damage and loss of life. If dams storing water fail because of faulting, ground shaking, liquefaction, or landslides, downstream flooding may potentially submerge large areas, causing great damage and many casualties. Dams formed by landslides are often later overtopped, which causes downstream flooding months after the causative earthquake. All these secondary hazards have the potential to cause greater losses than those occurring during the original ground shaking.

Finally, ocean-front properties may experience *tsunamis,* which are long and deep waves generated when a portion of the ocean's bottom drops or rises suddenly as a result of motion along a submerged fault. Tsunamis can travel large distances across oceans at great speeds. When the waves meet shallow water at a shore, they can rise up to immense heights and surge inland, causing enormous devastation. A similar type of phenomenon can also occur in large lakes.

B. Reduction and Prevention of Damage

Prevention of damage caused by geological hazards—faulting, liquefaction, landslides, and tsunamis—requires, first, identification of these hazards and, second, land use planning and regulation to avoid siting of structures in such locations. However, when private property is involved, taking such actions in the political arena is never simple and easy. Efforts to avoid damage from geological hazards become much more difficult when structures of any type already exist in areas identified as susceptible to these hazards.

The design of new buildings and other structures to resist the effects of ground shaking is a matter of good engineering plus the willingness of an owner to pay for whatever additional costs are required. Building codes in many cities and states require design against earthquakes. Engineers and architects implement these code provisions. However, there still are communities without adequate code requirements in regions subject to earthquakes. Convincing such communities to adopt suitable code requirements is a continuing task.

Building codes are aimed primarily at protecting occupants from injury. A building can be designed to meet code and fulfill the mission of protecting life safety and yet experience significant damage during an earthquake. Owners must bear the cost of expensive repairs, and buildings may be unusable until repairs are made. An increasing number of owners want to ensure continued usability of their buildings following an earthquake, and they are enlisting engineers and seismologists for advice.

However, existing buildings and structures designed before there were serious concerns about earthquakes constitute by far the greatest risk to life and property. To bring all such buildings just up to code requirements would involve prohibitive costs. Hence there are challenging

questions: Which buildings are so dangerous that some actions should be required? What types of strengthening actions are most cost-effective? Often the most dangerous buildings are those occupied by the poorest portion of the population. What happens to these people while buildings are strengthened? Higher rents charged for the improved buildings may force the occupants to seek shelter and business space elsewhere. The "existing building problem" thus involves social, economic, and political considerations in addition to engineering considerations.

This discussion of earthquake-related hazards and the challenges involved in dealing with these problems has emphasized the broad, multidisciplinary nature of earthquake engineering. It suggests the breadth of knowledge and skills required to assess the potential hazards and risks and their consequences and to develop and put into action measures to mitigate the earthquake threat. Buildings, housing residences, and offices represent the large bulk of the earthquake hazard mitigation challenge; therefore, the following discussion focuses primarily on the analysis and design of buildings. There are limited comments concerning other types of structures.

II. BASIC CONCEPTS CONCERNING STRUCTURAL RESPONSE

The theory of structural dynamics is fundamental to the understanding and analysis of the response of structures subject to earthquake ground motions. This theory identifies a number of basic concepts, the most important of which are as follows:

- A structural system has a number of *degrees-of-freedom*. The idealized building in Fig. 3a has three degrees-of-freedom in the plane of the paper that correspond to horizontal motions of the three floors where mass is concentrated. The number of important degrees-of-freedom typically increases with the height (or size) of a building and with the flexibility of the structural members.
- The structure possesses a number of *natural frequencies* (*or natural periods*) equal to the number of degrees-of-freedom. The smallest of these natural frequencies is the *fundamental frequency*. The fundamental frequency (and of course other natural frequencies as well) decreases with increasing building height and flexibility. If a structure is disturbed and allowed to vibrate freely without further application of force or displacement (*free vibrations*), the vibrations will occur at these natural frequencies.

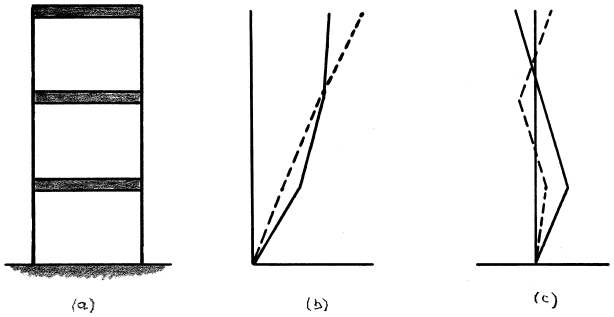


FIGURE 3 (a) Idealized three-degrees-of-freedom structure. (b) Typical shapes for fundamental mode. (c) Typical shapes for second (solid lines) and third (dashed line) modes.

- When a structure is vibrating at a natural frequency, there will be a characteristic pattern (*mode shape*) for the displacements. Figure 3b shows typical shapes of fundamental modes, which depend upon the distribution of mass and stiffness. Typical shapes for second and third modes are sketched in Fig. 3c.
- When a building is *forced* by a base motion such as that in Fig. 1, the response of the structure is a superposition of the mode shapes, with the relative importance of each mode shape depending upon the ratio of the predominant frequency in the base motion to the natural frequency of that mode.

When the earth beneath a building moves rapidly, the building attempts to move with the ground. Force must develop between the building and the ground to make it move and within the building to make the upper portions move with the lower part. It is these forces, and the associated deformations, that can cause the structural system and nonstructural features to be damaged or even to fail. If a building were completely rigid, the horizontal force at each level would equal the mass of the overlying building times the ground acceleration. Actual buildings, however, are flexible to some degree; as a result, accelerations within a building may be either greater or smaller than those at its base, and, as seen in Fig. 1, will have different frequencies.

The important role of flexibility is captured by a *response spectrum* (Fig. 4a), which for a particular ground motion plots the *spectral acceleration,* S_a (approximately the peak acceleration), experienced by the mass of a single-degree-of-freedom mass–spring–dashpot system versus the undamped natural period T of that system. T is given by

$$T = 2\pi(M/k)^{1/2}, \tag{1}$$

where $M =$ mass and $k =$ spring constant for the simple system. The product MS_a gives the force in the spring.

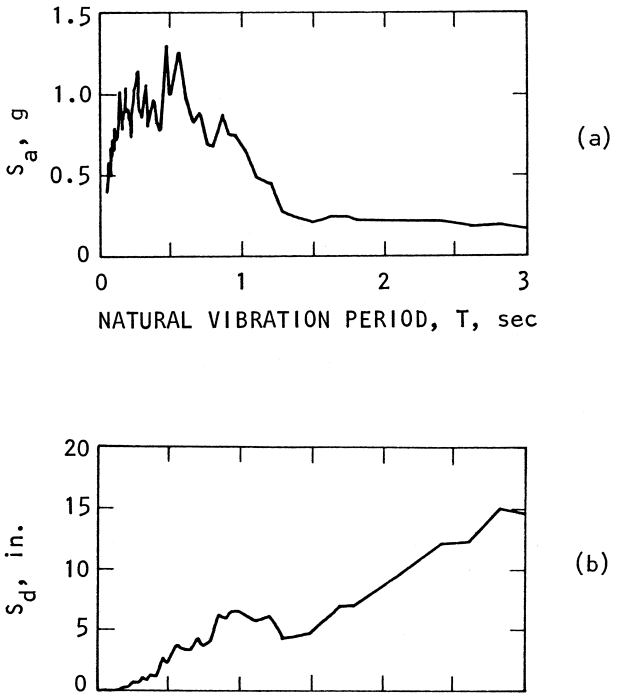


FIGURE 4 Response spectra. (a) Spectral acceleration, S_a. (b) Spectral displacement, S_d.

The case $T = 0$ corresponds to a truly rigid building, and in such a case the peak acceleration of the mass does equal the peak acceleration of the ground. As T increases, the peak acceleration of the mass generally becomes larger than that of the ground. There typically is some range of T for which the mass experiences accelerations several times greater than those of the ground. The peaked portion of the spectrum corresponds to the periods of ground motion that are strongest and most sustained. With further increase of T (which implies further increase of flexibility), the ratio of peak accelerations at mass to those at ground decreases and eventually becomes significantly less than unity.

Figure 4b is a related plot of maximum distortion in the spring (*spectral displacement, S_d*), which is related to spectral acceleration by

$$S_d = MS_a/k. \tag{2}$$

Comparing Figs. 4a and 4b reveals that stiff structures (small T) experience large forces but small distortions, whereas flexible structures (large T) may feel small forces but at the expense of large distortions.

These concepts remain useful for understanding and analyzing the response of actual buildings having several degrees-of-freedom. The acceleration and deformations experienced by a structure are also influenced by the damping present in the structure, and hence a series of curves (*response spectra*) for different values of damping are often plotted. Spectra are calculated from actual ground motions. Spectra from different motions have often been combined to produce "smoothed spectra" for use in design studies. Response spectra are engineering tools and are related to, but basically different from, the Fourier spectra that seismologists often use to represent the characteristics of ground motions.

A final, but very important, point concerns the duration of ground shaking. Such duration is primarily a function of the magnitude of an earthquake, being short (a few seconds) for earthquakes of magnitude less than 6 but long (several minutes) for great earthquakes of magnitude 8 or higher. Increasing duration tends to cause somewhat larger ordinates in response spectra. However, the greatest consequence occurs when the strength of a soil or structure decreases dramatically as shaking continues.

III. GENERAL PRINCIPLES OF DESIGN AGAINST EARTHQUAKES

A. Ductility

A very important, basic concept is that reaching the limiting strength of a structure during a rapid shaking does not necessarily mean that the structure as a whole will "fail" in the sense of collapsing. A structure is said to be "ductile" if it continues to hold together and support its weight after its strength has been momentarily reached. Conversely, a "brittle" structure that lacks ductility may come apart and collapse as a result of a shaking that causes its strength to be reached.

The ductility of a building depends not only upon the material used for beams, columns, floors, and walls, but also upon the way in which these elements are connected. Wood-framed walls with adequate bracing or sheathing are ductile. However, wood-framed houses inadequately connected to a foundation may collapse during an earthquake. Structural steel is a ductile material; a beam may yield and experience plastic deformations without cracking. However, connections between beams and columns must be specially designed to achieve ductile behavior. Concrete is inherently brittle; once its strength is reached, it begins to crack and split. Nevertheless, providing sufficient reinforcing steel within concrete allows a degree of ductility to be achieved. With precast concrete construction, having ductility depends upon the manner in which beams are connected to columns and the way in which walls are tied to floors and roofs. Unreinforced masonry construction is very brittle. Some ductility can be achieved by inserting steel bars into masonry walls.

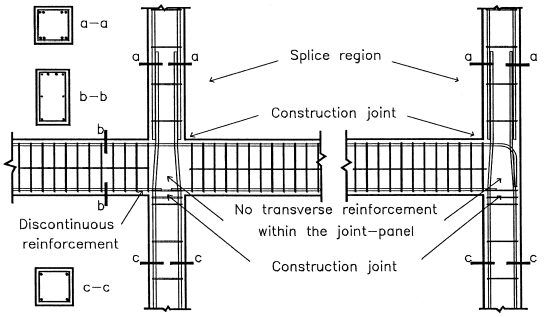
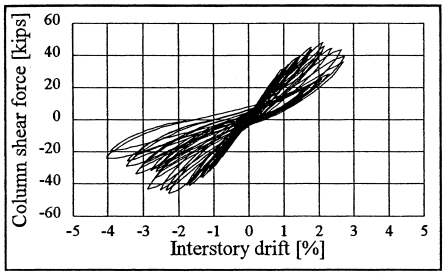


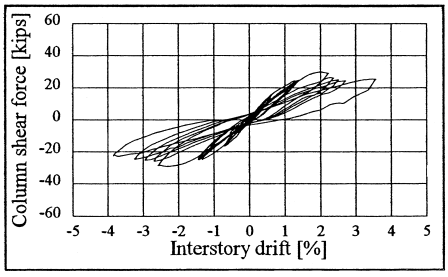
FIGURE 5 Elevation view of interior and exterior reinforced concrete beam–column joint regions. The reinforcement is typical of buildings designed without regard for earthquakes. [From Beres, A. *et al.* (1996). "Earthquake Spectra," Earthquake Engineering Research Institute, Oakland, CA.]

Past earthquakes have provided invaluable lessons concerning adequate and inadequate ductility. Following any major earthquake, teams of experts examine damaged and undamaged structures so as to understand the reasons for good or bad performance. "Learning from earthquakes" has been the key to achieving ever better design for buildings. In addition, many large-scale tests have been performed in laboratories. Various structural systems have been examined in this way, including reinforced concrete members and assemblages of steel and concrete members. These experiences form the basis for rules concerning adequate reinforcement for concrete members and adequate connections for steel members.

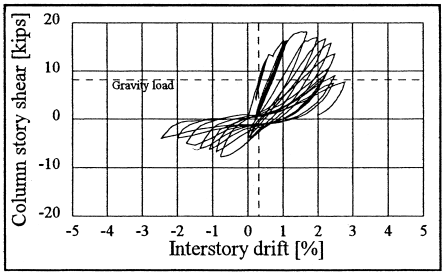
As an example of the challenge in providing ductility, Fig. 5 depicts test specimens representing portions of a building. The double lines within the concrete represent steel reinforcing bars, which give tensile strength to the concrete. The solid narrow lines are hoop steel that restrain the reinforcing bars from buckling outward and participate in resisting shear. This configuration is typical of a reinforced concrete structure that has been designed without regard to earthquakes. When these assemblies are tested, sustained vertical loads representing the weight of an overlying structure are applied, then cyclic horizontal loads are superimposed. Figure 6 shows typical results. Figure 6c illustrates inadequate behavior: After the second large cycle, the resistance falls off rapidly with subsequent cycles. Thus, failure of this beam–column assembly would occur during an earthquake with duration long enough to impose more than a few cycles of shaking. The behavior charted in Figs. 6a and 6b is better, but still inadequate for resisting intense ground shaking. Such results plus experiences during actual earthquakes show the need for additional reinforcement at and near beam–column junctions.



a. Interior joint, continuous positive bars



b. Interior joint, discont. positive bars



c. Exterior joint

FIGURE 6 Results from cyclic loading tests on the assemblages shown in Fig. 5: (a) interior joint, continuous positive bars; (b) interior joint, discontinuous positive bars; (c) exterior joint. Interstory drift is the relative displacement between adjacent floors. [From Beres, A. *et al.* (1996). "Earthquake Spectra," Earthquake Engineering Research Institute, Oakland, CA.]

B. Layout of Building

In addition to choosing materials and details that provide ductility, paying close attention to the layout of a building and structural system is essential. Irregularities in the layout of a building often cause difficulties. A change in the structural framing system at some elevation within a building can introduce a weak spot. A very common problem, and the cause of many collapses, is a "soft first story," where walls have been omitted so as to accommodate offices, stores, or garaging of vehicles. Irregularity in plan can also be the cause of severe damage and even collapse, in part by causing a building to twist. If one building is placed too close to another structure, they may pound against each other, which will likely cause damage and even partial collapse.

C. Controlling Damage

Although providing ductility can make it possible for buildings to undergo large deformations and resulting damage without collapsing, there are several reasons why large motions and deformations with associated damage may be undesirable. Large interstory deformations may cause the exterior cladding of a building, including windows, to fall from the building and threaten lives on the ground below. Ceilings may fall upon occupants of rooms. Elevators may become inoperable, and stairways blocked by fallen partitions, which would hinder escape from the damaged building. For these reasons, building codes generally limit the permissible interstory displacement. In addition, care in the design of nonstructural components is needed to reduce losses.

If essential facilities are to perform their critical roles following an earthquake, electrical and mechanical components must remain functional. Some elementary examples are the large doors of fire stations and the electrical supply to hospitals. There are enormous implications to the loss of large computerized databases in banks, and increasingly manufacturers are concerned about protecting production lines from disruption. Such key nonstructural features must themselves be robust enough to endure strong shaking without failing, but they must also be mounted properly.

At still another level, the release of toxic and poisonous substances as a result of earthquake shaking can threaten large populations in circumstances where escape from them is hampered by clogged roads and failed bridges. Nuclear power plants are an extreme example. In such cases, drastic and costly measures must be taken to ensure that containment vessels and structures do not experience significant damage and that control and safety equipment survive an earthquake.

Protecting against damage that is socially or economically unacceptable first suggests designing buildings to be stronger and stiffer. However, there are limits to the economic viability of such an approach. One possible alternative approach is *base isolation*. Conceptually, this approach means placing below a structure a weak surface that limits the shear force that can be transmitted upward into the structure from moving ground. The building as a whole may "slide" over the ground, with permanent displacement at the end of a major earthquake, but the structure above that level and its contents are undamaged. An immediately obvious requirement is to provide flexible connections where utilities enter the structure. Indeed many details need attention if the base isolation scheme is to function reliably and effectively. A number of proprietary base isolation systems have been developed and have been applied to key governmental buildings, manufacturing plants, banks, and so forth. They have also been used to achieve life safety requirements in existing, hazardous, historic buildings. In addition, special energy-absorbing devices, such as large dampers, have been developed for placement within structural systems to reduce deformations during earthquakes.

IV. ANALYSIS OF BUILDINGS

A variety of approaches may be used to analyze an existing structure or a trial design for a new structure, depending upon its nature and importance.

A. Equivalent Static Analysis

The great majority of buildings and other structures are simply designed in accordance with codes, which permit the use of horizontal static loads to simulate the forces associated with earthquakes. The procedure, which takes into account key aspects of dynamic response, is as follows:

- The horizontal shear, V, at the base of a building is computed from

$$V = ZC(T)W/R, \tag{3}$$

  where W is the weight of the building, $C(T)$ is a seismic coefficient depending upon the fundamental period T of the building, Z is a zone factor, and R is the force reduction factor dependent upon the structural system used for the building.
- The function $C(T)$ is a simplified, generic version of a response spectrum. Codes give simple equations for computing T depending upon the height of the building and the type of structural framing.
- Z is a zone factor that accounts for the appropriate local seismicity. It is read from a map of the nation. In the United States, Z is largest for states bordering the

Pacific Ocean and markedly less for many states in the eastern part of the country.
- Forces that sum to the base shear are distributed over the height of the building following procedures set forth in the code as a way of taking into account the distribution of mass and the expected shapes of the first several modes of the building.
- These forces are then distributed to the members of the beams, columns, and walls of the building in accordance with common structural engineering procedures.
- These pseudodynamic forces are combined with dead loads and other live loads, and, along with suitable safety factors, are then used to design the structural system.

If R were equal to 1 and no safety factors were to be used, then the ground shaking during a very strong (for the location) earthquake would bring a building essentially to the point of yielding. However, since such earthquakes are rare and the cost of designing against such a strong shaking can be significant, codes permit the use of values of $R > 1$; just how much greater depends upon using structural systems that will continue to support dead loads after being damaged by the strong shaking. For very ductile structural systems (e.g., steel frames with moment-resisting beam-to-column connections), $R = 8$ is allowed, whereas for building systems with poor ductility (e.g., unreinforced masonry or reinforced concrete lacking in adequate reinforcing steel bars), R is permitted to exceed 1 only slightly.

This simple approach, which is adequate for many buildings, has evolved since the 1930s. Values of $C(T)$ and R have been adjusted and readjusted to reflect experiences with buildings during actual earthquakes. Other factors—S, to account for the effects of soil conditions at the building site, and I, to reflect the importance of a building—have also been added, which leads to Eq. (4):

$$V = ZSIC(T)W/R. \tag{4}$$

The most important improvements in this equivalent static approach have been in the requirements that a designer must fulfill in order to qualify the building for values of R such as 4–8. Particular attention must be given to beam–column connections in steel structures and to the details of reinforcing steel in concrete or masonry structures. Codes also require attention to some nonstructural features, especially to ensure that the exterior cladding and windows will not fall off a building as a result of ground shaking.

B. Linear Dynamic Analysis

Today, computer codes for performing dynamic analysis of buildings subjected to ground motions are readily available and are found in the offices of nearly all structural engineering firms—at least in states with significant earthquake problems. It is necessary only to input the geometry of the framing system and the properties of the members, and often this can be done automatically by linking to a computer program used as part of the development and design of the building. The dynamic analysis code will compute natural periods and mode shapes and, once the motion at the base is specified, will determine forces in all members of the building for comparison with allowable forces. Using graphical interfaces, the user can watch the building move and deform as the shaking takes place.

Since the 1930s, the methods for performing dynamic analysis have evolved through many stages. A first major step was the theory for modal analysis. In this procedure, the response of each mode is determined separately and then these modal responses are superimposed. Although a full time-history of response can be obtained in this way, often it suffices to focus only upon the peak forces in members, and these can be estimated by using rules to combine the peak responses for each mode. The peak modal responses can be obtained by reading the response spectrum for ground motion at the natural periods of the several modes. Except for tall buildings, often it suffices to consider only the first three to five modes.

With today's powerful computer programs, response will more likely be calculated by direct integration of the simultaneous equations of motion. Fourier analysis is also sometimes used.

The use of linear dynamic analysis becomes important for buildings with unusual framing systems, very tall buildings, and especially buildings that are irregularly shaped in either plan or elevation. These are situations for which the approximate analysis following a building code may give misleading forces in some members. Often analyses will be made by using two different levels of shaking. The first is a shaking that can reasonably be expected to occur during the life of the structure. For this shaking, all stresses should remain within allowable stresses as specified by codes. The second shaking is the largest credible shaking, and it may lead to predicted forces that exceed the yield strength of the materials. Even though a linear analysis then is, strictly speaking, incorrect, experience has shown that the displacements and deformations predicted by the analysis will be approximately correct. These displacements and deformations can be compared with those that members can sustain and still remain intact.

C. Pushover Analysis

Pushover analysis is a static analysis used to investigate how far into the inelastic range a building can go before it is on the verge of a total or a partial collapse. A model for the building is assembled on a computer, with all

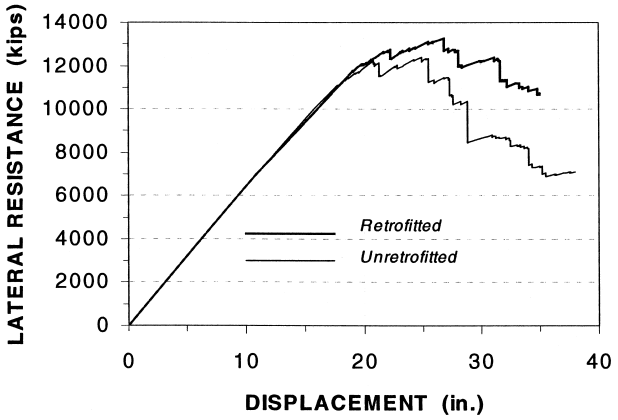


FIGURE 7 Pushover curves for a building before and after retrofit to increase ductility. [Courtesy of EQE, Oakland, CA.]

load-resisting elements together with their force–deformation relationships both before and after yielding and with dead loads plus average live loads. Then a small set of horizontal forces is applied so as to simulate the effects of ground motions, and deformations are calculated. The forces are then increased in steps so as to develop a plot of base motion versus deformation. Examination of this plot reveals the largest base motion that the building can resist.

This approach was developed to permit analysis of existing buildings and to study the effectiveness of schemes for strengthening these buildings and giving them greater ductility. Figure 7 shows typical plots for an existing building both before and after it is remediated. The existing building begins to lose resistance rapidly once its peak resistance is exceeded, and it would likely collapse during a very strong shaking. The strengthened building exhibits some ductility that will make collapse much less likely. Pushover analysis is now also used frequently to evaluate the expected performance of designs for new buildings.

D. Nonlinear Dynamic Analysis

There now exist computer programs that are capable of calculating nonlinear response in detail, including the effects of repeated yielding during successive cycles of ground motion. Analyses using these programs can consume large amounts of computer time. Such programs are used to evaluate designs for large and important buildings and those for which base isolation devices are to be used. The biggest challenge in the use of such programs is establishing realistic force–deformation relationships for all of the load-resisting elements, especially when these members are of reinforced concrete. Results from nonlinear analyses can be quite sensitive to the details of applied ground shakings. For this reason, analysts select a set of different "scenario earthquakes" that could cause severe shaking of the building so as to adequately check the performance of the building.

E. Response of Nonstructural Systems

A building may house and provide support for sensitive equipment and piping, the failure of which can have a severe impact on the functioning and safety of the building. Examples are hospitals, banks, and, at the extreme, the containment structure for a nuclear power plant. A dynamic analysis for the structure itself can be carried out in such a way as to provide dynamic motions at the points of support for such sensitive equipment. Again it is necessary to use multiple "scenario earthquakes" to account for potential uncertainties in earthquake ground motion and in the response of the basic structure. In effect, the task is to establish suitable support-point response spectra.

F. Soil–Structure Interaction

The response of a building to ground shaking can be affected by flexibility in the ground supporting or surrounding the base of the building. Rocking is an obvious example. The soil–structure interaction may have important consequences for especially tall or massive structures. Special methodologies for analyzing such interaction effects have been developed.

V. LIFELINE COMPONENTS AND SYSTEMS

A. Transportation Lifelines

Much of what has been said concerning the analysis and design of buildings also applies to bridges. Failures of highway bridges and viaducts in California during earthquakes in the late twentieth century have led to considerable research, testing, and development of new procedures for designing new bridges and strengthening existing ones. One of the first challenges has been to design supports that allow for temperature-caused changes in the length of girders and yet keep the girders on these supports during a major earthquake. A particular problem with analyzing very long bridges is that ground motions can be significantly out of phase at the several points of support. California and some other states have undertaken programs to replace or strengthen the most hazardous existing highway bridges.

Except where they cross faults, tunnels generally perform well provided that the liners are flexible enough to adjust to distortions experienced by the ground as it shakes.

Major attention has been given to underwater tunnels, such as the crossing of the Bay Area Rapid Transit system between San Francisco and Oakland, which performed well during a serious (but not severe) shaking in 1989.

Lateral spreading or landslides that take out sections of pavement can seriously interrupt highways and other transportation systems.

B. Dams and Earth Retaining Structures

Failure of a dam, with rapid release of stored water, can be a catastrophic event. Fortunately, there were no earthquake-caused failures of large dams during the last half of the twentieth century, but there were two near misses and several failures of smaller dams. One near miss, near Los Angeles in 1971, spotlighted the susceptibility of earth dams constructed by hydraulic filling techniques. This event led to the development of techniques for analyzing such dams and to continuing efforts to upgrade the performance of existing dams constructed in that manner. Similarly, a near failure of a concrete arched dam in India in 1967 stimulated study of such dams. It is worthy of note that Pacoima Dam, a large arch dam near Los Angeles, has twice experienced very strong earthquake shaking without failure-threatening damage.

Walls to retain highway cuts and fills have generally performed well during earthquakes, unless there has been liquefaction-susceptible soil in the backfill or foundation. As might be expected, there have been widespread problems with retaining walls along rivers and especially in seaports because of liquefaction. Failures during many recent earthquakes have stimulated the use and development not only of techniques for *in situ* improvement of liquefaction-susceptible soils, but also of improved methods for analyzing and constructing retaining structures.

C. Electrical Systems

While electrical generating plants have proven to be robust during earthquake shaking, transformers, circuit breakers, and other equipment in switchyards have often failed, which has caused widespread disruption of electrical service. At least in more earthquake-prone regions, electrical utilities have been taking steps to install more robust equipment.

D. Water and Waste Systems

Although pipes of water distribution systems can be broken just by stress waves passing through soil, most such failures are the result of lateral spreading, landslides, or fault slips intersecting pipes. The resulting system outages are of special concern to a community striving to recover from an earthquake-caused disaster. These problems are addressed by use of pipes that are more tolerant of lateral movements, automatic shutoff valves, increased redundancy in distribution systems, and special steps at fault crossings.

E. Other Lifelines

The seismic resistance of gas pipelines, and other gas-handling facilities, has recently been given serious attention, partly because of economic consequences but also because of the potential for explosions. Communication utilities have generally acted to protect their facilities from disruption by earthquake shaking, so that user overload rather than equipment failure is the primary reason for postearthquake difficulties.

VI. IDENTIFICATION AND EVALUATION OF EARTHQUAKE HAZARDS

For the great majority of buildings and other structures that are simply designed in accordance with codes, a national zonation map gives the ground-shaking hazard based upon a major effort by seismologists of the U.S. Geological Survey.

For some structures, the owner may want assurance that a building will experience little damage during an earthquake that will likely occur during the life of the structure, plus very strong assurance that life safety is protected even during the largest credible earthquake. In such cases, a seismologist will typically become involved in recommending ground motions that meet these specifications. The owner will then expect careful engineering of the structure and, for a building, careful attention to the design of nonstructural features.

When the failure of a structure would threaten many lives (e.g., a large dam or a nuclear power plant), a regulatory agency may become involved in the approval of ground motions assumed for analysis and design. If this happens, geologists will then be asked to assess the possibility that very large earthquakes might occur within a large area structure, and seismologists will be asked to study how large the resulting motions might be at the site. Very thorough and detailed engineering analyses will then be performed.

A. Effects of Local Soil Conditions

Ground motions atop a deposit of soil generally are stronger than those at underlying rock. Figure 8 shows spectra computed from motions measured at nearby soil (upper left recording) and rock (lower left recording) sites

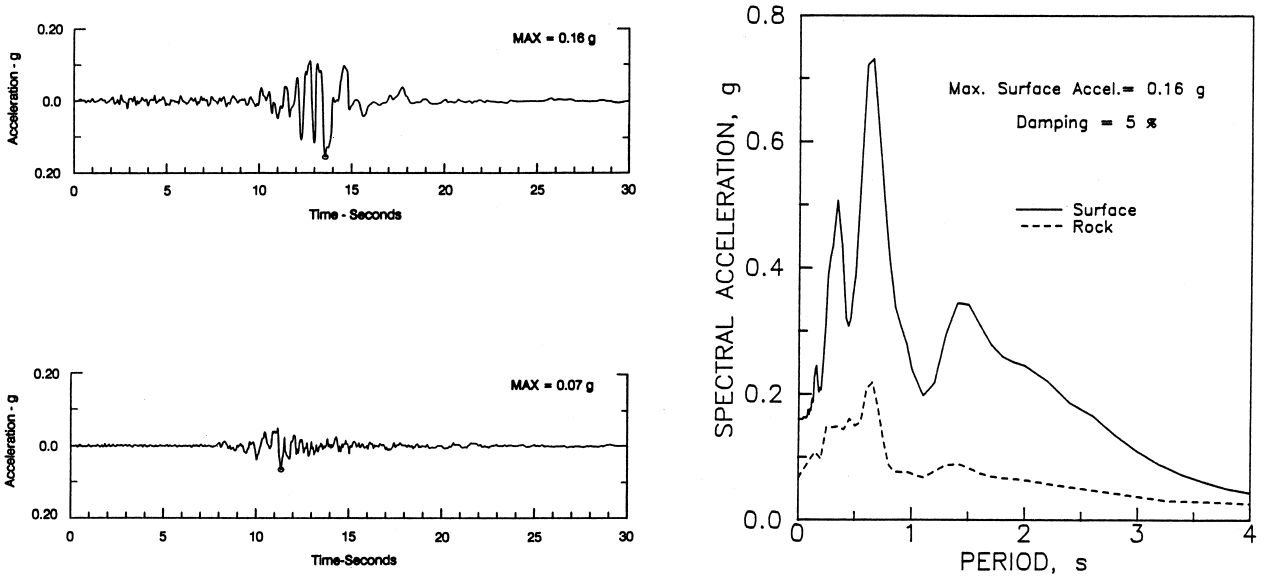


FIGURE 8 Accelerations measured on nearby islands in San Francisco Bay during the 1989 Loma Prieta earthquake. (Bottom left) Yerba Buena Island consists of rock, whereas (top left) Treasure Island is manmade fill over soft clay. The response spectra (far right) were calculated from the recorded accelerations. [From Seed *et al.* (1990). "EERC-90/05," University of California at Berkeley.]

in San Francisco Bay. The significant effect of soil is shown by the corresponding response spectra (Fig. 8, far right). The earthquake that caused great damage and loss of life in Mexico City in 1985 provides a striking example of the potential importance of local soil effects. Figure 9 shows the unusual response spectra recorded in the central part of the city. The large peak at the SCT site unfortunately coincided with the fundamental period of several large buildings, which led to very large deformations and then collapse.

Geotechnical engineers have developed theories for predicting the modifying effects of soil. In the simplest analysis, a column of soil is analyzed in much the same way as the idealized structure in Fig. 3. The stiffness of soils is represented by shear modulus, related to measured or estimated *in situ* shear wave velocity. Since soil is nonlinear, curves giving shear modulus reduction and damping ratio as a function of strain have been developed. Examples appear in Fig. 10. Usually linear analysis is used, with modulus and damping adjusted iteratively so as to be consistent with computed strains. Such calculations from theory have been reasonably well validated by the still scarce information from recordings during earthquakes. The calculations show that low-frequency components of ground motion are increased on soil, in accordance with Figs. 8 and 9. However, because of the large damping that occurs during very strong shaking, calculations suggest that high-frequency components of ground motion may be reduced between the bottom and the top of a soft soil.

Figure 11 summarizes peak ground accelerations measured on soft soils, plus theoretical predictions indicating that peak ground accelerations may be less atop soil than at rock. This aspect of the effect of site conditions is still somewhat controversial.

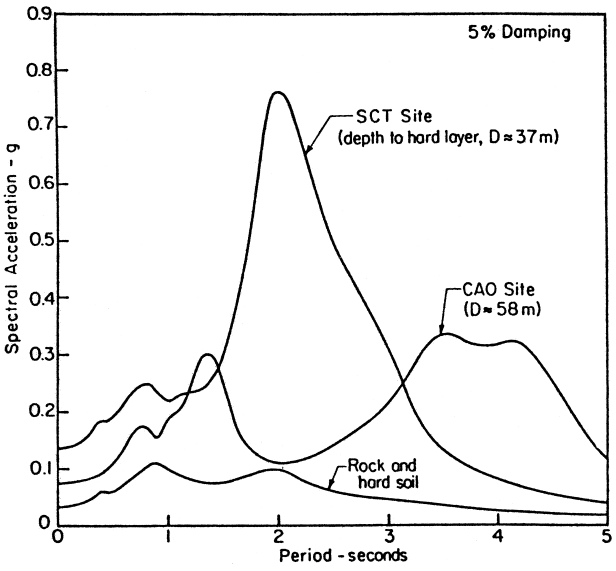


FIGURE 9 Response spectra computed from ground motions recorded in Mexico City during the earthquake of 1985. The SCT and CAO sites are two locations within the city, underlain by different depths of the very soft Mexico City clay. [From Romo, M., and Seed, H. B. (1987). "Int. Conf. Mexican Earthquake," American Society of Civil Engineers, Reston, VA.]

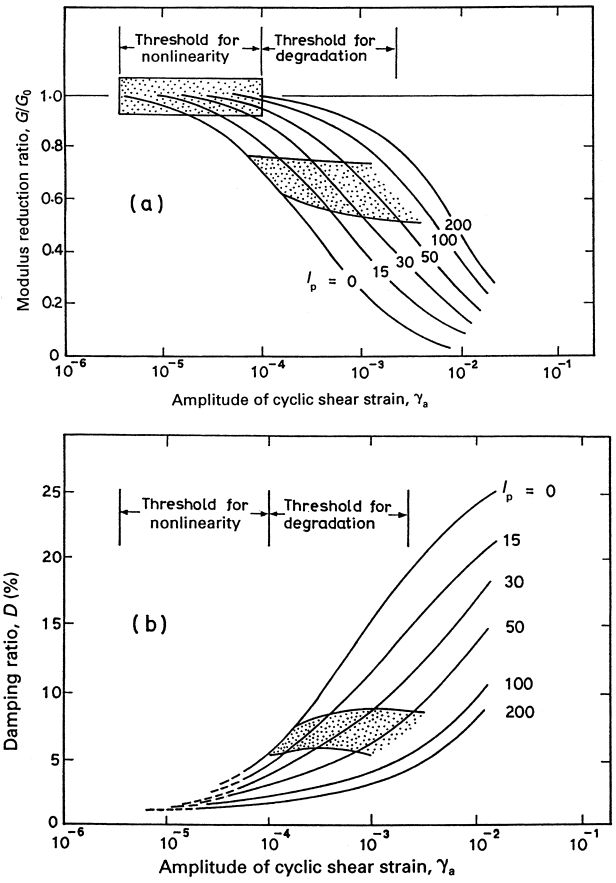


FIGURE 10 (a) Secant shear modulus during cyclic straining in relation to the shear modulus G_0 at very small strains. (b) Damping ratio, which is essentially the same as percentage of critical damping. I_p denotes the plasticity index, with $I_p = 0$ corresponding to cohesionless soil and $I_p = 200$ to a very plastic clay. At the threshold of degradation, the reduction factor and the damping ratio begin to be affected by the number of cycles of straining. [From Ishihara, K. (1996). "Soil Behavior in Earthquake Geotechnics," Clarendon Press, Oxford, after data by M. Vulcetic and R. Dobry.]

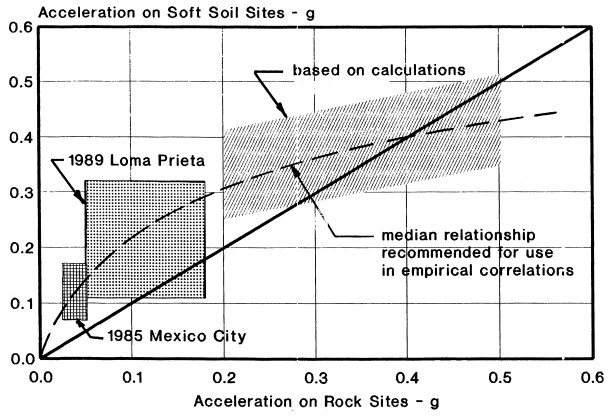


FIGURE 11 Relation of peak acceleration atop soft clay to peak acceleration at underlying rock. [From Idriss, I. M. (1990). "Seed Memorial," Vol. 2, BiTech Publishers, Vancouver, BC.]

Theoretical analyses and field experiences have been used to establish the soil factors S prescribed by building codes, and for important buildings, site-specific calculations are often made by geotechnical engineers. The fact that motions may be different at various elevations within a soil deposit has important implications concerning the behavior of tunnels or other structures embedded within soil, as well as for the dynamic response of buildings having deep foundations. In addition, it is the presence of dynamic stresses within soil and rock during ground shaking that can cause "failure," either liquefaction or landslides.

B. Liquefaction

Liquefaction is the consequence of the tendency for soils to become denser as a result of cyclic straining. When a soil is saturated, it cannot densify if water does not have time to flow out of the soil. This can cause the pressure in the fluid phase of soil to increase until there is little

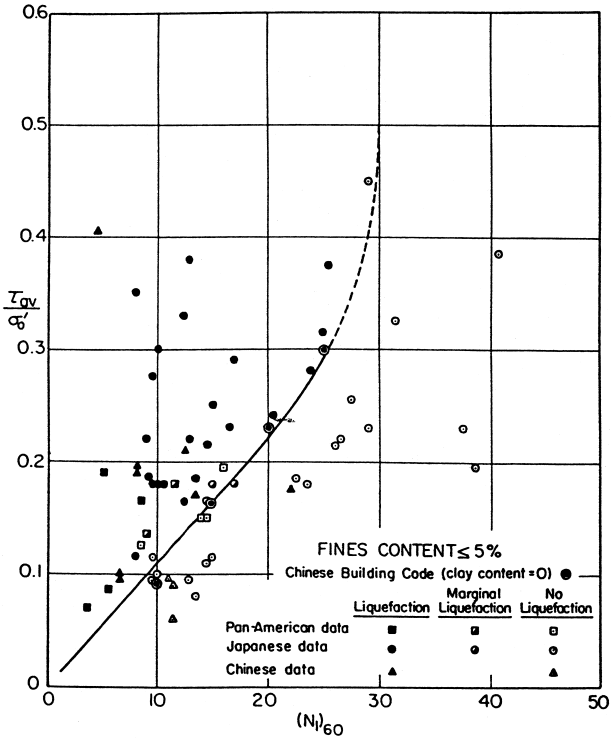


FIGURE 12 Instances of liquefaction and nonliquefaction observed in earthquakes, in relation to intensity of ground shaking (vertical axis) and resistance of soil (horizontal axis). τ_{av}/σ'_o is the ratio of earthquake-caused shear stress in the soil (generally calculated from estimated or measured accelerations) to effective overburden stress. $(N_1)_{60}$ is blows per foot of penetration during a standard penetration test, modified to account for the effective overburden pressure and the energy used to cause penetration. [From Seed, H. B. *et al.* (1985). *J. Geotech. Eng.*, American Society of Civil Engineers, Reston, VA.]

or no stress through the soil skeleton, and the soil loses stiffness and strength. Liquefaction is most likely to occur in fine sands, but it can also occur in gravels confined by less permeable layers and in silts.

Although liquefaction has always occurred during earthquakes, the phenomenon was not identified as such until the Niigata, Japan, earthquake of 1964. It has since been thoroughly studied in the laboratory in an attempt to understand the process and the factors that affect it. It is very difficult to measure directly the liquefaction resistance of a particular soil, unless very specialized sampling techniques (e.g., *in situ* freezing) are used. Hence, in practice, resistance is typically evaluated indirectly by measuring resistance to penetration as rods are pushed or driven into ground following standardized procedures. Figure 12 shows the form for assembling experience from past earthquakes. Recorded instances of liquefaction or nonliquefaction are plotted versus penetration resistance and peak acceleration. A curve separates the zones in which liquefaction is or is not expected at a new site.

Potential consequences of liquefaction—lateral spreading, landsliding, settlement, and so forth—are difficult to analyze and predict by theoretical analysis, but constitutive models and computer codes have been developed and used in connection with special projects. Centrifuge model tests have proved very useful in achieving better knowledge of such phenomena and in validating theoretical methods. These are specialized tests in which a small model of a soil mass is spun at high speeds so as to properly scale the effects of gravity and then is subjected to base shaking while it is spinning. Some useful empirical predictions have also been developed.

SEE ALSO THE FOLLOWING ARTICLES

DAMS, DIKES, AND LEVEES • EARTHQUAKE MECHANISMS AND PLATE TECTONICS • EARTHQUAKE PREDICTION • MECHANICS OF STRUCTURES • SEISMOLOGY, ENGINEERING

BIBLIOGRAPHY

Bolt, B. A. (1999). "Earthquakes," Freeman, San Francisco.

Chopra, A. K. (1995). "Dynamics of Structures," Prentice-Hall, Englewood Cliffs, N.J.

Federal Emergency Management Agency (1997a). "NEHRP Guidelines for the Seismic Rehabilitation of Buildings," FEMA 273, Washington, D.C.

Federal Emergency Management Agency (1997b). "NEHRP Recommended Provisions for the Seismic Regulation of New Buildings," FEMA 222A, Washington, D.C.

Ishihara, K. (1996). "Soil Behavior in Earthquake Geotechnics," Clarendon Press, Oxford.

Kramer, S. L. (1996). "Geotechnical Earthquake Engineering," Prentice Hall, Upper Saddle River, N.J.

Naeim, F., and Kelly, J. M. (1999). "Design of Seismic Isolated Structures," Wiley, New York.



Electric Machines, Design

Enrico Levi
Polytechnic University

I. Specifications
II. Parameterization of Performance
III. Losses
IV. Heating and Cooling
V. Scaling Laws
VI. Summary

GLOSSARY

Air gap Space separating the stationary structure from the moving structure.
Armature Winding carrying alternating current.
Bore Stator cavity that accommodates the rotor.
Field Winding carrying direct current.
Generator Machine designed to convert energy from mechanical to electrical form.
Induction Relating to a machine in which the currents in the secondary winding are excited by electromagnetic induction.
Motor Machine designed to convert energy from electrical to mechanical form.
Power factor Cosine of the phase angle between sinusoidally varying voltage and current.
Saturation Drastic reduction of the permeability in ferromagnetic materials that results from almost complete alignment of the domains with the exciting field.
Squirrel cage Relating to a winding consisting of straight parallel bars short-circuited at both ends by conducting rings.
Synchronous Relating to the angular speed equal to the radian frequency divided by the number of pole pairs.

THE BASIC DESIGN of electrical machines starts with the formulation of the performance specifications. These are translated first into a set of terminal characteristics and appropriate circuit parameters and later into a set of electric, magnetic, thermal, and mechanical stresses and corresponding machine dimensions. The combination of critical stresses and dimensions in simple scaling laws provides guidance for the optimization of the design and permits an estimation to be made of the bulk, weight, and cost of the machine.

I. SPECIFICATIONS

The aim of design is to ensure that the electric drive will perform a specified function with the least economic cost over the lifetime of the equipment. This cost consists of the first investment, or fixed charges, and running charges, the most important of which is the cost of energy. In general,

optimization of the design results from a trade-off between these two components of the cost. Because of the broad economic implications involved, the designer should carefully study the input data. Usually these define the needs the drive must satisfy in terms of the characteristics of the mechanical load and its duty cycles. Once the needs have been identified, they must be translated into a set of specifications, as outlined in the rest of this section.

A. Constant-Speed Motor Drives

The great majority of existing motor drives are fed by electric supplies at constant voltage and constant frequency. Also practically constant is their speed. Their power, instead, varies with the torque, but since the efficiency decreases with decreasing torque, they are preferably operated at constant torque.

These conditions result in the simplest set of performance specifications: the definition of a single rated point of operation. Accordingly, the load is characterized in terms of speed n_R [rpm] and power P_R [W], and the power supply is characterized in terms of voltage V_R [V], frequency f_R [Hz], and number of phases m. Other specifications may include efficiency η, power factor $\cos\phi$, starting current I_{st} [A], starting torque T_{st} [N m], maximum (breakdown or pull-out) torque T_{max} [N m], and type of enclosure.

The first task of the designer is to match the speed of the motor to the load speed by selecting the gear ratio r, if a gear is needed, and the number of pole pairs, p. Since it will be shown that the size and cost of a motor are almost inversely proportional to the shaft speed, one should select the highest economical supply frequency and the least number of poles.

B. Performance Envelopes

In most applications, operation at constant speed results in inefficient mechanical systems, so there is a need for electric drives that can operate efficiently at reduced speeds and torques. In these variable- or adjustable-speed drives there is no single point of operation that defines a nominal power and speed. Performance must be optimized over a duty cycle and an envelope of operating points.

Adjustable-speed drives can be classified into groups according to the general trend of their torque–speed characteristics. The limiting cases are as follows:

1. Torque increasing with increasing speed
2. Torque decreasing with increasing speed

The first group includes centrifugal-machine drives, such as pumps, blowers, and compressors. In terms of power consumption they represent the most important

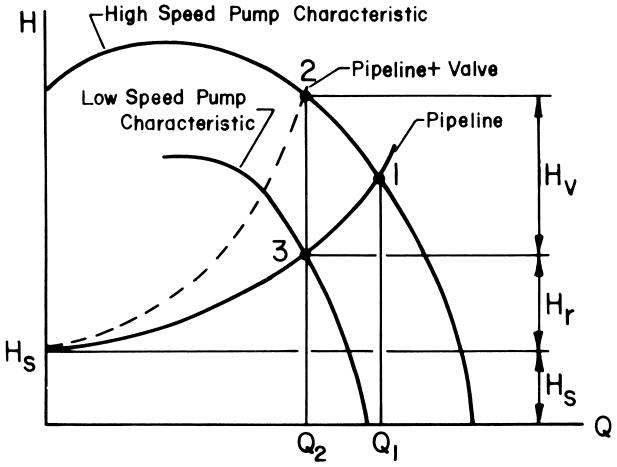


FIGURE 1 Comparison of operation of a centrifugal machine with speed and valve control.

electric drives. The power required by these machines is proportional to the pressure differential usually expressed in terms of the height of fluid column, or head H, and the volumetric flow rate Q. The steady-state point of operation is given by the intersection of the H–Q characteristics of the machine and the pipe network characteristic (Fig. 1). With the centrifugal machine operating at constant speed, a reduction in flow rate, say from Q_1 to Q_2, must be brought about by changing the characteristic of the pipeline. This is accomplished by throttling, or valve control. The composite pipe–valve characteristic is indicated by the dashed line, and the new operating point is 2. The valve increases the head by an amount H_v. The energy loss associated with this additional head can be avoided if, instead of modifying the pipe-network characteristic, one modifies the characteristic of the centrifugal machine by reducing its speed. By shifting the point of operation from 1 to 3, instead of 1 to 2, one can achieve energy savings of up to 60%.

The most important feature of centrifugal machines with regard to the design of the drive is that the absorbed power decreases in proportion to the cube of the speed. The low power requirement at low speed makes it particularly attractive to use a cascaded arrangement in which a wound induction motor is fed at constant frequency and its slip power is conditioned by changing its frequency and is fed back into the supply network. In this case the power conditioner is designed to handle only a small fraction of the peak power of the machine. In low-power applications it may even turn out to be economical to use a squirrel-cage drive at constant frequency but variable voltage. This represents a trade-off between the fixed and running charges by reducing the cost of the power conditioner at the expense of higher energy losses at reduced speeds and power.

High-inertia drives, of which traction applications are the most important, belong to the second group, which

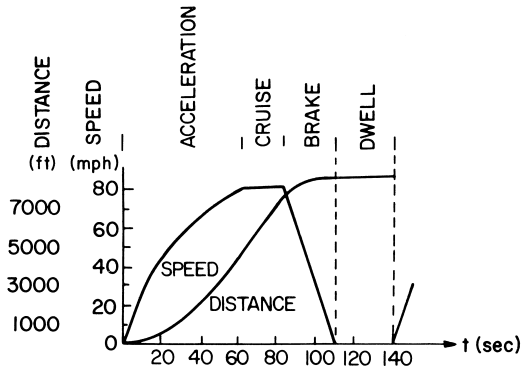


FIGURE 2 Duty cycle of rapid-transit car.

requires the highest torque at starting and, if regenerative braking is used, in stopping. As shown in the typical duty cycle of a rapid-transit car (Fig. 2), fast accelerations and decelerations are needed. To overcome the large inertia of the vehicle, these necessitate torques that may well exceed the cruising torque by one order of magnitude. The speed range is usually divided into two parts: (1) From zero speed to base speed the torque is maintained constant by maintaining constant both magnetic flux and current and by impressing a voltage proportional to speed and frequency, and (2) above base speed the power is maintained constant by keeping constant current and voltage and by reducing the flux and, therefore, the torque in inverse proportion to the speed.

In conclusion, an adjustable-speed drive involves a more complex system because of the presence of a power conditioner, which in general controls both voltage and frequency. This and the large range in power requirements necessitate the formulation of a more complex set of specifications.

A major decision facing the designer of an adjustable-speed drive is the selection of the motor type. The dc motor drive has been and still is the favorite; it is very likely to remain so for some time to come. Since the mid-1970s the appearance in the marketplace of competitors in the form of ac motors fed through solid-state power conditioners has led to a doubling and tripling of the output of the dc motor per unit weight. In the range above 20 kW, this has given the dc drive almost a 2:1 advantage in investment cost. However, this dramatic increase in output is not expected to continue in the future, and new and less expensive switching elements are being developed. Also weighted against the initial cost advantage of dc motor drives is the need for maintenance of the commutator and brushes at least about every 5000 hr.

Among the ac drives, the lower cost favors the induction motor in the smaller sizes, whereas the feasibility of overexcitation, and with it natural commutation of the switching elements, favors the synchronous motor. With regard to the switched-reluctance motor, see the Bibliography of Electric Machines, Rotating, Construction and Theory.

C. Generators

Most generators deliver a constant voltage at constant frequency and, as is the case with motors, operate at constant speed. Because of the need for the delivery of reactive volt amperes (VARs), as well as watts, to electric loads, which are generally inductive, most ac generators are of the synchronous type. When the prime movers are diesel engines or steam or gas turbines, machines that operate efficiently at high speed, the generators have two or four poles and are called turbogenerators. The rotors are cylindrical, and because of the high stresses resulting from centrifugal forces, they are built of solid alloy–steel forgings with diameters limited to ~1 m. The shafts are horizontal and operate at least 20% away from the nearest critical speed, which in the largest ratings correspond to the second natural frequency of oscillation. In contrast, generators driven by slow-speed water turbines, or waterwheel generators, are multipolar, large-diameter machines. Their rotors have salient poles built up of steel laminations in order to reduce field pulsation losses. Because of the delay in the closure of the water gate following a loss of electrical load, the rotors must be designed to withstand overspeed as high as 125% over rated speed. The shafts are often vertical. Induction generators are preferred in small remote and unattended hydroelectric plants, because of their low maintenance requirements. In contrast to the squirrel-cage induction machine, which is rugged, the dc machine is a delicate device and requires much attention and careful adjustment. For this reason shunt-and-compound-excited dc generators are being replaced by alternators coupled to solid-state rectifiers. This is particularly true in transportation applications in which bulk, weight, and ruggedness are important considerations. In aircraft applications, the prime movers usually receive their energy from the main engines. To deliver a constant frequency—400, 2000, or 3200 Hz according to military standard 704B—continuous, variable-ratio hydraulic transmissions are used at the mechanical input. In a recent development the generator is coupled directly to the prime mover and delivers 3000–4000 Hz, depending on engine speed. This is synthesized into a 400-Hz, 120 Y/208 ΔV output voltage either by using a cycloconverter or a dc link rectifier–inverter arrangement. Cooling by sprayed oil mist is becoming the standard.

D. Standards

Electric machines must satisfy standards set by appropriate professional organizations. In the United States

these are the National Electrical Manufacturers Association (NEMA) or the Institute of Electrical and Electronics Engineers (IEEE) and, when existing, by the all-inclusive American National Standards Institute (ANSI). The nameplates should abide by the National Electric Code. In particular, electric motors marketed in the United States should satisfy ANSI/NEMA Standards Publication No. MG1-1978 and more recent partial revisions. Similar standards prevail in other countries.

The standards specify power and speed ratings, voltage, and frequency. Rotating machines are classified according to size as fractional-horsepower, integral-horsepower (up to 500 hp at 3600 rpm for ac and 1.25 hp/rpm for dc), and large machines. NEMA standards specify the following powers: 1, 1.25, 1.5, 2, 2.5, 3, 4, 5, 6, 8, 10, 12.5, 16, 20, 25, 30, and 40 mhp, and $\frac{1}{20}$, $\frac{1}{15}$, $\frac{1}{12}$, $\frac{1}{10}$, $\frac{1}{8}$, $\frac{1}{6}$, $\frac{1}{5}$, $\frac{1}{4}$, $\frac{1}{3}$, $\frac{1}{2}$, $\frac{3}{4}$, and 1 hp for fractional-horsepower ac motors, and 0.5, 0.75, 1, 1.5, 2, 3, 5, 7.5, 10, 15, 20, 25, 30, 40, 50, 60, 75, 100, 125, 150, 200, 250, 300, 350, 400, 450, and 500 hp for integral-horsepower ac motors. It can be seen that in each class the ratio between two neighboring power ratings decreases with increasing size. This trend continues with large machines, which include 60 ratings with ratios varying from 1.25 at the 100-hp level to 0.9 at the 100,000-hp level. The synchronous speed is dictated by the number of pole pairs, p, which typically reaches 4 in fractional-horsepower motors, 7 in integral-horsepower motors, 16 in large motors, and 26 in salient-pole synchronous generators. Standard voltages are as follows: for universal and single-phase motors, 115 and 230 V at 60 Hz and 110 and 220 V at 50 Hz; for three-phase motors, 115, 200, 230, 460, 575, 2300, 4000, 4600, 6600, and 13,200 at 60 Hz and 220 and 380 V at 50 Hz. Similar ratings hold for dc machines, but the base speed of large industrial motors can be as low as 50 rpm.

The mounting dimensions have been standardized nationally in order to ensure interchangeability of machines of different manufacture. Standard NEMA frames are denoted as follows. In fractional-horsepower machines the frame number represents the height H from centerline of shaft to bottom of feet in inches multiplied by 16; in integral-horsepower machines the first two digits of the frame number are equal to $4H$, while the subsequent digits are related to the distance between centerlines of mounting holes in feet or in the base of the machine (side view). Following the frame numbers are letters denoting the type of mounting or the letter T denoting that the frame conforms with dimensions standards set by NEMA in 1964.

Rotating machines are designated according to their application as generators or motors and as general, definite, or special-purpose machines. They are designated according to environmental protection as open machines (drip-proof, splashproof, semiguarded, guarded, drip-proof guarded, and externally and pipe ventilated) or totally enclosed machines (nonventilated, fan-cooled, waterproof, pipe ventilated, water or oil cooled, water or oil–air cooled, air-to-air cooled, fan-cooled guarded, and air over machine). Special strengthening of the frames and special fitting are required to make the totally enclosed machines "explosionproof" and "dust–ignitionproof." The classification according to type includes squirrel-cage and wound-rotor induction, synchronous, and commutator ac machines and series, shunt, and compound-wound dc machines.

Standard designs are also classified according to performance. In the case of generators, this is assessed in terms of the voltage regulation between open-circuit and rated conditions and of the ratio of field currents in open and short-circuit conditions. In the case of motors, significant parameters are critical torques, such as starting or locked-rotor, pull-in, and pull-out torques for synchronous motors, and locked-rotor, pull-up or minimum, and breakdown or maximum torques for induction motors. Typical torque–speed and current–speed curves for various NEMA design classes of squirrel-cage induction motors are shown in Figs. 3a and 3b, respectively. The class A motor is the basic design in sizes below 7.5 and above 200 hp. Class B is characterized by normal starting torque, low starting current, and low slip. Class C has a double-cage rotor and produces a high starting torque at low starting current. Class D uses a single-cage rotor with high-resistance bars, produces an even higher starting torque with low starting current, but operates at high slip and consequently has a low running efficilency. Typical efficiency and power factor versus load curves for four-pole design B motors are shown in Figs. 4a and 4b, respectively, and typical full-load power factor versus horsepower rating curves are shown in Fig. 5.

Synchronous motors have higher efficiencies than induction motors of the equivalent rating. When efficiency is the primary consideration in the selection, the standard power factor is unity. When, instead, the primary consideration is compensation of the power factor of the system, the power factor is 0.8 leading (overexcited). This is also a standard power factor for synchronous generators.

Aircraft system characteristics are established by the requirements of specification MIL-G-21480A/AS for the generator and MIL-E-23001B for the power conditioner.

E. Testing and Maintenance

Rotating machines are subjected to spot checks in the smaller sizes and individual checks in the larger sizes, by the manufacturers. Compliance with the standards is

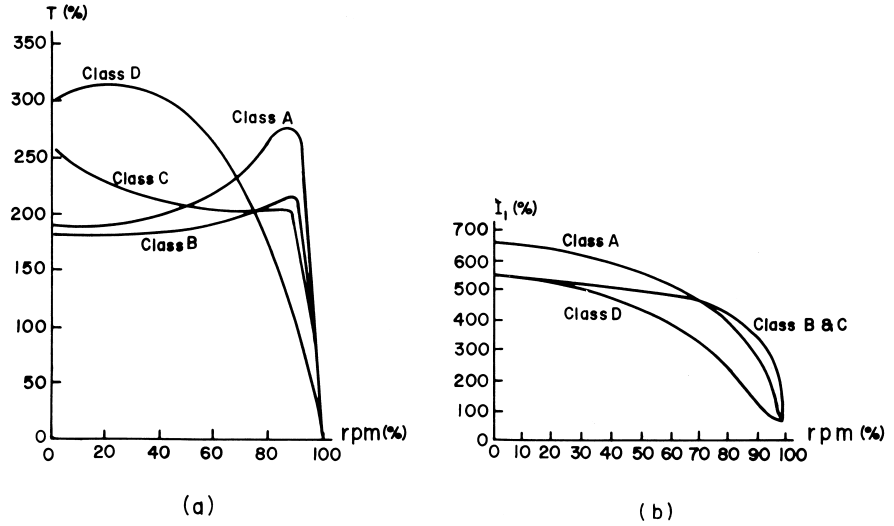


FIGURE 3 Typical characteristics of squirrel-cage induction motors. (a) Torque–speed curves; (b) current–speed curves.

determined by recommended test procedures, such as ANSI/IEEE St. 112-1978 for Polyphase Induction Motors and Generators, IEEE Publ. No. 115 and ANSI C50.10-1965 for Synchronous Machines, and IEEE St. 11-1980 for Rotating Electrical Machinery for Rail and Road Vehicles. Routine tests consist of measurements of winding resistances, high-potential tests at twice the rated voltage plus 1000 V; no-load readings of current, power, and nominal speed at rated voltage and frequency in induction motors; and measurement of open-circuit voltage ratio in wound-rotor induction motors. Specific tests include measurements of; (1) the acoustic noise, which, when certain frequencies are omitted to match the frequency response of the human ear (A network), should not exceed ~120 dB referred to 10^{-12} W, (2) the critical torques in motors, and (3) the harmonic content of the voltage wave in ac

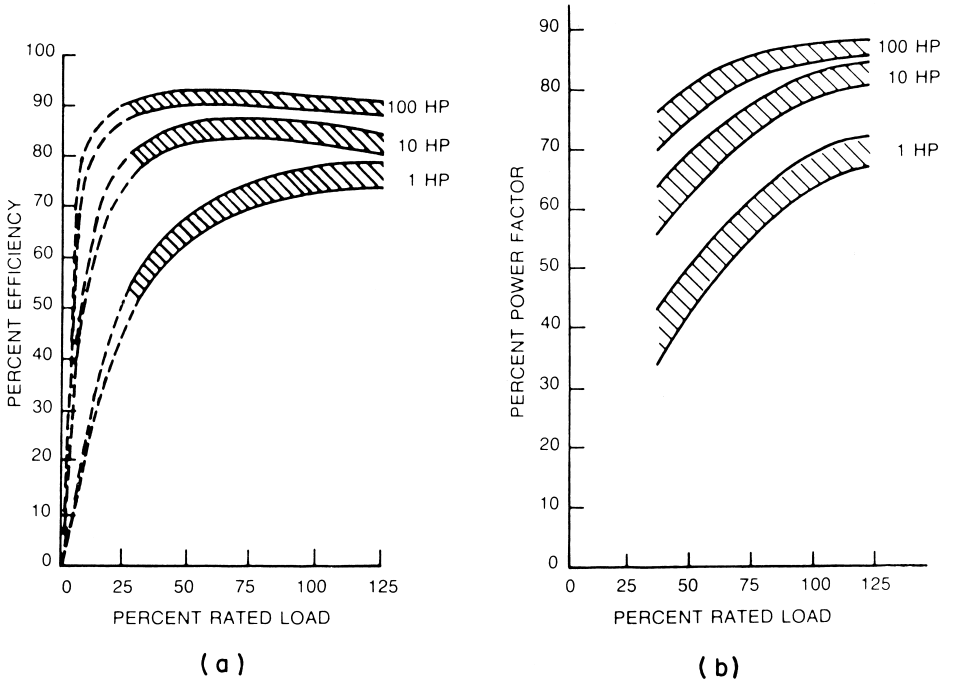


FIGURE 4 Typical efficiency (a) and power factor (b) versus load curves.

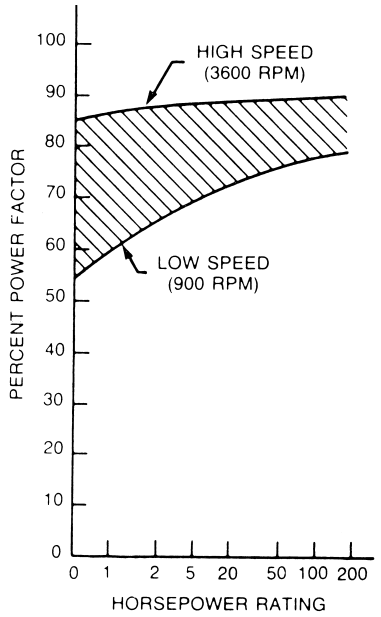


FIGURE 5 Typical full-load power factor versus horsepower rating curves.

generators. The latter is specified in terms of two factors, which are defined in ANSI C42.10 with limiting values given in ANSI C50.12, C50.13, and C50.14 as follows:

1. The deviation factor is the maximum difference between the ordinates of the open-circuit terminal voltage wave and of the equivalent sine wave divided by the amplitude of the equivalent sine wave. Its limiting values are of the order of 0.1.
2. The telephone influence factor (TIF) is the sum of all the harmonics in the voltage wave weighted to reflect their relative objectional effect on telephone communications. The weighing factor ranges from 0.5 at 60 Hz to a peak of 10,600 at 2580 Hz. The limiting values of TIF are of the order of 100.

There are a number of test methods for determining the efficiency of rotating machines. Because efficiencies are relatively high—in large sizes and in special-design small sizes they exceed 90%—very accurate measurements of the electrical quantities, torque, and speed are needed. A common method is to couple a motor and a generator in a set called the dynamometer and to measure the power input and output. Alternatively, one can assess the individual components of the losses. These are discussed in detail in Section III. The recent concern with energy conservation has prompted NEMA to improve the evaluation of efficiency in polyphase induction motors between 1 and 125 hp. A new standard MG1-12.53 has been adopted. Part (a) deals with testing procedures and aims at determining the efficiency with precision better than 1%. Part (b) deals with the variances that result from materials, components, and manufacturing limitations. It establishes a table defining a nominal full-load efficiency, which is the average value in a large population of motors of the same design, and a minimum value, which is based on an increment of 20% in the losses.

The losses not only are an important factor in the economic evaluation process, but also lead to a thermal limitation in the rating of the motor. In general, the lifetime of the insulation decays exponentially with the temperature rise and, in the range of practical interest, it is reduced by about one-half with every increase in temperature of $\sim$10 K. As a result, standards establish that the temperature rise over a 40°C ambient temperature, when measured by the increase in resistance of the winding or by thermocouples, should not exceed prescribed limits. Typical values are as follows: 60 K for class A insulation, which comprises organic materials; 80 K for class B insulation, which comprises inorganic materials and synthetic films; 105 K for class F, which comprises materials similar to those of class B but with higher-temperature binders; and 125 K for class H, which includes silicon elastometers, as well as inorganic materials (MGl-12.41, 12.42, 21.40, 22.40, 23.40, 24.40, 1.65). The stated temperature rises ensure a life expectancy of 20,000–40,000 hr.

The allowable temperature rise defines the rating of the machine. The most common is the continuous rating defining the output that can be carried indefinitely. For intermittent, periodic, or varying duty a machine may be given a short-time rating defining the load that can be carried for a specific time. Standard periods for short-time ratings depend on the thermal time constant and therefore on the size of the machine. They are as follows: 5 min up to 39 hp, 15 min up to 50 hp, 30 min up to 60 hp, and 60 min up to 70 hp. The short-time load test commences only when the windings and other parts of the machine are within 5 K of the ambient temperature at the time of the starting of the test (MG1-10.35). The permissible temperature rises quoted above apply to open, general-purpose motors for which service conditions are unknown. Higher temperature rises are allowed for special-purpose motors that are designed with either operating characteristics or mechanical construction or both for a particular application. As a way to compensate for the lower rise, general-purpose motors are allowed a service factor of 1.15 when operated at rated voltage. Service factor is a multiplier that, when applied to the rated output, indicates a permissible loading that may be carried under the conditions specified for the service factor (MG1-14.35).

In general, electrical machines require much less maintenance than other types of machinery. The insulation resistance to ground is a good index of the state of the

windings and should be checked periodically. Its value in megohms should exceed 1 MΩ plus the rated machine potential in kilovolts. The only parts requiring maintenance are surfaces that are in relative motion, such as bearing surfaces. Sliding contacts that carry electrical currents, such as the contacts between brushes and slip rings and commutators, operate at higher temperatures and may be subjected to pitting due to sparks. As a result they may require more frequent maintenance. This consists of cleaning and blowing out the conducting dust resulting from brush wear, adjusting and replacing brushes, and polishing, grinding, and turning slip rings and commutators. Whether the mica between commutator segments should be left flush with the copper surface or be recessed (i.e., undercut) is debatable. In any case, properly designed rotating machines can operate for thousands of hours without requiring maintenance.

II. PARAMETERIZATION OF PERFORMANCE

Once the desired performance of the rotating machine has been formulated in terms of a set of specifications that conform to the prevailing standards, one must find the values of the model parameteres that satisfy those specifications. In terms of the network model, this means, primarily, establishing a relation between the inductances and the critical torques in motors, and regulation and short-circuit current ratios in generators. In the case of motors fed at constant voltage, the specification that is most important with regard to the overall stability of the drive is the overload capability (OC); that is, the ratio of maximum torque $T_{\max}$ to rated torque T_{R}.

A. Evaluation of Inductances

If one expresses all quantities in per unit of their rated values, the overload capability of induction motors is, according to Section IV.C of Electric Machines, Rotating, Construction and Theory,

$$\mathrm{OC} = \frac{T_{\max}}{T_{\mathrm{R}}} \sim \frac{1/(2x_{\mathrm{sc}})}{\cos\phi_{1\mathrm{R}}}. \tag{1}$$

This relation determines the values of $x_1 = x_2 = x_{\mathrm{sc}}/2$ corresponding to the specified values of OC and $\cos\phi_{1\mathrm{R}}$. The magnetizing reactance x_{m} can then be determined with the help of the phasor diagram of Fig. 28 of Electric Machines, Rotating, Construction and Theory. One observes that the per-unit magnetizing current is

$$|i_\mu| = 1/(x_{\mathrm{m}} + x_1) = |\sin\phi_{1\mathrm{R}}| - |i_2 \sin\phi_{2\mathrm{R}}|, \tag{2}$$

and since

$$\begin{aligned} |i_2 \cos\phi_{2\mathrm{R}}| &= |\cos\phi_{1\mathrm{R}}|, \\ |i_\mu| &= |\sin\phi_{1\mathrm{R}}| - |\tan\phi_{2\mathrm{R}}| \cos\phi_{1\mathrm{R}}, \end{aligned} \tag{3}$$

where $|\tan\phi_{2\mathrm{R}}| = x_{\mathrm{sc}}/(r_2/S_{\mathrm{R}}) \sim x_{\mathrm{sc}}$.

For synchronous motors and neglecting the small contribution of the variable reluctance if the poles are salient, one obtains from Eq. (56) of Electric Machines, Rotating, Construction and Theory

$$\begin{aligned} \mathrm{OC} &= \frac{i_{\mathrm{f}}}{\cos\phi_{\mathrm{aR}}} \\ &= \frac{1}{\cos\phi_{\mathrm{aR}}} \sqrt{i_\mu^2 - 2|i_\mu| \sin\phi_{\mathrm{aR}} + 1}, \end{aligned}$$

where

$$|i_\mu| = \frac{1}{x_{\mathrm{m}}(1 + x_{\mathrm{a}}/x_{\mathrm{m}})}, \tag{4}$$

i_{f} is the per-unit field current, and $x_{\mathrm{a}}/x_{\mathrm{m}}$ can be assumed to be approximately equal to 0.1. Similar relations can be obtained for synchronous generators where the regulation is

$$\mathrm{Reg} = (V_0 - V_{\mathrm{R}})/V_{\mathrm{R}} = x_{\mathrm{m}} i_{\mathrm{f}} - 1$$

or

$$\sqrt{\left(\frac{1}{1 + x_{\mathrm{a}}/x_{\mathrm{m}}}\right)^2 - 2\frac{x_{\mathrm{m}} \sin\phi_{\mathrm{aR}}}{1 + x_{\mathrm{a}}/x_{\mathrm{m}}} + x_{\mathrm{m}}^2 - 1} \tag{5}$$

and the short-circuit ratio is

$$\mathrm{SCR} = |i_\mu|. \tag{6}$$

The conditions imposed by the performance specifications on the design of machines operating at adjustable frequency through the interposition of a power conditioner are much less stringent. The dc machines belong to this group. They are essentially synchronous machines in which the frequency is controlled by a closed, feedback loop that senses the position of the brushes and therefore that of the poles. Thus, they differ from adjustable-frequency synchronous-machine drives only with respect to the power conditioner, which is mechanical instead instead of electrical.

Up to this point the rotating machine has been considered as a block in the system diagram, or a "black box." The next step in the design is to provide the details of its physical realization in terms of topology, geometry, and materials.

B. Field Model

An electrical machine consists of interlinked electric and magnetic circuits. Their dimensions are dictated by the

permissible densities of current and magnetic flux, respectively. The current density is limited by Joule loss considerations, and the flux density by consideration of the saturation of the ferromagnetic materials, as well as hysteresis and eddy current losses. Since current and flux densities are field quantities as opposed to their corresponding integral quantities, the design differs from the performance analysis, which is usually based on integral or terminal quantities and lumped circuit parameters.

For the purpose of design it is expedient to introduce an idealized model of the machine that retains only its salient features. First is the air gap, which is required so that relative motion can take place between stator and rotor and in which the electromechanical power conversion process takes place according to the field concept. The air gap is bounded by two sets of conductors, which are attached to ferromagnetic structures. For the sake of simplicity the length g of the air gap is assumed to be constant; that is, pole saliency and slots are disregarded. The actual windings are replaced by two layers of active conductive material of equivalent thickness a, and the ferromagnetic structures are assumed to have infinite permeability. Finally, the curvature of the air gap is neglected. The resulting model is shown in Fig. 6, where D is the bore diameter, L the effective width of the lamination stack, $\mathbf{v}$ the relative velocity, and $\mathbf{K}$ the surface current density, which is related to the volume current density $\mathbf{J}$ as

$$\bar{\mathbf{K}} = k_{\mathrm{dp}} a \bar{\mathbf{J}},$$

where k_{dp} is a winding factor of the order of 0.9 that accounts for the distribution of the actual winding in discrete slots and for the chording of the coil pitch.

If one adopts the convention that the field vectors are oriented as shown in Fig. 7 and focuses attention on a unit surface of the active conductor that moves with velocity v relative to the observer, the equations describing its electrical and mechanical performance reduce to two scalar relationships.

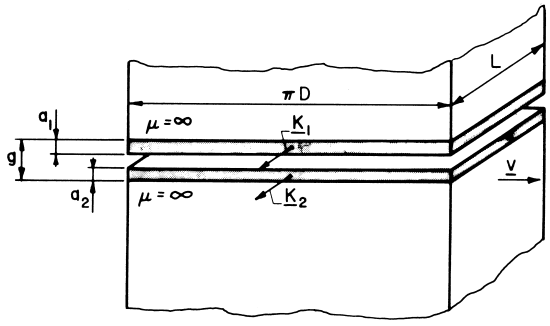


FIGURE 6 Geometry of electromechanical interaction region in developed form.

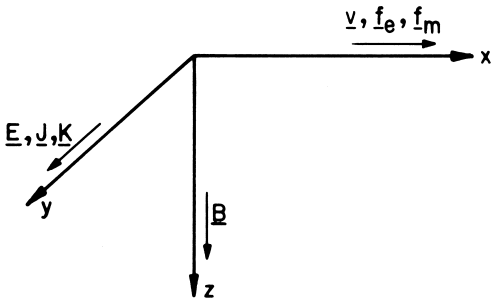


FIGURE 7 Orientation of electrical and mechanical vectors.

One is Ohm's law, which, for a conductor seen to be moving with velocity v in a magnetic field of flux density B, is

$$E - vB = J/\gamma = K/a\gamma k_{\mathrm{dp}}. \tag{7}$$

The other relationship is Newton's law,

$$a\xi(dv/dt) = f_{\mathrm{e}} + f_{\mathrm{m}}, \tag{8}$$

where ξ is the mass density of the conductor, f_m the mechanical force per unit surface, and $f_{\mathrm{e}} = KB$, the electromagnetic force per unit surface. The terminal quantities can be recovered by integrating each term of these equations over the surface of the conductor. Thus, assuming a sinusoidal distribution of the field quantities along the periphery of the gap and introducing phasor notation, one obtains for the electromotive force of one-phase winding

$$-\bar{e}_{\mathrm{ph}} = 2N_{\mathrm{eff}} v L \bar{B} = j\omega_{\mathrm{e}} N_{\mathrm{eff}} \bar{\Phi}, \tag{9}$$

where

$$\begin{aligned} \omega_{\mathrm{e}} &= (2p/D)v, \\ \bar{\Phi} &= -j\frac{DL}{p}\bar{B}, \\ N_{\mathrm{eff}} &= k_{\mathrm{dp}} N, \end{aligned} \tag{9}$$

and p is the number of pole pairs, Φ the flux per pole, and k_{dp} a winding factor relating the number of turns per phase N to that of an equivalent concentrated full-pitched coil N_{eff}. The phase current is

$$\bar{I}_{\mathrm{ph}} = (2mN_{\mathrm{eff}}/\pi D)^{-1}\bar{K}, \tag{10}$$

where m is the number of phases. Similarly, integration of f_{e} yields

$$\begin{aligned} T_{\mathrm{e}} &= \frac{D}{2}(\pi DL)\langle f_{\mathrm{e}}\rangle = 2\left(\frac{\pi D^2 L}{4}\right) K B \cos(\bar{K}\bar{B}) \\ &= \frac{mp}{\omega_{\mathrm{e}}}(-e) I_{\mathrm{ph}} \cos[(-\bar{e})\bar{I}_{\mathrm{ph}}], \end{aligned} \tag{11}$$

where the angled brackets denote average value. This equation puts in evidence the proportionality of the torque to the bore volume $\pi D^2 L/4$. The design problem of determining the dimensions of the bore volume for a specified torque is thus reduced to the assignment of appropriate values to the components of $\langle f_e \rangle$, the surface current density K, and the flux density B. Since they are representative of the stresses on the materials, they are called the electric and magnetic loadings, respectively.

The other dimension of the air gap, its length g, is obtained by applying Ampere's law to a loop enclosing both conductors. The result is

$$\frac{\bar{B}}{\mu_0} g = j \frac{D}{2p} \bar{K}_\mu, \tag{12}$$

where $\bar{K}_\mu = \bar{K}_1 + \bar{K}_2 =$ magnetizing surface current density and μ_0 is the permeability of vacuum. Multiplying both sides by v and integrating over the surface of the conductors yields the terminal relation

$$(-e_{ph}) = j\omega_e M \bar{I}_\mu, \tag{13}$$

where $M = \mu_0 m N_{eff}^2 DL/\pi pg$, the magnetizing inductance, and $\bar{I}_\mu$ is the magnetizing current. Also, multiplying Eq. (12) by μ_0/g and dividing it by the rated value K_R, one obtains an important relation between B and K. This is

$$\frac{B}{K_R} = i_\mu \frac{\mu_0}{2p(g/D)}, \tag{14}$$

where i_μ, the per-unit value of the magnetizing currents, is seen to play a pivotal role as an index of performance.

Equations (9) and (10) establish a one-to-one correspondence between terminal and field quantities and provide the means by which the representation of the machine performance can be switched from a network model to a field one. This is illustrated by the phasor diagrams of Fig. 8 in the most general case of a salient-pole machine. The coefficients C_f, C_d, and C_q are the ratios of the peak values of the fundamental harmonics for field, direct, and quadrature excitations, respectively, to those of the corresponding actual B waves.

C. Magnetic and Electric Loadings

The magnetic flux density B, which is relevant to the electromechanical power conversion process, is the effective or rms value of the radial component of B at the air gap. Except for special cases, such as superconducting field excitation and printed windings, this value is determined by the characteristics of the ferromagnetic structure into

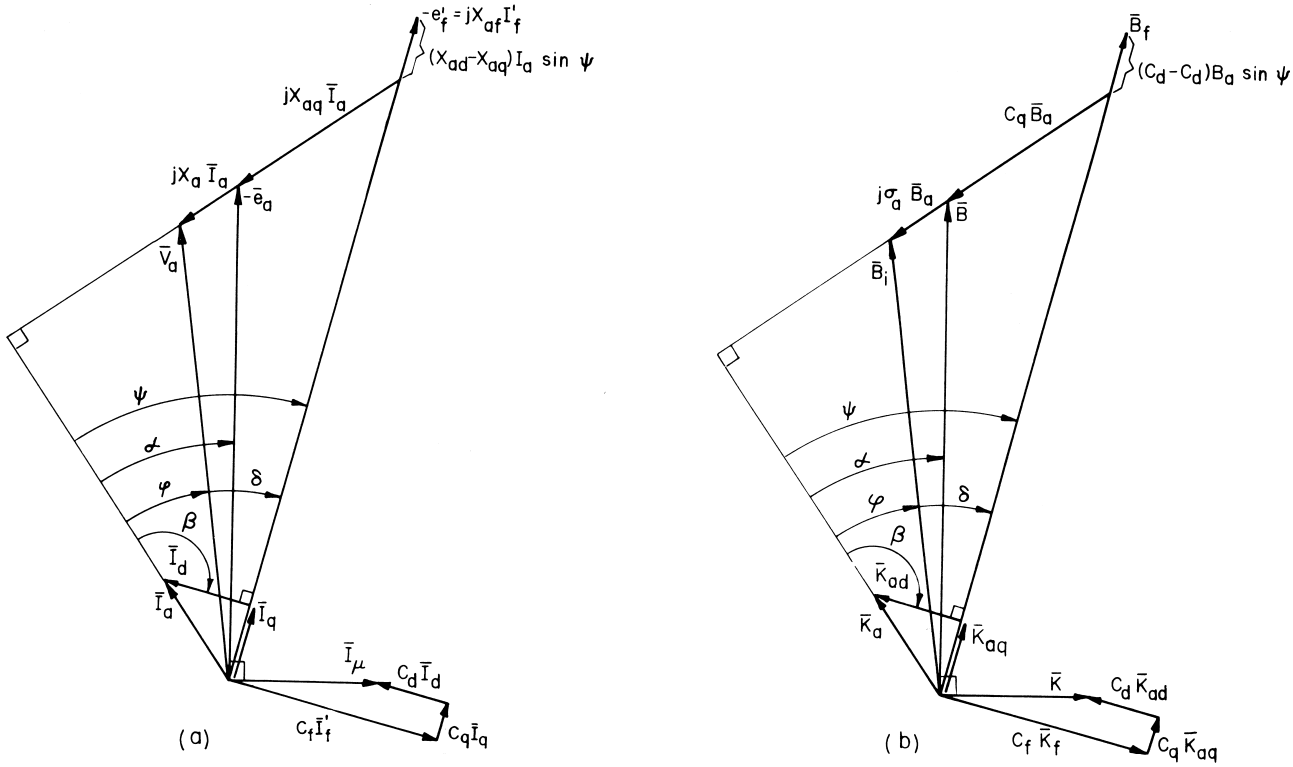


FIGURE 8 Phasor diagram of salient-pole machine in terms of (a) terminal quantities and (b) field quantities.

which the conductors are embedded. This consists of a core or yoke, which provides both physical integrity and a path for the magnetic flux, and a slotted portion adjacent to the air gap, which accommodates the active conductors. The slot dimensions represent a compromise between the conflicting requirements of conduction of current in the copper and of flux in the iron teeth. It turns out that there is a value of the ratio between the slot and tooth widths that minimizes the volume and weight of the ferromagnetic structure. This value is unity, so that the tooth width should be half the slot pitch. If one assumes that all the flux crossing the air gap passes through the iron teeth, the flux density B_t in the teeth is related to B as

$$B = \tfrac{1}{2} B_t. \tag{15}$$

When the tooth is driven too far into saturation, the magnetizing current and the iron loss rise steeply to unacceptable values and the wave shape of the flux distribution is deformed. Also, part of the flux is diverted to the slot, which causes an increase in the additional copper losses and the transfer of the force from the tooth to the conductor. As a result the electrical insulation is stressed mechanically. For these reasons rms values of B_t in excess of 1.4 T are not recommended and B is practically limited to about 0.7 T.

In contrast to the magnetic loading, there is no optimal value for the electric loading. Its value is determined primarily by thermal considerations, which are related to the losses in the machine.

III. LOSSES

The power dissipated in the machine determines its efficiency and its cooling requirements. Heat dissipation comprises the following components: iron, stray-load, copper, and mechanical losses.

A. Iron Losses

Losses occur in the iron when the magnetic field is alternating because of the hysteresis in the B–H magnetization cycle and because of eddy currents. The losses in watts per kilogram mass of lamination can be expressed as

$$p_{Fe} = af B_p^2 + b(\Delta f B_p)^2, \tag{16}$$

where f is the frequency, Δ the thickness of the laminations in millimeters, and B_p the peak value of B. The first term accounts for the hysteresis loss and the second for the eddy current loss. The constants a and b depend on the properties of the material. Typically, for laminations used in motors such as A1S1 grade M-15, the constants are

$$a = 0.019; \qquad b = 6.2 \times 10^{-4}. \tag{17}$$

Because of the strong dependence on frequency, the losses based on the fundamental wave are significantly augmented by the effects of higher harmonics in the currents and due to the presence of slots. Additional losses are caused by flaws in manufacturing. As a result, the iron loss based on the fundamental should be increased at least by 50–100%. Iron losses usually amount to about 20–30% of the total losses.

B. Stray-Load Losses

Stray-load losses appear only with the flow of load currents. They include losses at the surface of the poles and teeth, losses in the teeth due to flux pulsations, and losses in the end bells. They usually represent ~10% of the total loss.

C. Copper Losses

The Joule losses in the windings go by this name, because copper is the material most commonly used. The Joule loss per unit volume of the conductor is

$$p_{Cu} = J^2/\gamma = K^2/a^2\gamma k_{dp}^2, \tag{18}$$

and for the whole winding it is

$$P_{Cu} = \pi D l_{co} a \frac{J^2}{\gamma} = \pi D L \left(\frac{l_{co}}{L}\right) \frac{K^2}{a\gamma k_{dp}^2},$$

where

$$l_{co} \sim \left(1 + \frac{2D}{pL}\right) L \sim 2.27 p^{-0.16} L \tag{19}$$

denotes the average length of the conductors and γ is the electrical conductivity of the material calculated at two-thirds the maximum temperature allowed by the insulation class. Additional losses due to skin and proximity effects and harmonics may augment the dc loss by more than 50%. The copper losses amount to ~35% of the total loss in the winding carrying the magnetizing current and ~25% in the other winding.

A good starting guess for the copper loss in the conductor, which is critically stressed thermally, is

$$P_{Cu} = 0.42 P_i^{0.75}, \tag{20}$$

where the ideal power P_i is given by the rated power P_R divided by the product of the efficiency η_R and power factor $\cos \phi_R$.

D. Mechanical Losses

Mechanical losses are due to friction and windage and include the power required by the mover of the cooling

medium, such as the fan, blower, or pump. The friction and windage losses in watts can be estimated as

$$P_{\mathrm{F+W}} = 5 \times 10^{-5}(\mathrm{rpm})^{0.7} P_{\mathrm{R}}, \tag{21}$$

where P_{R} is the rated power in watts. The power required by the fluid mover is given by

$$P_{\mathrm{fm}} = (Q\Delta p)/\eta_{\mathrm{m}} = Q\xi g H/\eta_{\mathrm{m}}, \tag{22}$$

where $Q = \Sigma P_{\mathrm{loss}}/c\vartheta_{\mathrm{c}}$ is the volumetric flow rate, Δp the pressure differential, g the acceleration of gravity, η_{m} the mechanical efficiency, c the specific heat of the coolant per unit volume, and ϑ_{c} the temperature rise of the coolant. The mechanical losses typically amount to ~10% of the total loss.

E. Evaluation of the Resistances

From the estimation of the losses one can determine the efficiency as

$$\eta = P_{\mathrm{R}}/(P_{\mathrm{R}} + \Sigma P_{\mathrm{loss}} + P_{\mathrm{fm}}). \tag{23}$$

Also, the partition of the losses determines the value of the resistive circuit parameters. Accordingly the resistance associated with the excitation current is

$$R_{\mathrm{e}} = me^2 \big/ P_{\mathrm{Fe}}, \tag{24}$$

and the resistance of the nth winding is:

$$R_{\mathrm{n}} = P_{\mathrm{cu}\,n} \big/ m I_n^2. \tag{25}$$

This completes the determination of the lumped circuit parameters.

IV. HEATING AND COOLING

The energy dissipated in a part of the machine in the interval of time dt is equal to the sum of two components: (1) the quantity of heat required to raise its temperature by $d\vartheta$, and (2) the heat transferred to the surroundings, or

$$P_{\mathrm{diss}}\, dt = C\, d\vartheta + (1/R)\vartheta\, dt, \tag{26}$$

where $C\,[\mathrm{JK}^{-1}]$ is the thermal capacity and $R\,[\mathrm{KW}^{-1}]$ the thermal resistance. The solution of this differential equation is

$$\vartheta(t) = \vartheta_{\infty}(1 - e^{-t \neq \tau_{\mathrm{T}}}) + \vartheta(0) e^{-t \neq \tau_{\mathrm{T}}}, \tag{27}$$

where

$$\vartheta_{\infty} = R P_{\mathrm{diss}}$$

and

$$\tau_{\mathrm{T}} = RC.$$

A. Thermal Transients

The curve describing the temperature rise versus time can be approximated by two asymptotes. For $t \ll \tau_T$ the temperature rise is small and the transfer of heat to the surroundings can be neglected (adiabatic approximation). This situation prevails under transient conditions and is particularly relevant to machines operating with intermittent duty cycles and during starting of high-inertia drives. Considering a unit volume of the conductor and denoting its specific heat by c, one finds that the energy dissipated over a period of time T is

$$w_{\mathrm{diss}} = \int_0^T p_{\mathrm{diss}}\, dt = \int_0^T \frac{J^2}{\gamma}\, dt = c\vartheta. \tag{28}$$

When the machine is supplied at constant current, J is constant and the integration gives the maximum allowable value of J for a specified ϑ_{max} as

$$J_{\mathrm{max}} = \sqrt{c\gamma\vartheta_{\mathrm{max}}/T}. \tag{29}$$

In the more common case, in which the voltage remains constant, the electric field E and the induction B can be considered constant. The electric energy input is

$$w_{\mathrm{elec}} = \int_0^T EJ\, dt = E\int_0^T J\, dt, \tag{30}$$

and using Eq. (8) and $f_{\mathrm{m}} = 0$ yields

$$w_{\mathrm{elec}} = E\int_0^T \frac{1}{B}\xi\frac{dv}{dt}\, dt = \frac{E}{B}\xi[v(T) - v(0)]. \tag{31}$$

Of this input, the share absorbed in kinetic form is

$$w_{\mathrm{kin}} = \tfrac{1}{2}\xi[v^2(T) - v^2(0)]. \tag{32}$$

The remainder is the energy dissipated:

$$w_{\mathrm{diss}} = \xi[v(T) - v(0)]\big\{E/B - \tfrac{1}{2}[v(T) + v(0)]\big\}. \tag{33}$$

During starting, $v(0) = 0$ and $v(T)$ approaches the asymptotic velocity E/B, so that

$$w_{\mathrm{diss}} = w_{\mathrm{kin}} = c\vartheta. \tag{34}$$

This equality between the energy stored in kinetic form and that dissipated as heat is maintained even when the inertia of the masses attached to the rotating conductor is taken into account. If Θ_{tot} represents the total inertia and Θ_{c} the inertia of the active conductor, their ratio cannot exceed

$$\Theta_{\mathrm{tot}}/\Theta_{\mathrm{c}} = c\vartheta_{\mathrm{max}} \big/ \tfrac{1}{2}\xi v^2(T). \tag{35}$$

Since this relation does not depend on machine type, electrical quantities and parameters, or acceleration time T, it represents a general limitation of high-inertia drives.

The thermal stress is much higher when the motor is brought to a standstill by reversing the sequence of the

phases (plugging). Letting $v(T) = 0$ and $v(0) = -E/B$, one finds that the energy dissipated into heat equals three times the energy initially stored in kinetic form.

B. Thermal Steady State

The other asymptote of the temperature curve is given by the steady-state condition, when the power dissipated equals the heat transferred to the environment by conduction, convection, and radiation. Conduction is the dominant mechanism in metallic materials, and it is so effective that their temperature may be considered uniform. Convection occurs at the interface between a solid and a fluid, in this case the coolant. The heat flow rate $P\,[W]$ resulting from convection is related to the temperature rise $\vartheta\,[K]$ of the solid over the coolant as

$$P = \vartheta/R = hS\vartheta, \tag{36}$$

where $h\,[\mathrm{W\,K^{-1}m^{-2}}]$ is the coefficient of heat transfer and $S\,[\mathrm{m^2}]$ is the surface of the interface between the solid and the coolant. Radiation by electromagnetic waves is proportional to the difference in the fourth powers of the absolute temperatures. In electrical machines, where this difference is small, the heat-flow rate can be assumed to be proportional to the temperature rise, and the radiation and convection coefficients are lumped together into an equivalent h.

With these approximations and making use of Eq. (36), one can reduce the machine to a discrete network of parts represented by heat sources and resistances and determine the temperature rises by applying a lumped-parameter analysis. This approach is exemplified by considering the common case of a radially cooled machine (Fig. 9a). The two fans attached to the rotor push the cool air through the end windings and along the surface of the lamination packet. The major sources of heat are usually the losses in the copper and the losses in the iron. In general, the transfer of heat between the embedded portion of the copper and the iron is not very significant, because of the conductor and slot insulation and the poor contact caused by manufacturing tolerances. Therefore, one can suppose, as a worst-case working assumption that leads to conservative designs, that all the heat P_{Cu} developed in the winding is transferred to the cooling air through the end winding resistance R_{ew}. Similarly, the iron losses are transferred through the iron resistance R_{Fe}. One thus obtains the equivalent circuit of Fig. 9b. Of particular interest are the branches representing the windings. If one denotes by $P_{\mathrm{s\text{-}diss}}$ the power dissipated per unit surface of the active conductor and introduces an overall transfer coefficient h, Eqs. (18) and (19) yield

$$p_{\mathrm{s\text{-}diss}} = \frac{P_{\mathrm{Cu}}(L/l_{\mathrm{co}})}{\pi DL} = \frac{(K/k_{\mathrm{dp}})^2}{\gamma a} = \frac{K/k_{\mathrm{dp}}}{\gamma} J = h\vartheta_{\mathrm{w}}. \tag{37}$$

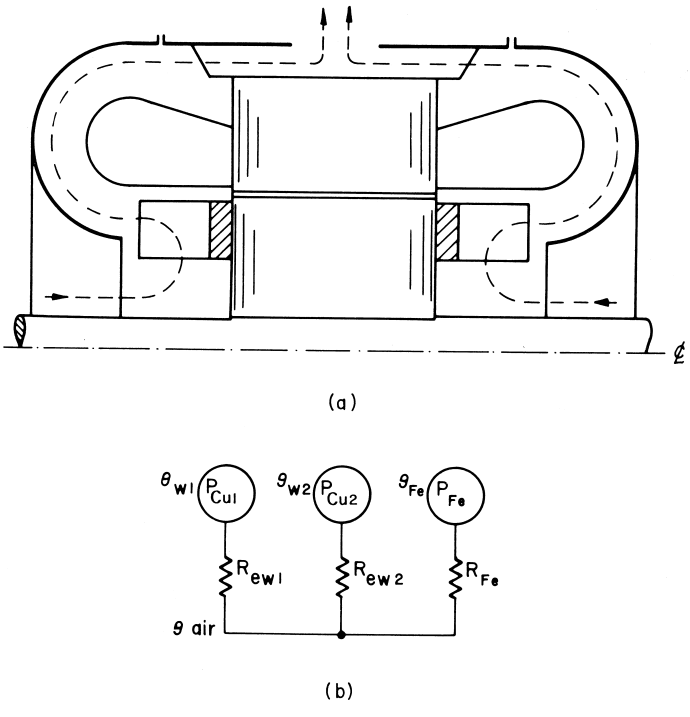


FIGURE 9 Sketch of airflow (a) and thermal equivalent circuit (b) for drip-proof, squirrel-cage motor with internal ventilation.

Typical approximate values of h are as follows:

30–70 for internal air ventilation,
70–100 for forced air draft,
70–150 for hydrogen and liquid cooling over the conductor,
100–200 for heat pipes, and
200–250 for liquid cooling inside the conductor.

In the most common case of internal air ventilation, one finds that

$$h \sim 16(Df/p)^{0.6}. \tag{38}$$

V. SCALING LAWS

Once the parameterization of the performance has been completed, it becomes possible to undertake the final step in the design; that is, to relate the specifications to the main dimensions of the machine and to the magnetic and electric loadings.

A. Derivation of Main Dimensions

The heat balance relation [Eq. (37)] can be used to determine the bore diameter as

$$D = \left[\frac{P_{\mathrm{Cu}}(L/l_{\mathrm{co}})}{\pi h \vartheta_{\mathrm{w}}(L/D)}\right]^{1/2}, \tag{39}$$

where a typical value for the bore aspect ratio is

$$\frac{L}{D} \sim \frac{\pi}{2} p^{-2/3}. \tag{40}$$

Making use of Eqs. (19), (20), (39), and (40), one obtains for internally ventilated machines

$$D \sim 0.097 \vartheta_{\mathrm{w}}^{-0.385} P_{\mathrm{i}}^{0.29} p^{0.55} f^{-0.23}. \tag{41}$$

Equation (40) then gives L. Introducing Eq. (39) into Eq. (11) yields

$$K = \frac{p}{f} \frac{P_{\mathrm{i}}}{(P_{\mathrm{Cu}})^{3/2}} \left[\frac{h\vartheta_{\mathrm{w}}}{(L/l_{\mathrm{co}})}\right]^{3/2} \left(\frac{L/D}{\pi}\right)^{1/2} \frac{1}{B}. \tag{42}$$

Equation (37) then yields J and a.

It appears that if B is specified—for instance, by choosing its optimal value according to Eq. (15)—then the main dimensions of the machine and the current densities can be selected solely on the basis of thermal considerations and a torque equation that essentially expresses Biot-Savart's law. However, a selection based on such narrow grounds may not satisfy Eq. (14), as needed in order to meet important performance specifications. By combining Eqs. (14) and (42), one obtains

$$B = \left[\frac{\mu_0 i_\mu}{2f(g/D)}\right]^{1/2} \left(\frac{L/D}{\pi P_{\mathrm{i}}}\right)^{1/4} \left[\frac{h\vartheta_{\mathrm{w}}}{P_{\mathrm{Cu}}(L/l_{\mathrm{co}})P_{\mathrm{i}}}\right]^{3/4}. \tag{43}$$

This equation reveals the dependence of B on i_μ and on g/D.

In general-purpose squirrel-cage motors of NEMA class A, Eq. (3) leads to

$$i_\mu \sim 0.28 P_{\mathrm{i}}^{-0.1} f^{0.15} p^{0.6}. \tag{44}$$

In induction motors, in which good coupling between primary and secondary is essential, the air-gap length is made as small as mechanical considerations allow. As a first approximation one may assume

$$g/D \sim 7.25 \times 10^{-3} P_{\mathrm{i}}^{0.01} p^{-0.5}. \tag{45}$$

Larger gaps are needed in special applications, such as traction motors that are subjected to large accelerations (g's). Introducing Eqs. (44) and (45) into Eq. (43), one can calculate B. If B turns out to be smaller than the optimal value given in Eq. (15), the motor is "thermally limited." In this case B could be raised and, therefore, the motor dimensions reduced by improving the cooling or by switching to a higher class of insulation. If, instead, B turns out to be too large, the motor is "saturation limited." Then, one must reduce ϑ_{w}. This is accomplished by increasing the bore dimensions. A lower ϑ_{w}, however, allows one to switch to a less expensive class of insulation and reduces the copper losses because of better conductivity.

Different considerations hold for synchronous machines. Here the performance requirements of Eqs. (4) and (6) lead to

$$i_\mu \sim 1.27 C_{\mathrm{q}}, \tag{46}$$

where C_{q} is the nondimensional reactance coefficient in the quadrature axis. This value of i_μ is more than twice the value recommended for induction machines by Eq. (44). If the ratio g/D were the same, a synchronous machine with the same magnetic loading B would allow a much smaller electrical loading K_{R} than the induction machine would, and therefore would be much larger for the same developed torque. It follows that synchronous machines should be built with larger gaps than those permitted by purely mechanical considerations, according to Eq. (45). Large air gaps are not as objectionable as in induction motors, because the availability of a second excitation provided by the field permits maintaining an adequate power factor. Moreover, the field winding can tolerate higher electric loading and thermal stresses than the armature, because its lower voltage reduces the insulation requirements.

B. Bulk, Weight, and Cost

Once the electric and magnetic loadings have been selected, all the dimensions of the bore and of the winding are determined. Another dimension that is very important, because it determines the frame size, is D_o, the outside diameter of the laminations. As a first approximation one may assume

$$D_o \sim (1.1 + 0.8/p)D. \tag{47}$$

The outside diameter can also lead to rough estimates of the motor's overall length, weight, and cost. The overall length is approximated as

$$L_o \sim 2.7 \times p^{-0.25} P_R^{-0.05} D_o, \tag{48}$$

the weight is approximated as

$$W \sim 2.63 \times 10^3 D_o^{3.1}, \tag{49}$$

and the original equipment manufacturer's (OEM) cost in 1985 U.S. dollars is approximately \$10/kg.

The design procedure outlined above in Sections V.A and V.B is illustrated by two examples.

C. Numerical Examples

1. Example 1

A general-purpose squirrel-cage motor is designed to develop 250 hp at 1776 rpm when fed at 60 Hz with a power factor of 0.889. Considering internal ventilation with an ambient temperature of 40°C and class F insulation, the temperature rise is, according to Sections I.E and III.C, $\vartheta_w = 75$°C. For the specified speed, the machine is built with four poles ($p = 2$). Introducing into Eq. (41) the ideal power

$$P_i = (250 \times 746)/0.889 = 209{,}786\,\text{W},$$

one obtains

$$D = 0.097 \times 75^{-0.385} \times 209{,}786^{0.29} \times 2^{0.55} \times 60^{-0.23} = 0.366\,\text{m},$$

and from Eq. (40),

$$L = (\pi/2) \times 2^{-2/3} \times 0.366 = 0.363\,\text{m}.$$

This yields an effective bore area

$$\pi DL = \pi \times 0.366 \times 0.363 = 0.417\,\text{m}^2,$$

and according to Eq. (47) an outside diameter

$$D_o = (1.1 + 0.8/2) \times 0.366 = 0.549\,\text{m}.$$

The per-unit magnetizing current is, according to Eq. (44),

$$i_\mu = 0.28 \times 209{,}786^{-0.1} \times 60^{0.15} \times 2^{0.6} = 0.23\,\text{per unit}.$$

The effective air-gap length is, according to Eq. (45),

$$g = 7.25 \times 10^{-3} \times 209{,}786^{0.01} \times 2^{-0.5} \times 0.366 = 2.12 \times 10^{-3}\,\text{m}.$$

The heat dissipated in the primary is, from Eq. (20),

$$P_{Cu} = 0.42 \times 209{,}786^{0.75} = 4117\,\text{W}.$$

Introducing into Eq. (37) the ratio of average conductor length to bore length according to Eq. (19),

$$l_{co}/L = 1 + (2 \times 0.366)/(2 \times 0.363) = 2.01,$$

and the aspect ratio $L/D = 0.988$, one obtains the power dissipated per unit surface

$$p_{s\text{-diss}} = h\vartheta_w = 4117/(2.01 \times 0.417) = 4911\,\text{Wm}^{-2}.$$

and the heat transfer coefficient

$$h = 4911/75 = 65\,\text{WK}^{-1}\text{m}^{-2}.$$

Introducing the appropriate values into Eq. (43), one finds that the effective flux density in the gap is

$$B = \left[\frac{4\pi \times 10^{-7} \times 0.23}{2 \times 60 \times (2.12 \times 10^{-3}/0.366)}\right]^{1/2} \times \left(\frac{0.988}{\pi \times 209{,}786}\right)^{1/4} \left(\frac{4911 \times 2.01}{4117/209{,}786}\right)^{3/4} = 0.426\,\text{T}.$$

According to Eq. (42), the surface current density is

$$K = \frac{2}{60}\frac{209{,}786}{(4117)^{3/2}}(4911 \times 2.01)^{3/2} \times \left(\frac{0.988}{\pi}\right)^{1/2} \frac{1}{0.426} = 34{,}176\,\text{Am}^{-1}.$$

The conductivity of copper at 115°C is $\gamma = 4.185 \times 10^7\ \text{Sm}^{-1}$. Assuming that the stator has 72 slots and that the coils have a $\frac{5}{6}$ throw, the winding factor is $k_{dp} = 0.923$. Introducing these values into Eq. (37), one obtains the volume current density

$$J = \frac{4911 \times 4.185 \times 10^7 \times 0.923}{34{,}176} = 5.55 \times 10^6\,\text{Am}^{-2}$$

and the equivalent thickness of the conductor

$$a = \frac{34{,}176}{0.923 \times 5.55 \times 10^6} = 6.67 \times 10^{-3}\,\text{m}.$$

These data are now compared with those of a machine of recent construction and having the same power and speed ratings. Its data are as follows:

$$D = 0.317\,\text{m}, \qquad L = 0.381\,\text{m},$$
$$a = 7.2 \times 10^{-3}\,\text{m}, \qquad l_{co}/L = 1.85,$$
$$B = 0.51\,\text{T}, \qquad K = 36{,}271\,\text{Am}^{-1},$$
$$J = 5.43 \times 10^{6}\,\text{Am}^{-2}, \qquad P_{Cu} = 3581\,\text{W}.$$

This machine is characterized by a higher aspect ratio, $L/D = 1.2$ instead of 0.988, and a higher effective flux density, $B = 0.51$ T instead of 0.426 T. The selection of a higher aspect ratio was motivated by the desire to fit the machine within a NEMA frame, such as the 447 T one, which has an 11-in. shaft height from base to center.

Also, the specific power dissipation $p_{s\text{-diss}}$ and, therefore, the thermal loading $h\vartheta_w$ are about 4% higher than in the sample calculation, despite the lower copper losses in the primary. Clearly, both machines are thermally limited, because their gap flux densities are much less than 0.7 T. This tends to be the case in large squirrel-cage motors, because power factor considerations prevail over the need for large starting and pull-out torques.

2. Example 2

A synchronous motor is designed to develop 2500 hp at 1200 rpm when fed at 60 Hz with a power factor of unity. Considering internal ventilation with an ambient temperature of 40°C and class H insulation, the temperature rise is, according to Sections I.E and III.C, $\vartheta_w = 90$°C. According to Eq. (41) the bore diameter is

$$D = 0.097 \times 90^{-0.385} \times (2500 \times 746)^{0.29} \times 3^{0.55} \times 60^{-0.23} = 0.806\,\text{m},$$

and from Eq. (40), the effective bore length is

$$L = (\pi/2) \times 3^{-2/3} \times 0.806 = 0.609\,\text{m}.$$

This yields an effective bore area

$$\pi DL = \pi \times 0.806 \times 0.609 = 1.542\,\text{m}^2,$$

and, according to Eq. (47), a ratio of outside diameter to bore diameter

$$D_o/D = (1.1 + 0.8/3) = 1.36.$$

The weight is then, according to Eq. (49),

$$W = 2.63 \times 10^{3} \times (1.36 \times 0.806)^{3.1} = 3496\,\text{kg}.$$

A machine of recent construction and similar performance has the following dimensions:

$$D = 0.914\,\text{m}, \qquad L = 0.381\,\text{m},$$
$$D_o = 1.295\,\text{m}, \qquad W = 3487\,\text{kg}.$$

One observes that the effective bore area is

$$\pi DL = \pi \times 0.914 \times 0.381 = 1.094\,\text{m}^2,$$

which results in a thermal loading that is about 40% higher and that may require forced air draft. The ratio of outside diameter to bore diameter is $D_o/D = 1.295/0.914 = 1.41$, or 3% higher. The almost equal weight, despite the smaller bore area, is due to the selection of a lower aspect ratio L/D.

3. Summary

Examples 1 and 2 show that there is some latitude in the choice of the aspect ratio L/D. Moreover, the selection of temperature rises ϑ_w, which are about two-thirds of the maximum allowable ones, leads to conservative designs, as was intended.

VI. SUMMARY

The tools needed for the design of electrical machines consist of two classes:

1. General relations that develop insights into the structural components of the machine. These insights may guide the development of very unconventional machines, as well as conventional ones.
2. Equations that are empirical or based on judgment. These are derived from conventional designs and are contingent on the material used. Therefore, they have restricted application.

The first class includes the basic electromechanical power conversion relation [Eq. (11)]; the steady-state thermal balance equation [Eq. (37)], which relates the performance specifications regarding ohmic losses to the parameters of the electric circuit; and the air-gap characteristics [Eq. (14)], which relate the performance specifications regarding overload capability in motors and regulation in generators to the parameters of the magnetic circuit. These equations can be modified and applied to linear motion simply by substituting the bore diameter D with $2p\tau/\pi$, where τ is the pole pitch.

The second class includes the dimensional ratios [Eqs. (19), (40), (45), (47), and (48)] and the other suggestions for the design [Eqs. (15), (20), and (38)].

All these relations can be combined to reveal the interrelationships among the various parameters, so that the designer can estimate at a glance the effect of a change in a particular value. Simple scaling laws allow the designer to derive the main dimensions of the ferromagnetic

structure and windings of the machine (D, D_o, L, L_o, g, a, and l_{co}) and its electrical and magnetic stresses (K, J, and B) directly from the performance expressed in terms of power, frequency, number of pole pairs, and per-unit magnetizing current.

Although the basic design, thus obtained, must be checked by verifying all the assumptions on which it was based, before detailed drawings of the machine can be made, it provides adequate information and accuracy for the optimization of the overall drive.

SEE ALSO THE FOLLOWING ARTICLES

ELECTRIC MACHINES, DRIVES • ELECTRIC MACHINES, ROTATING, CONSTRUCTION AND THEORY • MACHINE DESIGN • POWER ELECTRONICS • TRANSFORMERS, ELECTRICAL

BIBLIOGRAPHY

Chapman, S. J. (1999). "Electric Machinery Fundamentals," McGraw-Hill, New York.

Electric Machines and Drives Conference (1997). IEEE International.

(1999). "Electrical Machines, Drives, and Power Systems," 4th ed., Prentice Hall, New York.

(1999). "Electrical Transformers and Rotating Machines," Delmar, Albany, N.Y.

(1999). "Electromechanical Systems, Electric Machines, and Applied Mechatronics," CRC Press, Boca Raton, Fla.

Hamdi, E. S. (1994). "Design of Small Electrical Machines," Wiley, New York.

Hindmarsh, J., and Renfrew, A. (1996). "Electrical Machines and Drives," Butterworth-Heinemann, Newton, Mass.

Naidu, M. S., and Kamakshaiah, S. (1995). "Introduction to Electrical Engineering," Tata McGraw-Hill, New Delhi.

Nasar, S. (1994). "Electric Machines and Power Systems" (Vol. I, Electric Machines), McGraw-Hill (Higher Education), New York.

Rexford, K. (1997). "Electrical Control for Machines," 5th ed., Delmar, Albany, N.Y.



Electric Machines, Drives

Zivan Zabar
Polytechnic University of New York

I. Basic Concepts
II. Performance Features
III. Speed Control

GLOSSARY

Acceleration Rate of change of speed; also, the process of running a motor up to speed.
Chopper Static device for interrupting a current at regular intervals. This reduces the output voltage from a constant-voltage source in a controlled manner.
Commutating reactance Reactance that opposes the transfer of direct current from one winding to another during switchover.
Converter Static device for changing alternating current power to direct current power or vice versa, or from one frequency to another.
Cycloconverter Static device for direct changing of alternating current power at one frequency to alternating current power at a lower frequency.
Inertia Property of a body that causes it to resist changes in its motion.
Inverter Static device that changes direct current power to alternating current power.
Mass Property of a body that determines the acceleration it will have when acted on by a given force.
Rectifier Static device for converting alternating current power to direct current power.
Rheostat Adjustable resistor.
Thyristor Bistable semiconductor device that can be switched at will from the conducting state to the nonconducting state or vice versa.
Transistor Active semiconductor device with three or more terminals.

AN ELECTRIC DRIVE consists of the equipment used for converting electric power to mechanical power suitable for the operation of a machine. Mechanical drive (also called working machine or simply mechanism) is a general term given to such equipment as machine tools, hoists, winches, cranes, conveyers, electrical traction outfits, pumps, and fans. Mechanical drives are powered by windmills, water wheels and turbines, steam engines, internal-combustion engines, and electric motors. The latter have gradually become the most important and extensive means of producing motive power.

I. BASIC CONCEPTS

A. Main System Components

An electric drive consists of three fundamental elements: the electric motor, which converts electrical energy to mechanical energy; the transmission, which transfers the torque, speed, and power to the working machine; and the control system, which conditions the electric power delivered to the motor in accordance with system requirements.

B. Power Conditioning

Electric drive control is involved in starting, speed regulation, braking, and reversal and also in maintaining the drive conditions required by the process being carried out or the work operation being performed. The element that performs this function is called the power conditioning unit.

This unit consists mainly of three equipment types: relays and contactors, rotating and magnetic amplifiers, and electronic or solid-state switching elements. The last-named have become the most common devices in electric power conditioning units because of their faster response, greater reliability, and smaller size and weight.

Solid-state equipment in electric drives can be divided into two main parts: the first is the power part, which includes silicon controlled rectifiers (SCRs or thyristors), triacs, gate turnoff (GTO) devices, and power transistors and diodes. The second is the logic part, which includes integrated circuits and microprocessors.

The power electronic equipment is classified into four main categories: rectifiers, inverters, choppers, and cycloconverters (frequency converters) (Table I).

C. Energy Conversion

The electric motor converts the electric energy supplied by the power conditioning unit into mechanical energy, which is transferred to the mechanism via the transmission element.

In studying the behavior of a motor selected to drive a mechanism, one of the first problems is to determine whether the speed–torque characteristic of the motor suits the requirements imposed by the speed–torque characteristic of the driven unit. The motor speed–torque characteristic is defined as the relationship between the speed ω at which it runs and the torque T it develops, $\omega = f(T)$. Electric motor speed–torque characteristics can be classified into three main categories:

1. Absolutely hard (flat) characteristic, which exhibits no change in speed with a change in torque. Synchronous motors operate with such a characteristic (Fig. 1a).
2. Hard speed characteristic, where the speed drops only slightly with an increase in torque. Induction motors over the working portion of the speed–torque characteristic (Fig. 1d) and shunt-wound direct

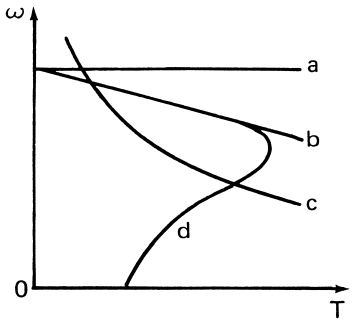


FIGURE 1 Speed–torque characteristics of several kinds of electric motor: (a) synchronous motor, (b) shunt-wound dc motor, (c) series-wound dc motor, (d) induction motor.

TABLE I Fields of Application of Power Electronic Equipment

Rectifiers	Inverters	Choppers	Frequency converters
Electrolysis	Regenerative electric drives	Field control of electric machines	Induction-motor drives
Rectifier substations of traction systems	Regenerative electric locomotive	Speed control of battery-driven vehicles	Welding equipment
Galvanizing	Direct current power transmission	Temperature control in electric furnaces	Heating converter for diesel electric locomotive
Corrosion protection	Induction-motor drives	Light dimming in theaters	
Electrophoresis	Emergency power supplies	Advertising displays	
Excitation of synchronous generators	Medium-frequency tools	Alternating current welding equipment	
Excitation of magnets	Circuit-control systems	Voltage regulation in electrostatic gas purifiers	
Electric locomotives	Electronics of vehicles, ships, aircraft	Induction-motor drives	
Battery charging	Ultracentrifuges		
Stabilized power supplies	Dielectric drying Surface heat treatment, tempering		
Welding equipment	High-frequency melting furnaces		
Direct current theater lighting	Direct current voltage transformers		
Reflectors			
Electrostatic gas purifiers			
X-ray equipment			
Electric drives			
Direct current power transmission			

current (dc) motors (Fig. 1b) operate with such a characteristic.

3. Soft speed characteristic, which exhibits considerable drop in speed with rise in torque. Series-wound dc motors operate with such a characteristic (Fig. 1c).

D. Typical Mechanisms

Driven mechanisms can be divided into three classes:

1. Speed-related machines, or mechanisms whose load is a function of speed. This dependence can be expressed by the empirical equation:

$$T_{\mathrm{L}} = T_0 + (T_{\mathrm{R}} - T_0)(\omega/\omega_{\mathrm{R}})^x, \qquad (1)$$

where T_{L} is the resisting torque developed by the load at speed ω, T_0 is the friction (resisting) torque, T_{R} is the rated (resisting) torque, ω_{R} is the rated speed, and x is an exponential coefficient. The value of x may be a whole or fractional number, but in practice the values 2, 1, 0, and -1 are used (Fig. 2).

Mechanisms having the parabolic characteristics ($x = 2$) are fans, centrifugal pumps, and ship propellers. Mechanisms exhibiting the linearly rising characteristics ($x = 1$) are viscous friction-type loads and separately excited dc generators connected to constant-resistance loads. Mechanisms in which the torque is independent of speed ($x = 0$) are cranes during hoisting, piston pumps operating against a constant-pressure head, and, with some approximation, friction torque-type machines. Finally, mechanisms having the hyperbolic characteristic ($x = -1$) are mainly of the machine-tool type, such as lathes, boring, milling, and metal-cutting machines, and ground transportation vehicles.

2. Position-related machines, or mechanisms whose load torque is a function of the position or path they are in. To this class belong, for instance, piston-type compressors whose torque depends on the piston travel and metal-cutting shears with a load torque varying according to the angle the crank pin travels. For this class, the dependence of the load torque on the path $T_{\mathrm{L}} = f(l)$ is most frequently determined by a graphical method.

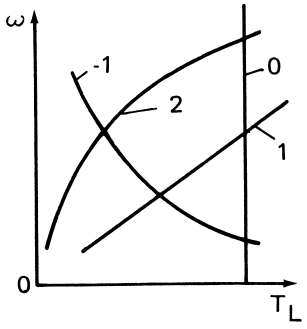


FIGURE 2 Speed–torque characteristics typical of different kinds of machines and mechanisms.

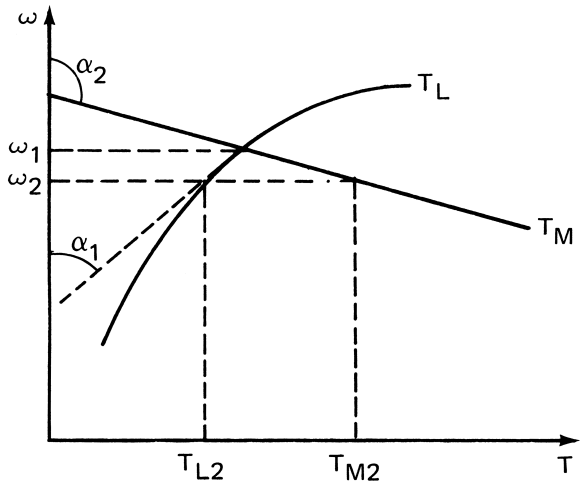


FIGURE 3 Steady–state stability conditions.

3. Random-related machines, or mechanisms whose load torque varies randomly. Examples are rock crushers, clay mills, and ball-grinding mills. In order to carry out calculations for units of this class, it is necessary to use an average value of the load torque, usually established experimentally.

A classification of typical machines and mechanisms is given in Table II.

The joint point of operation of the electric motor and of the unit it drives, when the speed has a steady value, corresponds to a condition of balance between the driving torque of the motor and the resisting torque developed by the driven unit at the given value of speed. The operating point ω_1 (Fig. 3) is in stable equilibrium when $\tan\alpha_1 > \tan\alpha_2$. This means that, when the resisting (load) torque applied to the motor shaft by the driven unit undergoes some change, the speed and torque developed by the motor will change automatically so as to restore stable operation. For instance, at ω_2 (Fig. 3), $T_{\mathrm{M2}} > T_{\mathrm{L2}}$; therefore, the speed will increase to ω_1. This automatic governing is performed by the counter electromotive force (emf) of the motor.

II. PERFORMANCE FEATURES

The torque developed by the driving motor tends to promote motion in the driven system. This motion may be uniform if the linear velocity, angular velocity, or both are constant (called steady-state conditions), or it may be nonuniform during transient conditions.

TABLE II Classification of Typical Machines and Mechanisms

Application	Special features required	Type of motor
Large ships	Large power demand	Synchronous or slip ring
Small ships	Medium power demand	Slip ring or dc
Centrifugal pumps	Normally started with closed valve; about 40–50% rated torque required for starting	Standard squirrel cage
Reciprocating pumps	Normally started with closed valve; about 100–200% rated torque required for starting	Slip ring
Rotary compressore	Unloaded starting	Squirrel cage or slip ring
Reciprocating compressors	Starting with partial load	Slip ring
Centrifugal fans	Started with closed dampers or valves	Squirrel cage or slip ring
Blowers	Starting with valves closed	Squirrel cage or slip ring
Stirrer for liquids	Light starting duty	Squirrel cage or slip ring
Cranes and hoists	Heavy starting torque, frequent starts and stops	High-torque squirrel cage of slip ring
Lifts	Smooth uniform acceleration, quiet running, frequent starts and stops	High-torque squirrel cage or slip ring
Line shafts, load conveyors, pulp grinders	Started on load; normal starting torque of 200% rated torque required	High-torque squirrel cage or slip ring
Flour and rubber mills, circular saws, chippers, beaters, crushers without flywheels, etc.	High peak loads; a maximum torque of 200–250% rated torque required; frequent starts, stops, and reversals	High-torque squirrel cage or slip ring
Crushers, punch presses, power hammers, etc., all with fly-wheels	High peak loads; heavy starting torque	High-slip, high-torque squirrel cage or slip ring
Domestic appliances: mixers, sewing machines, vacuum cleaners	Simple speed control	Universal
Domestic refrigerator compressor, window air conditioners	Good starting torque	Single-phase capacitor with direct on-line start

A. Steady-State Characteristics

In the steady-state condition, only the static resisting torque (load torque) is to be counter balanced by the torque developed by the driving motor. Load torques can be of two types: active and passive. Active torques are due to either gravitational forces, such as developed in hoists, elevators, and cars operating on gradients, or deformation in elastic bodies, such as developed during compression of springs. These torques can act in the direction of motion (loaded hoist down) or against the motion (loaded hoist up). Passive torques are those due to friction, shear (as with lathes), or deformation in inelastic bodies (fans and pumps). These are always opposed to the motion.

In view of the fact that both passive and active torques can be present in a drive system, the motor may operate in different regimes—not only as a motor but, under specific conditions, also as a generator and as a brake. In addition, the motor may be required to run in both directions. Therefore, in sketching the speed–torque characteristic of either the load or the motor, it may be necessary to use all four quadrants of the speed–torque plane for plotting (Fig. 4).

In the first quadrant the torque acts in the forward direction of motion, and therefore, the mechanical power $P_{\text{mech}} = T\omega$ is positive. This means that the electrical machine operates as a motor delivering mechanical power to a mechanism that represents a passive load. The same situation prevails in the third quadrant, with the motor running in the opposite or reverse direction of motion. In the second and fourth quadrants, instead, the torque acts against the direction of motion. The electric machine operates in the braking mode, absorbing mechanical

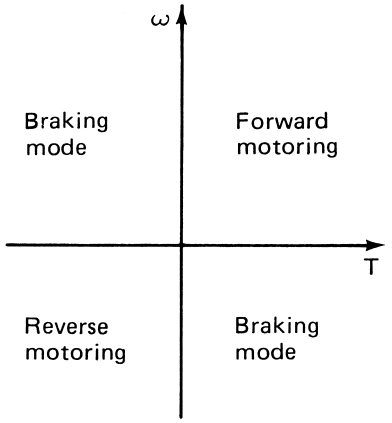


FIGURE 4 Four-quadrant operation of a drive.

power and converting it to electricity. The braking is dynamic when the electrical energy is dissipated into heat either in the resistance of the winding or in an external rheostat. It is regenerative when the electrical energy is fed back into the utility network. There also exists a third braking mode called plugging, in which power derived from both the electrical and the mechanical sources is dissipated in the resistances. Mechanical energy for this case may come from a reduction in the speed of the machine, that is, from the kinetic energy of rotation; or it may come (at fixed speed) from an active load, such as a descending elevator connected through a cable to the shaft of the motor.

B. Electromechanical Transients

Transients in electric drives result from short-time mechanical, electrical, and thermal events that come into play simultaneously and are mutually dependent on one another. Due to the relatively short duration of the electromechanical transients, the thermal change usually has no significant effect on the other two processes. For this reason, the thermal effects during transient conditions are considered to be negligible.

A transient condition occurs when a transition is made from one steady state of operation to another in which the speed, torque, and current undergo certain changes. The transient may arise as a result of a load surge, control command, supply voltage change, and so on. The nature of the transient condition must always be taken into account, because it produces a specific effect on the drive performance.

The transient processes are related to the dynamic behavior of both the driving motor and the driven machine. The following three subsections discuss the fundamentals of electric drive dynamics.

C. Equations of Motion

In translational motion, the driving force F is counterbalanced by a resisting force F_r set up by the driven machine and by an inertia force Ma arising from the change in speed, or

$$F - F_r = Ma = M\,dv/dt, \tag{2}$$

where the mass M is expressed in kg. the velocity v in m/sec, the acceleration a in m/sec^2, and force F in newtons (N).

In rotational motion, the driving torque T_M (usually developed by the electric motor) is counterbalanced by a resisting torque T_L (usually developed by the load and referred to the motor shaft) and by an inertia or dynamic torque $J\,d\omega/dt$,

$$T_M - T_L = J\,d\omega/dt = (WR^2/94)\,dn/dt, \tag{3}$$

where the inertia $J = WR^2/g$ is expressed in kg m^2. WR^2 is a technical term used in industry instead of J. It is sometimes called flywheel torque or moment. Here, W is the weight in N. R is the radius of gyration in m, $g = 9.81$ the acceleration of gravity in m/sec^2, ω is the angular velocity in rad/sec, n is the speed in revolutions per minute (rpm), and T is the torque in Nm.

The signs to be associated with T_M and T_L in Eq. (3) depend on the regime of operation (Section II.A) of the driving motor and the nature of the load torque.

D. Prime Movers–Load and Torque–Inertia Relationships

A motor generally drives an industrial machine through some transmission system whose individual parts operate at different speeds. For the purpose of analysis it is expedient to refer all the torques and masses of the individual parts to some convenient element, for instance, the motor shaft.

The load torque can be referred from one shaft to another on the basis of the power balance in the system.

$$T'_L\omega_M\eta_t = T_L\omega_L,$$

so that T'_L, the load torque T_L referred to the motor shaft, is

$$T'_L = T_L/k\eta_t, \tag{4}$$

where $k = \omega_M/\omega_L$ is the speed (transmission) ratio and η_t is the transmission efficiency.

The inertia can be dealt with on the basis that the total amount of kinetic energy stored in the moving parts remains unchanged and when referred to the motor shaft is

$$J'\frac{\omega_M^2}{2} = J_M\frac{\omega_M^2}{2} + \sum_i^m J_i\frac{\omega_i^2}{2}. \tag{5}$$

In words, the equivalent total kinetic energy stored in the motor is the sum of all the kinetic energies in the m parts moving at different speeds ω_i and that of the motor itself (Fig. 5).

It follows that the equivalent polar inertia referred to the motor shaft is

$$J' = J_M + \sum_i^m J_i\frac{1}{k_i^2}, \tag{6}$$

where $k_i = \omega_M/\omega_i$ and the subscript i represents the individual part that rotates at a speed of ω_i.

A translational motion can be referred to a rotating shaft on the basis of the power balance in the system

$$T'_L\omega_M\eta_t = F_L v_L, \tag{7}$$

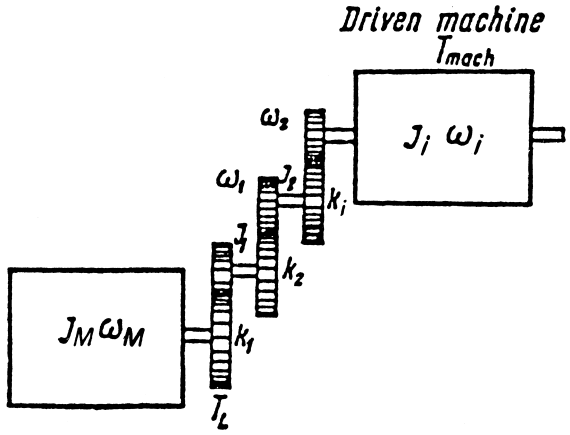


FIGURE 5 Gearing diagram showing transmission links provided between a motor and the machine it drives.

where F_L is the resisting force developed by a given machine against the translational motion at linear speed v_L (Fig. 6). The load torque T_L referred to the motor shaft is then

$$T'_L = F_L v_L / \omega_M n_t. \tag{8}$$

A translating mass is referred to a rotating shaft on the basis of equality of kinetic energy,

$$M v_L^2/2 = J' \omega_M^2/2,$$

and J, the equivalent polar inertia on the motor shaft rotating at a speed of ω_M, is

$$J' = M(v_L/\omega_M)^2. \tag{9}$$

When the machine has both rotating and translating parts, the aggregate inertia referred to the motor shaft can be expressed by the sum of Eqs. (6) and (9).

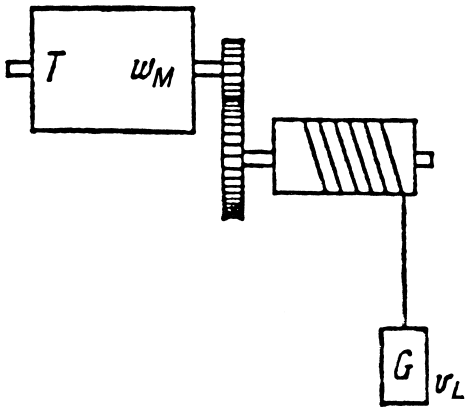


FIGURE 6 Gearing diagram of a hoist drive.

E. Acceleration and Deceleration Times

The duration of the transient conditions in a drive, that is, the duration of starting, backing, and transition from one speed to another, affects the performance of the driven machine. The duration of the transient is determined by integrating the equation of motion [Eq. (3)]. The time interval during which the speed varies between ω_1 and ω_2 is then

$$t_{1,2} = \int_{\omega_1}^{\omega_2} \frac{J' \, d\omega}{T_M - T'_L}, \tag{10}$$

where J' and T'_L are the inertia and load torque referred to the motor shaft. In particular, this equation gives the time required for starting when $\omega_1 = 0$ and $\omega_2 = \omega_M$, where ω_M is the steady-state speed of the motor.

To perform the integration in Eq. (10) it is necessary to know the motor and the load torques, T_M and T'_L, as functions of speed. In complicated cases, numerical computation must be used. Two examples are discussed below.

1. Starting of a Shunt-Wound Direct Current Motor

The shunt-wound dc motor in Fig. 7 is started by closing the switch S. It is reasonable to assume that the armature inductance is negligible and that the supply voltages V, excitation flux ϕ, and load torque T'_L are constant. Let R denote the sum of R_a, the resistance of the armature, and R_{rh}, that of the series-connected starting rheostat.

According to Eq. (10) of "Electric Machines, Rotating, Construction and Theory," the motor torque is proportional to the armature current I, or

$$T_M = k\Phi I = cI. \tag{11}$$

The Kirchhoff voltage equation is

$$V = E + IR. \tag{12}$$

According to Eq. (11) of "Electric Machines, Rotating, Construction and Theory," the counter emf E is proportional to the speed ω, or

$$E = c\omega. \tag{13}$$

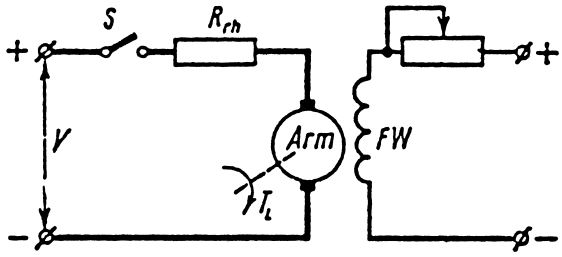


FIGURE 7 Basic starting circuit of a shunt-wound dc motor.

Equations (10)–(13) can be solved for the speed $\omega = f_1(t)$, the armature current $I = f_2(t)$, and the starting time t_{st}.

In general, if the initial speed is denoted by ω_1 and the steady-state speed by ω_L, one obtains

$$\omega = \omega_L + (\omega_1 - \omega_L)e^{-t/\tau_{mec}} \tag{14}$$

and in particular for starting, when $\omega_1 = 0$, the speed becomes

$$\omega = \omega_L(1 - e^{-t/\tau_{mec}}), \tag{15}$$

where $\tau_{mec} = JR/c^2 = J\omega_0/T_{st}$ is the mechanical time constant of the drive, with ω_0 as the no-load speed of the motor (see Section II.F of "Electric Machines, Rotating, Construction and Theory") and T_{st} as the starting torque (short-circuit torque).

In practice, it is adequate to assume that steady state is reached after three time constants, or

$$t_{st} \cong 3\tau_{mec}. \tag{16}$$

The armature current I is

$$I = \left(\frac{V}{R} - \frac{T_L}{c}\right)e^{-t/\tau_{mec}} + \frac{T_L}{c} \tag{17}$$

and in steady state is $I = T_L/c$.

2. Starting of an Alternating Current Motor

The starting process of an ac motor-fan drive is illustrated in Fig. 8. In this case Eq. (10) can be solved only by means of numerical computations.

The reciprocal of the accelerating torque $1/(T_M - T_L)$ is plotted as a function of speed in Fig. 8b. The shaded area is the value of the integral in question [Eq. (10)] and, when multiplied by the total inertia J, gives the starting time t_{st}, At exactly ω_L, the quantity $T_M - T_L$ equals zero (Fig. 8a); therefore, its reciprocal value goes to infinity. This singularity is removed by considering the accelerating time about $0.98\omega_L$. The horizontal axis ω of the curve $1/(T_M - T_L)$ can be divided into small increments in order to construct the complete speed vs time curve (Fig. 8c).

F. Electric Braking

Using the electric motor as a means of applying a braking torque to the drive has several advantages over mechanical braking: (1) less maintenance is required: (2) no special means are needed for cooling; (3) the operation is smooth and chattering is avoided; (4) the control is easier; and (5) the efficiency is higher. The main limitation is that, normally, electric braking cannot provide a holding torque.

As stated in Section II.A there are three types of electric braking: (1) dynamic braking, when the motor operates as a generator but the electrical energy is dissipated as heat in the winding or in an external rheostat; (2) regenerative braking, when the motor operates as a generator and returns electrical energy back to the utility network; and (3) plugging or countercurrent braking, when the motor is operated as a generator connected in series with the utility source and the combined electrical energy is wasted in the

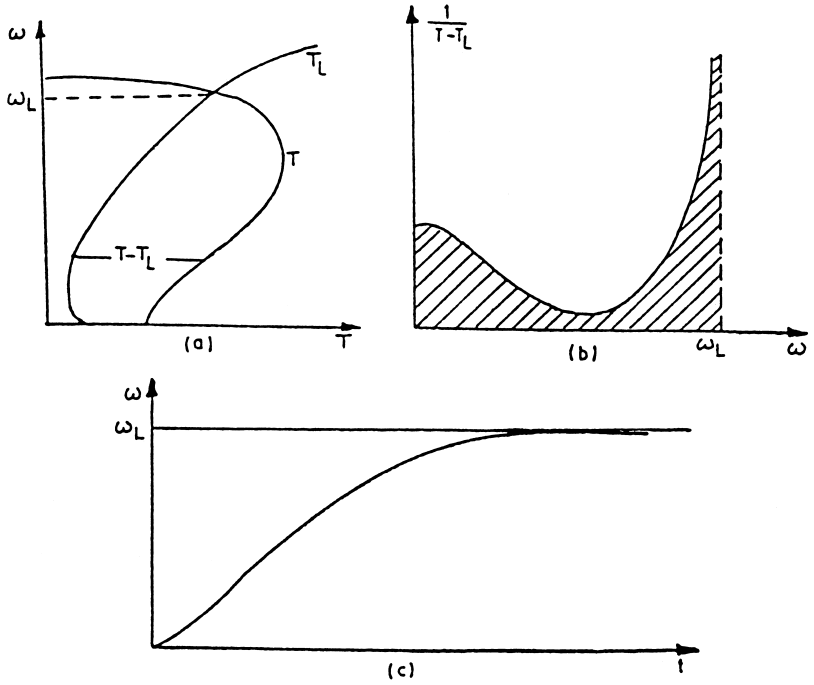


FIGURE 8 (a) Speed–torque characteristics of ac motor and load; (b) area under the curve $1/(T_M - T_L)$ vs ω; (c) speed–time curve.

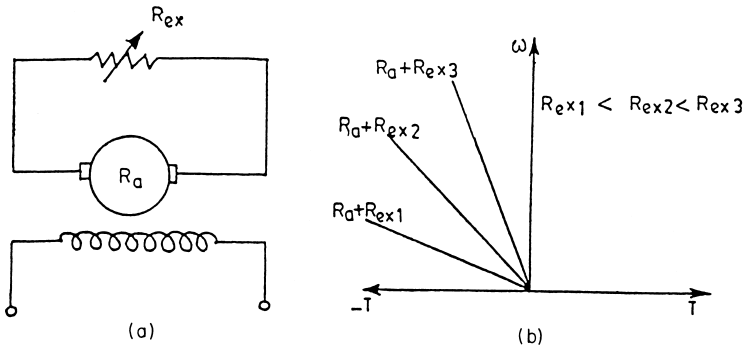


FIGURE 9 Dynamic braking: (a) basic circuit, (b) speed–torque characteristics.

motor winding or in an external rheostat. Two examples are discussed below.

1. Braking of a Shunt-Wound Direct Current Motor

The excitation flux in a shunt-wound dc motor and in a separately excited motor can be considered constant. This means that the motor torque T_M depends only on the armature current I, $T_M = cI$ [Eq. (11)].

Dynamic braking is performed by switching the armature connection from the source to an external resistor, while the field winding remains connected to the source (Fig. 9a). Since the source voltage is now $V = 0$, the polarity of the motor's counter emf will reverse the direction of the armature current I.

$$I = (V - E)/R = -E/R, \tag{18}$$

where R is the combined resistance of the armature circuit. The torque–speed characteristic is then

$$\omega = -TR/c^2 \tag{19}$$

so that the dynamic braking characteristics are straight lines starting at the origin (Fig. 9b).

This mode of operation is useful when a complete stop within a short, but controlled braking time is required, as in the case of an electric vehicle. The energy can be dumped into a heater or a rechargeable battery.

Regenerative braking, in which the energy is fed back to the supply, occurs in those cases when the motor is made to run faster than its ideal no-load speed. The counter emf E then exceeds the impressed voltage V, causing the input current to reverse direction (Fig. 10a). According to Eq. (18) with $E > V$, the current I becomes negative and therefore the torque $T_M = cI$ also changes its sign. The speed–torque characteristic.

$$\omega = \frac{V}{c} - T\frac{R}{c^2} \tag{20}$$

then becomes an extention into the second and fourth quadrants of the curves representing motor operation (Fig. 10b).

This braking mode is realizable with conveyers and hoisting mechanisms when the load is being lowered or in cases of speed control when the motor no-load speed $\omega_0 = V/c$ is reduced below the shaft speed ω. This is the most efficient mode of electrical braking since part of the electrical energy converted from mechanical form is returned to the supply.

Countercurrent braking is accomplished in two ways: (1) by a short-time reversal of the connection of the motor armature to the supply so as to bring about a rapid stopping called plugging (curve A in Fig. 11) and (2) by inserting a relatively large resistor in the armature circuit when lowering the load of a hoist mechanism while the motor is still connected so as to raise the load (curve B in Fig. 11). In this case the voltage V reverses its polarity, but the added series resistance limits the current. The speed–torque characteristic is given by Eq. (20) with negative V and is an extention (with a relatively large R) into the second and fourth quadrants of the characteristic representing motor operation.

This is the most inefficient technique of electric braking since, in addition to dissipating the electrical energy converted from mechanical from in the resistances of the circuit, the electrical energy drawn from the supply is also wasted.

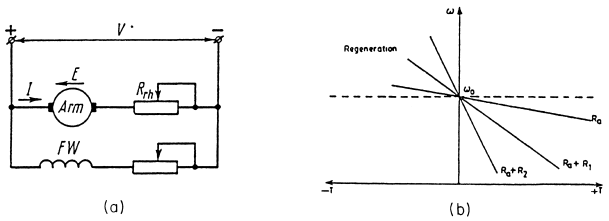


FIGURE 10 Regenerative braking: (a) basic circuit, (b) speed–torque characteristics.

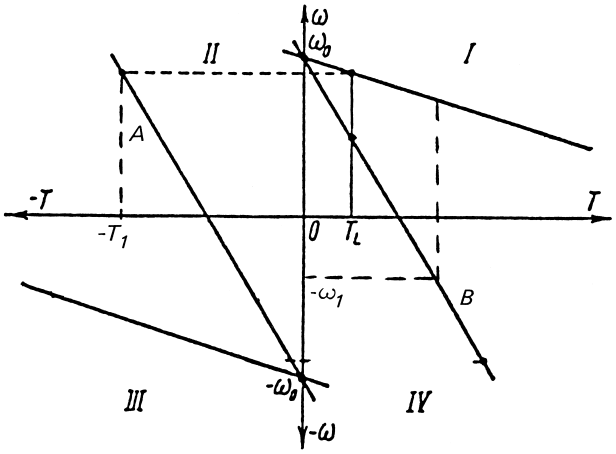


FIGURE 11 Speed–torque characteristics of countercurrent braking of a shunt-wound dc motor.

2. Braking of Induction Motors

A slip-ring induction motor has a greater variety of braking mode controls than a squirrel-cage motor. This is due to the additional flexibility in controlling the resistance of the rotor circuit.

Dynamic braking is usually obtained by switching over the stator winding to a dc voltage source. In order to limit the current magnitude and to control the braking rate, one can either vary the dc voltage or, if the rotor has slip rings, connect it to external resistors (Fig. 12).

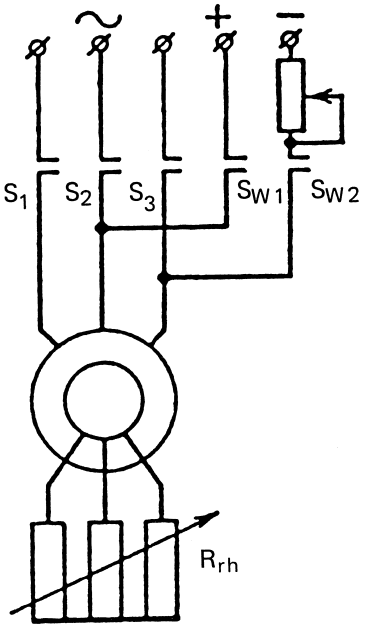


FIGURE 12 Dynamic braking of an induction motor with slip rings, where S_1, S_2, and S_3 are the main switches and SW_1 and SW_2 are auxiliary switches connected to the dc source.

The dc current in the stator generates a stationary magnetic flux in the motor air gap. Rotation of the rotor in this field induces ac currents in its windings. The frequency of these currents decreases with rotor speed so as to set up a magnetic field that is stationary with respect to the stator. The interaction of the two fields creates a braking torque whose magnitude depends on the dc field generated by the stator, the rotor resistance, and the speed of the rotor. The braking torque T as a function of the slip S given by Eq. (81) of "Electric Machines, Rotating, Construction and Theory", with $T_{\max}$ proportional to the square of the dc current, and $S = \omega/\omega_0$ (where ω is the rotor shaft speed and ω_0 is the synchronous speed of the motor). This braking mode is applied mainly to hoisting and crane machines.

Regenerative braking can take place only when the rotor rotates in the same direction as that of the stator magnetic field but faster than its synchronous speed ω_0. Several torque–speed characteristics are plotted in Fig. 13, where the different resistance values refer to a slip-ring rotor. The curve $R_r = 0$ is the natural motor characteristic, which is the only one realizable with a squirrel-cage motor.

Any increase in the rotor speed ω above ω_0, $\omega > \omega_0$, due to some external torque, will force the motor to operate in a generating mode. When $\omega > \omega_0$, the slip S becomes negative. Therefore, rewriting Eq. (78) of "Electric Machines, Rotating, Construction and Theory,"

$$\bar{I}'_2 = \frac{\bar{V}_1 R'_2 S}{(R'_2)^2 + (\omega_1 L_{sc} S)^2} - j\frac{\bar{V}_1 \omega_1 L_{sc} S^2}{(R'_2)^2 + (\omega_1 L_{sc} S)^2}, \quad (21)$$

the active component of the current $\bar{I}'_2$ reverses direction while the reactive one stays in the same direction. This means that the motor operates as a generator connected in parallel with the utility network and returns electrical energy to it while drawing reactive power for its excitation.

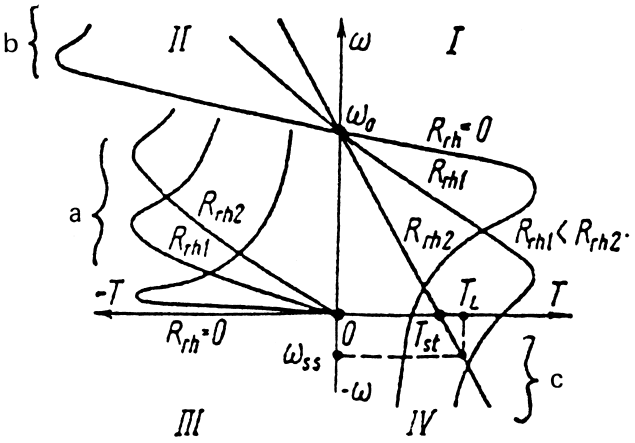


FIGURE 13 Speed–torque characteristics of an induction motor for different conditions of operation: (a) dynamic braking, (b) regenerative braking, (c) countercurrent braking.

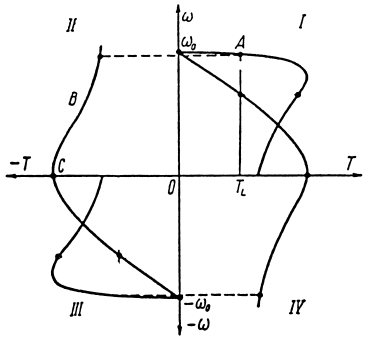


FIGURE 14 Speed–torque characteristics of an induction motor for the case when two of the supply leads are interchanged.

This braking mode is used in drives with pole-changing motors, as well as in hoisting mechanisms, and so on.

Countercurrent braking is widely used in drive practice, mainly when the motor is of the wound-rotor type. The reason is that the presence of slip rings permits the introduction of resistances in the rotor circuit to limit the magnitude of the current and to control the braking rate (Fig. 14).

This plugging condition can be set up by interchanging any two phases of the stator winding so as to reverse the direction of the rotating magnetic field in the motor air gap while the rotor still rotates in the initial direction. Since the speed has a negative sign, the slip S is larger than 1, $S > 1$. The torque, which is given by Eq. (77) of "Electric Machines, Rotating, Construction and Theory," is relatively small. Since it acts in the opposite direction to the rotor motion, the latter begins to slow down (curve B in Fig. 14). When the speed drops to zero (point C on curve B), the motor should be deenergized; otherwise, it will begin to rotate in the reverse direction.

III. SPEED CONTROL

Modern industry employs many machines that require different and adjustable speeds, such as metal-cutting machine tools, electric cranes, electric traction units, and rolling mills.

A. Main Features

The fundamental features by which various methods of electric speed control can be characterized are as follows:

1. Range of speed control is defined as the ratio of maximum required speed $\omega_{\max}$ to minimum required speed $\omega_{\min}$, $K_r = \omega_{\max}/\omega_{\min}$. In general, K_r is expressed as a numerical proportion such as 2:1 or 100:1.
2. Continuity (smoothness) of speed control depends on the total number of steady operating speeds required by the drive within a given range. The continuity factor is expressed as a ratio of any adjacent steps of speed control, $K_{cf} = \omega_n/\omega_{n-1}$, where ω_n is the speed of the nth and ω_{n-1} is the speed of the $(n-1)$th step.
3. Stability of operation at a given speed is defined as the change in speed caused by a change in the load torque. For a given mechanism, the harder the motor speed–torque characteristic is, the greater the stability.
4. Economy is concerned with capital cost, energy losses, reliability, and availability of the equipment needed to set up the system.
5. Direction of speed control is defined as the direction in which the speed of the motor is changed from a base speed and depends on the method of control. The base speed is a designated speed that the motor develops at rated voltage for a predetermined value of load torque. For instance, ω_b in Fig. 15 is the base speed that separates two zones: constant-torque and constant-power control zones.
6. The permissible load of any adjustable-speed characteristic is determined by the rated current of the motor. Each method of speed control corresponds to specific characteristics of the load. For instance, hoist equipment requires speed control at constant torque. Since $P = T\omega$, this means that the output power varies in proportion to the speed (zone I in Fig. 15). This kind of speed control is usually accomplished by maintaining the current at a constant level. A second example is a metal-cutting lathe in which the linear velocity v of the workpiece surface is kept constant to sustain a constant cutting speed and force f. Since the output power is constant, $P = fv$, the torque will be inversely proportional to the speed. $T = P/\omega$ (zone II in Fig. 15). Again, the shaft speed is usually varied while the current is kept constant.

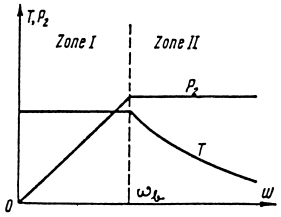


FIGURE 15 Power–speed and torque–speed characteristics of a shunt-wound dc motor for two zones of speed control.

B. Direct Current

The speed of a dc motor is, according to Eq. (21) of "Electric Machines, Rotating, Construction and Theory",

$$\omega = (V - IR)/k\Phi. \tag{22}$$

It follows that there are three essentially different ways of controlling the speed of the motor: (1) by varying the resistance R in the armature circuit, (2) by varying the

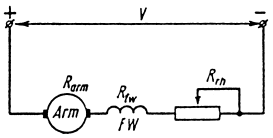


FIGURE 16 Basic connections of a series-wound dc motor.

field excitation current and, therefore, the flux Φ, and (3) by varying the voltage V applied to the armature.

1. Series-Wound Direct Current Motors

Speed control by varying the resistance in series with the armature is used to vary the angular velocity down from the base speed (Fig. 16). A narrow speed control range of about 3:1 is feasible, depending on the load. This method of speed control is widely used in traction drives because of its simplicity.

Speed control by varying the field excitation is used to increase the speed above the base speed (Fig. 17). For a given armature current I, any change made in the shunting resistance R_{sh} will result in a corresponding change in the exciting current $I_{ex} = I - I_{sh}$. Since the losses in the shunt resistor are of the same order as those in the excitation coil and both have relatively low resistance, the efficiency is quite high.

This method of speed control is widely applied to universal motors (series-wound motors that operate in ac systems), such as those used in electric hand drills, kitchen mixers, and other equipment. In practice, a static solid-state switch is connected across the shunt resistor, and its on–off operation controls the excitation current.

Speed control by varying the applied voltage can be accomplished by feeding the motor from a variable-voltage source or using a series–parallel motor connection scheme. The latter is applied when the machine is driven by two motors, each rated for one-half the required load capacity. Figure 18a shows the sequence of connections used to control the speed of two identical series-wound dc motors, and the speed control characteristics are plotted in Fig. 18b.

The advantages of using two motors instead of one are (1) shorter starting and braking times by reducing the total moment of inertia, which, according to Eq. (3), is proportional to the square of the diameter; (2) higher reliability of operation in case of fault in one motor; (3) attractiveness

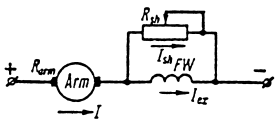


FIGURE 17 Speed control of a series-wound dc motor by shunting the field winding with a resistor.

of small motors for applications with space limitations; and (4) higher efficiency with the two steps of speed control (series and/or parallel connections) as compared with a single motor having a series-connected resistor. Drives such as this are used in electric trains, street cars, heavy-duty metal-cutting shears, and so on. Both motors in these cases have the same power ratings and a common shaft.

2. Shunt-Wound Direct Current Motors

Speed control by varying the resistance in the armature circuit varies the hardness of the speed–torque characteristics and therefore results in a different stability of operation for a given speed. This type of control is used to reduce the speed below the base speed at constant load torque. The speed range is not constant and depends on the load (Fig. 19). This control method is not economical due to the relatively high power losses in the series-connected resistor.

Speed control by varying the excitation current increases the speed above the base speed. The no-load speed is inversely proportional to the flux $\omega_0 = V/k\Phi$, and the starting torque is $T_{st} = k\Phi I_{st}$ (Fig. 20). The armature current $I_{st} = V/R_{arm}$ is the short-circuit current at start.

This is a simple and economical control method, and its speed range is within 2:1 to 5:1. The extreme speeds are determined by saturation of the magnetic circuit (lower limit) and by the mechanical strength for which the armature is designed (high limit). The main application is in motor–generator sets when there is a need for a relatively high speed.

Speed control by varying the supply voltage provides a wide range and smooth speed control (Fig. 21). Since the speed gradient R/c^2 [Eq. (20)] is not a function of the voltage V, the speed–torque characteristics consist of a family of parallel straight lines (parallel to the natural characteristic). The speed range can be as high as 100:1, and this range can be increased further by a factor of 2:1 by varying the field excitation of the motor as well.

The classical system that utilizes the voltage control method is called the Ward–Leonard system (Fig. 22). In this system, the generator G serves as a dc supply to the motor M to drive a production machine PM. The armatures of the two machines, G and M, are directly connected, and both are separately excitated by the exciter E. The generator G and the exciter E are driven at a fixed speed by the induction motor IM.

The speed–torque equation is [see also Eq. (20)]

$$\omega = \frac{E_g}{k\Phi_m} - T\frac{R_m + R_g}{k^2\Phi_m^2}, \tag{23}$$

where E_g is the generator emf, R_m and R_g are the armature resistances, and Φ_m is the motor flux.

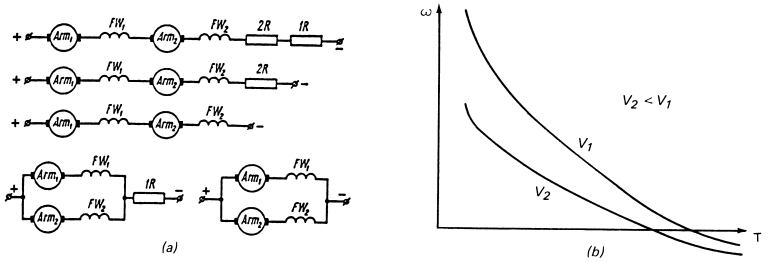


FIGURE 18 Voltage speed control of a series-connected dc motor: (a) sequence of connections, (b) speed–torque characteristics.

All speed control operations, including starting, braking, and reverse direction, are governed by the excitation circuits (e.g., starting is performed by increasing the excitation current of the generator G from zero level). The low power level required for speed control of the Ward–Leonard system is its main advantage. This makes the system simple and reliable. The drawbacks are relatively low efficiency (∼70%) and high investment and maintenance costs.

3. Static Converters

Static (solid-state) converters for the dc motor speed control can be divided into two categories: dc–dc (or dc chopper) and ac–dc (or controlled rectifier) converters. The power solid-state devices are simple diodes: silicon controlled rectifiers (SCRs), also called thyristors, and triacs, which control the start of the conduction period, and gate turnoff (GTO) thyristors; metal–oxide–silicon field-effect transistors (MOSFETs); and power transistors that can control both the beginning and the end of the conduction period.

Chopper circuits are used to control the speed of dc motors, which are fed from fixed-voltage sources, such as batteries or uncontrolled rectifiers.

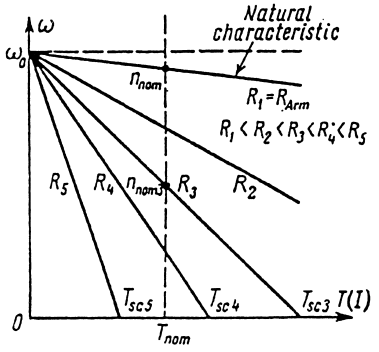


FIGURE 19 Natural speed–torque characteristics and a family of rheostatic speed–torque of characteristics of a shunt-wound dc motor.

In order to supply a load that requires a variable voltage from a dc source, it is necessary to insert, between the source and the load, a resistance that absorbs the difference between the source and load voltages. The voltage across the resistor varies as the load current does. Hence, the only way of supplying a constant voltage to a variable load is to change the series resistor. There are two extreme cases in which dissipation in the series resistor is zero, when its value is either zero or infinite. In the first case the load receives the full supply voltage. In the second case no current flows and the voltage across the load is zero. A voltage control that is loss free, in principle, can be achieved by varying the series resistor in abrupt steps between the values zero and infinity, thus reducing the average value of the voltage across the load below the supply voltage. An abrupt change in the voltage can be realized, in principle, by an ideal switch, which lets no current pass when open and has no voltage drop when closed (Fig. 23).

If the switch is closed for an interval of time t_{on} during a period T, then the average value V_l, of the voltage across the load is less than the supply voltage and is given by

$$V_L = V_s t_{on}/T = V_s a, \tag{24}$$

where $a = t_{on}/T$.

The voltage across the load includes both dc and ac components. The latter is undesirable and can be reduced by means of either a series connected inductance or a parallel-connected capacitor. In power applications of dc chopper circuits, this reduction is implemented almost exclusively

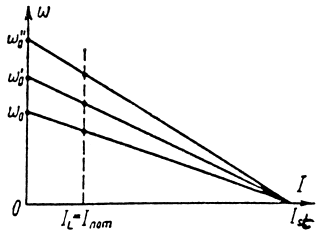


FIGURE 20 Speed characteristics of a shunt-wound dc motor whose speed is controlled by varying the excitation current.

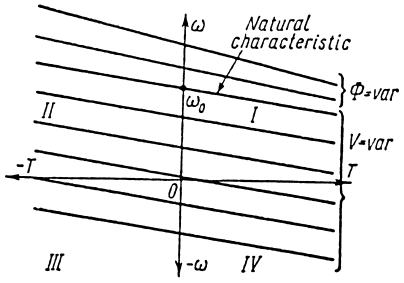


FIGURE 21 Speed–torque characteristics of a shunt-wound dc motor for adjustable-voltage speed control.

by means of inductances. In these cases, a freewheeling diode should be connected across the load (Fig. 24) to allow flow of the current driven by the inductance when the switch S is open.

The full supply voltage V_s appears across the series-connected resistance R and inductance L of the armature and across the emf E at the instant when the switch S is closed. Assuming ideal components, the current i_1 is

$$i_1(t) = \frac{V_s - E}{R}(1 - e^{-t/T_e}) + I_{min}e^{-t/T_e}. \quad (25)$$

When the switch S is open, the load current flows through the diode D and the current i_2 is

$$i_2(t - aT) = I_{max}e^{-(t-aT)/T_e}, \quad (26)$$

where $T_e = L/R$ and

$$I_{max} = \frac{V_s - E}{R}\frac{1 - e^{-aT/T_e}}{1 - e^{-T/T_e}} \quad (27)$$

$$I_{max} = \frac{V_s - E}{R}\frac{(1 - e^{-aT/T_e})e^{-T(1-a)T_e}}{1 - e^{-T/T_e}}. \quad (28)$$

The average current through the load is

$$I_{av} = \frac{1}{T}\left(\int_0^{aT} i_1\,dt + \int_{aT}^{T} i_2\,dt\right) = \frac{aV_s - E}{R}. \quad (29)$$

This result could be expected because the average voltage absorbed by the inductance L over a period must be

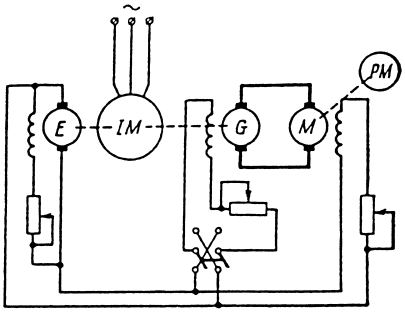


FIGURE 22 Schematic diagram of an adjustable-voltage system of speed control for a shunt-wound dc motor.

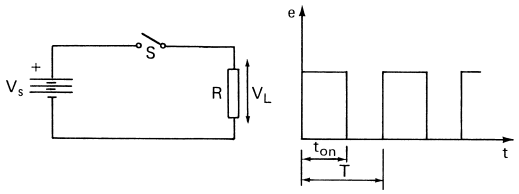


FIGURE 23 Direct current voltage adjustment by means of a switch.

zero. The average voltage across the load consisting of the resistor and the battery is given by Eq. (24). In practice, the period T is chosen to be much shorter than the mechanical time constant τ_{mec} of the drive, so that the speed remains practically constant at the value determined by the average voltage and load torque. The torque–speed characteristic is given by Eq. (20) with $V = aV_s$. It follows that the only control parameter is the normalized pulse width a. The speed–torque characteristics are similar to those of Fig. 21 with variable voltage. This can be varied in three ways:

1. Pulse-width control. In this mode, the switching time T in Fig. 23 is kept constant, and the on time t_{on} is varied. Since the ratio T/τ_{mec} (period/mechanical time constant) is constant, the ripple of the load current can be kept within predetermined limits.
2. Pulse-frequency control. In this mode, the on time t_{on} is kept constant, but the off time t_{off} and with it the period T are varied. Since a minimum on time is kept constant at all times, the switching no-load (low-frequency) losses are comparatively small.
3. Two-position control. In this mode, neither the switching frequency nor the on time is kept constant. This is a closed-loop control in which a sensor monitors the load current and adjusts the instants of time t_{on} and t_{off} so that the instantaneous load current i will be kept within a preset range, $I_{max} - I_{min} = \text{const}$ (see Fig. 24). The advantage is that the size of the smoothing reactor (to reduce the ac components) can be optimized.

The various chopper configurations and their operation in different quadrants are shown in Table III. In the first quadrant, when the chopper CH1 is on, the load voltage

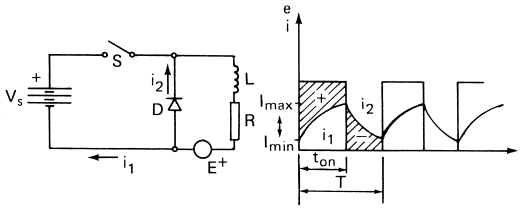
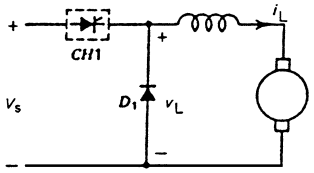
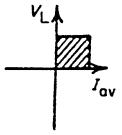
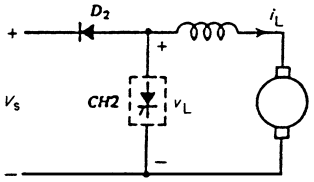
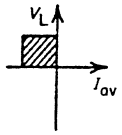
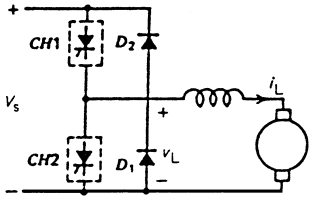
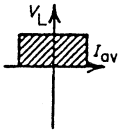
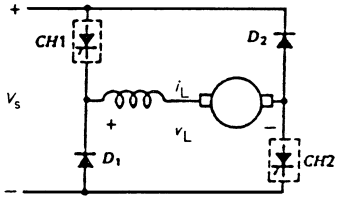
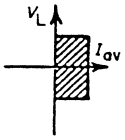
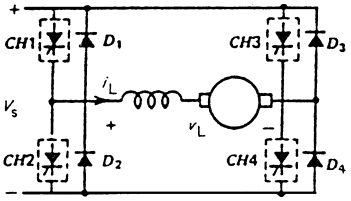
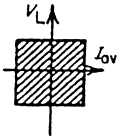


FIGURE 24 Operation of a dc chopper.

TABLE III Chopper Configurations

Type	Chopper configuration	$V_L - I_{av}$ characteristics	Function
First-quadrant chopper			$v_L = V_s$ CH1 on $= 0$, CH1 off D_1 on
Second quadrant or regenerative chopper			$v_L = 0$, CH2 on $= V_s$, CH2 off D_2 on
Two-quadrant type A chopper			$v_L = V_s$ CH1 or D_2 on $= 0$, CH2 or D_1 on i_L = positive CH1 or D_1 on = negative CH2 or D_2 on
Two-quadrant type B chopper			$v_L = -V_s$ CH1 and CH2 on $v_L = V_s$ CH1 and CH2 off D_1 and D_2 on
Four-quadrant chopper			CH4 on and CH3 off CH1 and CH2 operated v_L = positive I_{av} = reversible CH2 on and CH1 off CH3 and CH4 operated v_L = negative I_{av} = reversible

$v_L = V_s$: when the chopper is off, $v_L = 0$. Both average load voltage V_L and current I_{av} are positive, and power flows from the source to the load. This configuration is used for motoring operation.

In the second quadrant, when the chopper CH2 is on, the load voltage $v_L = 0$; when the chopper is off, $v_L = V_s$ (when D_2 conducts current). The current i_L flows from the load either through CH2 charging the reactor L or through D_2 back to the source V_s. The average voltage V_L across the load is positive, but the average current I_{av} is negative: therefore, power flows from the load to the source. This configuration is used for the regenerative braking mode. Braking current and torque can be varied by controlling the pulse width a.

Two-quadrant (I and II) operation can be achieved by combining the above two circuits. Chopper CH1 controls motoring operation, whereas D_1 serves as the freewheeling diode. Chopper CH2 controls the regenerative mode, whereas D_2 returns the current back to the source.

Two-quadrant (I and IV) operation can also be achieved in a bridge configuration of two choppers and two diodes. The two choppers CH1 and CH2 operate simultaneously.

When they are on. the instantaneous load voltage v_L is positive. When they are off, the current flows back to the source through the diodes D_1 and D_2, and the instantaneous load voltage v_L is negative. The average voltage V_L across the load is either positive or negative depending on whether the on time is greater or smaller than the off time. Since the load current flows in the same direction, the machine operates either as a motor (positive power) or as a generator (negative power).

Four-quadrant operation can be achieved by a bridge configuration with four choppers and four diodes. With CH4 always on and CH3 always off, choppers CH1 and CH2 provide operation in quadrants I and II. With CH2 always on and CH1 always off, choppers CH4 and CH3 provide operation in quadrants III and IV.

Controlled rectifiers are used to control the speed of dc motors that are fed from ac single- or three-phase systems at constant voltage. The rectifier changes the fixed ac input voltage to a controllable dc output voltage. The most common switching devices are SCRs (thyristors), because commutation (i.e., transfer of current from one device to the next) is easily achieved by natural (line) commutation. When an incoming thyristor is turned on, it immediately biases the outgoing thyristor in the reverse direction and turns if off.

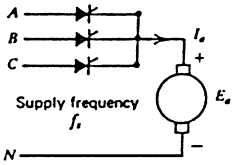
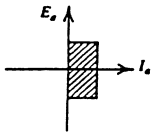
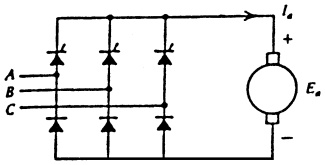
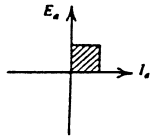
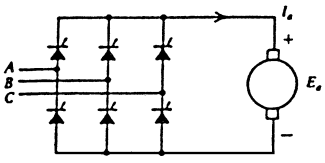
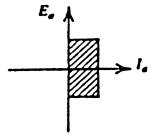
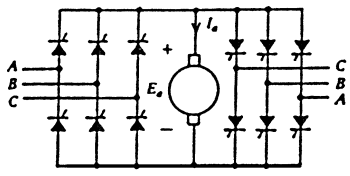
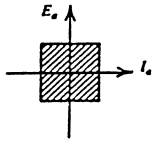
Phase-controlled rectifiers are broadly classified as single-phase and three-phase units. When a three-phase unit is compared with a single-phase unit, the ripple frequency of the voltage at the motor terminals is found to be higher and the amplitude is smaller. Consequently, the filtering requirements for smoothing out the motor current are less stringent. With three-phase units the motor current is continuous (it does not drop to zero), and therefore, it can be compared with that of a Ward–Leonard system. In a single-phase unit the current may drop to zero, and this causes the speed–torque characteristic to become softer. Some of these three-phase rectifier circuits are shown in Table IV.

The most commonly used unit is the three-phase, six-pulse, fully controlled rectifier (Fig. 25). The thyristors are fired in an interval of 60°, and the motor terminal voltage has a six-pulse-per-cycle ripple. The thyristors are numbered in sequence in the order in which they are fired.

The ideal output voltage across the load is

$$V_{d\alpha} = V_{d0} \cos \alpha, \tag{30}$$

TABLE IV Three-Phase Thyristor Phase-Controlled Converter Circuits

Circuit	Type	Typical hp	Ripple frequency	Quadrant operation
	Half-wave	10–50	$3f_s$	
	Semi-converter	15–150	$6f_s$	
	Full converter	100–150	$6f_s$	
	Dual converter	200–2000	$6f_s$	

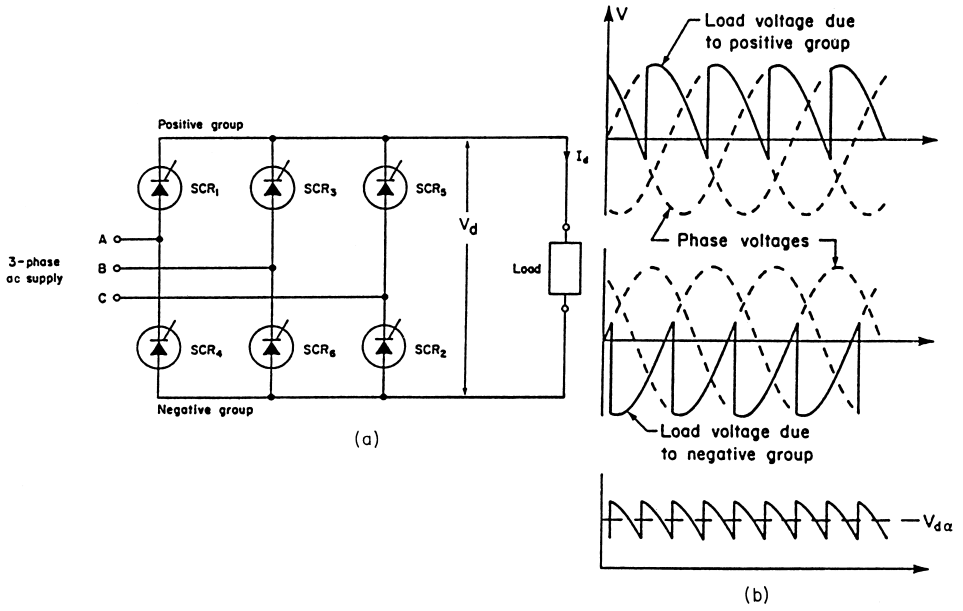


FIGURE 25 Three-phase, six-pulse, fully controlled rectifier: (a) power circuit; (b) phase and load voltages.

where

$$V_{d0} = E_{max}\frac{q}{\pi}\sin\frac{\pi}{q} \tag{31}$$

and α is the delayed firing angle, q is the number of dc pulses (in this case $q = 6$), and E_{max} is the amplitude of the ac voltage. In practice, the reactive voltage drop across the transmission lines and the leakage reactance of the transformers reduce the output to a dc level V_d,

$$V_d = V_{d0}\cos\alpha - \frac{q}{2\pi}X_c I_d, \tag{32}$$

where the commutating reactance is $X_c = \omega L_c$, ω is the input frequency, L_c is the commutating inductance (combining the leakage inductance of the transformer with that of the transmission line), and I_d is the dc current in the load.

Applying Eq. (32) to the dc motor speed–torque equation [Eq. (20)] will produce the desired relation between the thyristor firing angle α and the motor speed ω:

$$\omega = \frac{V_{d0}\cos\alpha}{k\Phi} - T\frac{R_a + qX_c/2\pi}{(k\Phi)^2}. \tag{33}$$

The slope of the speed–torque characteristic of the motor driven by the controlled rectifier is higher than the natural characteristic of the motor due to the voltage drop in the commutating reactance.

C. Induction Motors

The torque–speed relationship of an induction motor is expressed in terms of the slip S by Eq. (81) of "Electric Machines, Rotating. Construction and Theory." The values for S, T_{max}, and S_{max} are given by Eqs. (61), (80), and (79) of the same article. It follows that there are four essential ways of controlling the speed (slip S): (1) increasing rotor resistance R_2; (2) changing the number of pole pairs, p; (3) varying the supply voltage V_{ph}; and (4) varying the supply frequency f.

1. Squirrel-Cage Motor

An increase in rotor resistance cannot be achieved with squirrel-cage motors because the rotor conductors are shorted internally. A change in the number of pole pairs, p, Affects n_0, the synchronous speed of the motor in rpm, because

$$n_0 = 60f_1/p, \tag{34}$$

where f_1 is the source frequency. For instance, a two-speed motor with $p_1 = 1$ and $p_2 = 2$ runs, in a 60-Hz system, at $n_{01} = 3600$ rpm and $n_{02} = 1800$ rpm. Such a motor is manufactured with each phase winding divided into two equal parts with terminal connections shown in Fig. 26. A speed change is accomplished by switching between those parts from a series to a parallel connection. This causes the current direction in the conductors of one part to change and the number of pole pairs to be reduced to one-half.

In practice, the switchover between winding parts of each phase is done either by a single-wye to double-wye connection method (Fig. 27a) or by a delta to double-wye connection method (Fig. 27b). In going from a single-wye to a double-wye connection, the magnitude of the

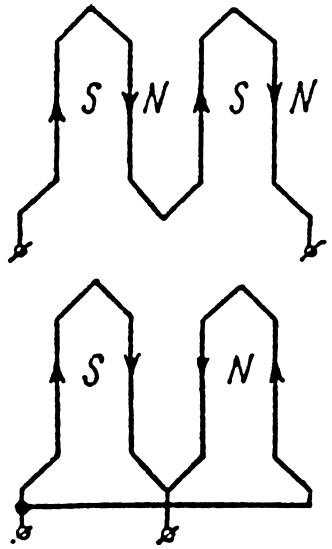


FIGURE 26 Method of stator lead changeover from series to parallel connection of phase winding.

torque is kept constant for corresponding speed steps, because the input power for each case is

$$P_{\mathrm{Y}} = \sqrt{3} V_n I_n \cos\phi_{\mathrm{Y}} \tag{35}$$

$$P_{\mathrm{YY}} = 2\sqrt{3} V_n I_n \cos\phi_{\mathrm{YY}}, \tag{36}$$

where V_n is the rated line voltage and I_n is the rated phase current. The torque for each case is (keeping in mind that $\omega_{\mathrm{YY}} = 2\omega_{\Delta}$)

$$T_{\mathrm{Y}} = P_{\mathrm{Y}}/\omega_{\mathrm{Y}} \tag{37}$$

$$T_{\mathrm{YY}} = P_{\mathrm{YY}}/2\omega_{\mathrm{Y}}. \tag{38}$$

If, as can be assumed, $\cos\phi_{\mathrm{Y}} \cong \cos\phi_{\mathrm{YY}}$, then $T_{\mathrm{Y}} = T_{\mathrm{YY}}$.

In going from the delta to double-wye connection, the torque for corresponding speed steps will change, because the power for each case is

$$P_{\Delta} = 3V_n I_n \cos\phi_{\Delta} \tag{39}$$

$$P_{\mathrm{YY}} = 3\left(V_n/\sqrt{3}\right)2I_n \cos\phi_{\mathrm{YY}}$$

$$= 3.46 V_n I_n \cos\phi_{\mathrm{YY}} \tag{40}$$

and the torque for each case is (keeping in mind that $\omega_{\mathrm{YY}} = 2\omega_{\Delta}$)

$$T_{\Delta} = P_{\Delta}/\omega_{\Delta} \tag{41}$$

$$T_{\mathrm{YY}} = P_{\mathrm{YY}}/2\omega_{\Delta}. \tag{42}$$

Assuming that $\cos\phi_{\Delta} \cong \cos\phi_{\mathrm{YY}}$, then $T_{\Delta} = 1.73 T_{\mathrm{YY}}$.

The speed–torque characteristics of the above cases are shown in Fig. 28.

Beside the two-speed motors described above, pole-changing motors are available up to four speeds. The speed control range usually does not exceed 8:1, because of the need for oversized motor construction. Multispeed motors are used when there is no need for smooth speed control, as is the case with fans, pumps, home washing machines, and so on.

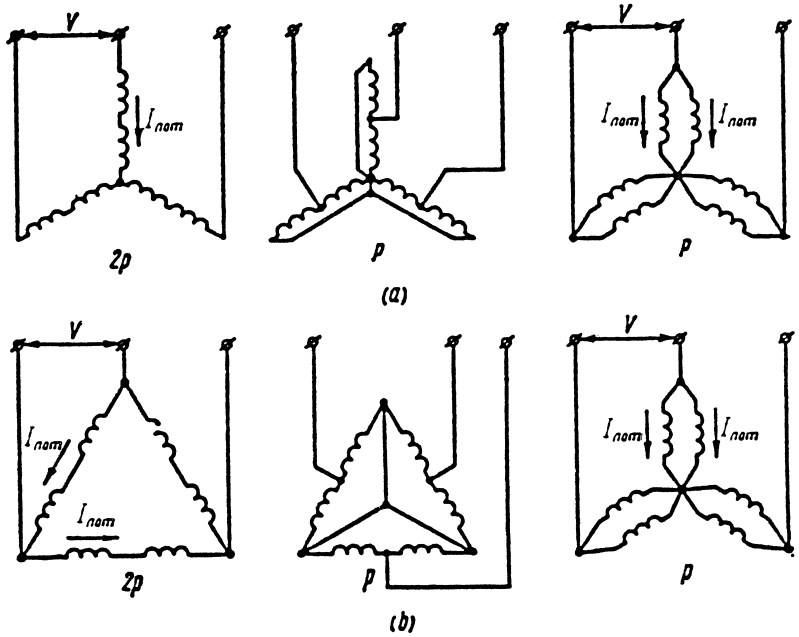


FIGURE 27 Stator winding lead switching for changeover from single-wye, or delta, to double-wye connection.

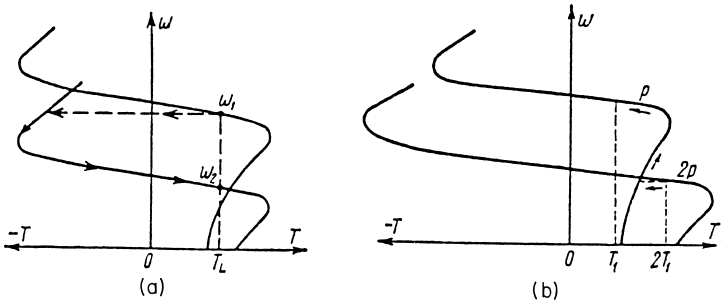


FIGURE 28 Speed–torque characteristics of a two-speed motor: (a) single to double wye, (b) delta to double wye.

Varying the supply voltage to an induction motor affects the magnitude of the electromagnetic torque. According to Eq. (80) of "Electric Machines, Rotating, Construction and Theory," the maximum torque T_{max} is directly proportional to the square of the applied voltage V_{ph}. Using this relationship, one can compute a family of speed–torque curves (Fig. 29).

Since the critical slip S_{max} is independent of the terminal voltage, the maximum torque is obtained at the same speed ω_{max}. The available speed control is limited between $\omega_0 < \omega < \omega_{max}$ (at most 1.5:1 speed range). This mode of speed control finds use in mechanisms having mechanical torques that are proportional to the square of the speed, such as centrifugal pumps and fans.

Varying the supply frequency changes the synchronous speed n_0 of the motor in accordance with Eq. (34). To provide maximum available torque with this mode of control, the air gap flux in the machine has to be maintained constant at all frequencies. Keeping the flux constant also prevents saturation of the iron at low frequencies. To maintain the magnetic flux unchanged, it is necessary to control the supply voltage V_{ph} simultaneously with the frequency, so that

$$V_{ph}/f_1 = \Phi = \text{const.} \qquad (43)$$

Equation (80) of "Electric Machines, Rotating. Construction and Theory" then shows that the maximum torque T_{max} remains constant. This, however, is true only in the high-frequency range in which R_1 can be neglected when compared with X_{sc}.

Since the value of the short-circuit reactance X_{sc} decreases with frequency, in the low-frequency range the above assumption is no longer valid. The voltage drop across the stator resistance becomes significant. To compensate for this change, the ratio V_{ph}/f_1 must be increased as the frequency f_1 decreases. The speed–torque characteristics with frequency control are shown in Fig. 30.

Typical applications of frequency-controlled ac motors are textile machines, certain types of conveyer installations, and traction mechanisms.

2. Wound-Rotor Induction Motor

The three control modes discussed above—change of number of pole pairs. variation of the supply voltage, and variation of the supply frequency—can also be used with wound-rotor induction machines. In addition, a fourth control mode, increase in rotor resistance, can be applied to this machine.

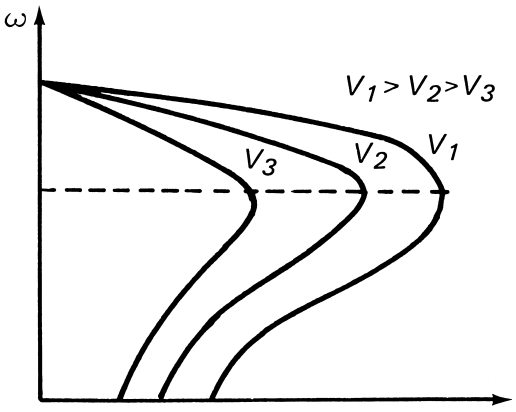


FIGURE 29 Speed–torque characteristics of an induction motor obtained with input voltage control.

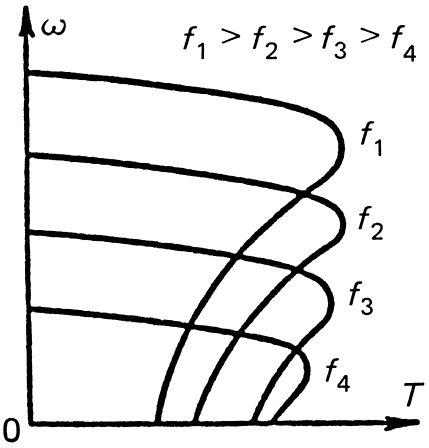


FIGURE 30 Speed–torque characteristics of an induction motor obtained with adjustable-frequency speed control.

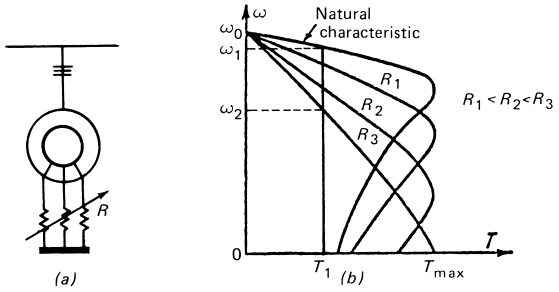


FIGURE 31 Rheostatic speed–torque characteristics of a slip-ring induction motor.

An increase in rotor resistance provides a relatively simple control mode. According to Eq. (79) of "Electric Machines, Rotating, Construction and Theory," the slip $S_{\max}$ at maximum torque increases in proportion to the equivalent resistance R_2' of the secondary circuit. The change in this resistance is accomplished by inserting external rheostats between the rotor slip rings and the short-circuit bar (Fig. 31a).

The value of the external resistance per phase is

$$R_{\text{ext}} = R_2(S_1/S_2 - 1), \tag{44}$$

where

$$R_2 = S_{\text{R}} E_{2\text{R}} \big/ \sqrt{3} I_{2\text{R}}, \tag{45}$$

R_2 is the resistance of the rotor winding, s_{R} is the rated slip, $E_{2\text{R}}$ is the rated rotor open-circuit voltage (between two slip rings), $I_{2\text{R}}$ is the rated rotor current, S_1 is the slip at a given torque on the natural characteristic, and S_2 is the required slip (speed) of the same torque but on the rheostatic curve (Fig. 31b).

An increase in the rotor resistance can provide maximum torque at motor startup (curve R_3 in Fig. 31b). Therefore, this method is widely used when the motor operates in an intermittent duty cycle mode, such as is the case with cranes.

3. Static Converter

Solid-state control units corresponding to each of the above control techniques have been developed.

1. The effective value of the external resistance inserted in the rotor circuit can be varied by introducing a high-frequency chopper across the resistance and by varying the time for which the chopper is on during the cycle.
2. Stator voltage variation is accomplished by means of static ac switches (regulators), which vary the rms value of the ac voltage applied to the motor.

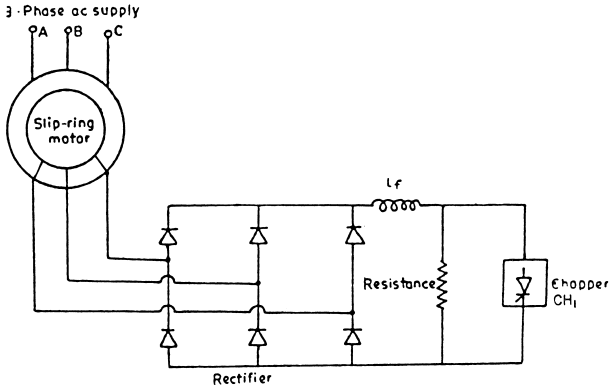


FIGURE 32 Pulsed rotor resistance control.

3. Variable-frequency power is obtained either by means of a cycloconverter, which directly converts ac at fixed frequency to variable-frequency ac, or by an inverter that converts electric power from dc to ac.

Pulsed rotor resistance control can be accomplished by the circuit in Fig. 32. The rotor slip power is rectified in a diode bridge rectifier and fed through a filtering choke to an external resistance. The chopper CH1 connected across the resistor is switched on and off at a relatively high frequency (the switching cycle must be much smaller than the mechanical time constant of the motor). The effective value R_{eff} of the external resistor R_{ext} is determined by the ratio of on time t_{on} to the cycle time T, or

$$R_{\text{eff}} = R_{\text{ext}} t_{\text{on}} / T. \tag{46}$$

Static ac regulators control the rms value of the ac voltage applied to the motor. The static switches chop out a preset portion of each cycle of the supply voltage (Fig. 33). SCR cells are used because they turn off naturally at current zero crossing. Two SCR devices or one triac constitute a single switch to allow a bidirectional control capability. Each is fired once every cycle to obtain the required rms voltage across the motor. The operation can be explained as follows. If we denote the phase angle between the source voltage V_{s} and the motor back emf E_{m} by ψ and neglect the circuit resistance, the maximum current $I_{\max}$ available at full drive (when the two SCRs are in the on state at all times) is

$$I_{\max} = \frac{1}{\omega L} \sqrt{V_{\text{s}}^2 + E_{\text{m}}^2 - 2V_{\text{s}} E_{\text{m}} \cos \psi} \tag{47}$$

and its phase angle is given by

$$\cos \phi' = (E_{\text{m}} / \omega L I_{\max}) \sin \psi. \tag{48}$$

The rms value of the fundamental component I_1 of the current as a function of the firing angle α is

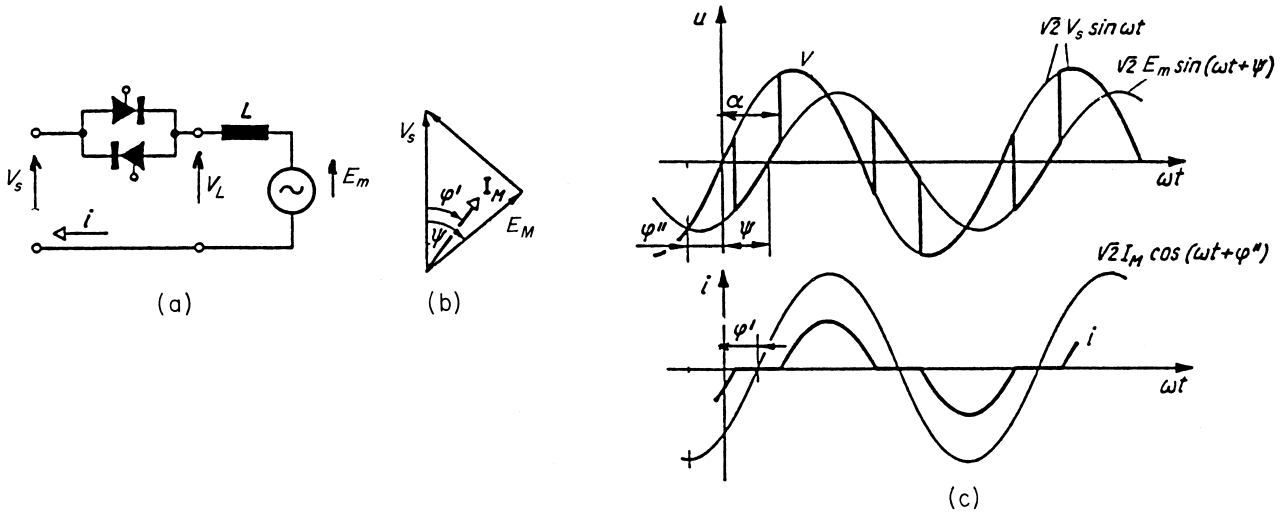


FIGURE 33 Speed control of a single-phase induction motor by static ac regulator: (a) basic circuit, (b) phasor diagram at full drive, (c) voltage and current waves.

$$I_1 = I_{\max}\left(\frac{\pi - \alpha}{\pi} + \frac{\sin 2\alpha}{\pi}\right) \tag{49}$$

and the source power is

$$P = V_s I_1 \cos\phi'. \tag{50}$$

A three-phase ac regulator is illustrated in Fig. 34

The cycloconverter is a direct frequency changer that converts ac power at one frequency to ac power at another frequency. In practice, in order to keep the harmonic content of the output voltage waveform within acceptable limits, cycloconverters are operated in the output frequency range below one-third of the supply frequency.

Since line commutation at the zero crossing of the current is possible, the best available static switch is the SCR (thyristor). Figure 35 shows the circuit diagram and the principle of operation of a single-phase cycloconverter. The circuit is similar to that of a single-phase center-tapped rectifier, except that two groups of controlled rectifiers are provided to permit current reversal through the load. The thyristors S_1 and S_2 provide positive current to the load, and S_3 and S_4 provide negative current. Firing pulses are provided simultaneously to the pair S_1 and S_3 and to the pair S_2 and S_4. As a result, when the positive group (S_1 and S_2) is advanced in firing angle to produce more positive voltage, the negative group (S_3 and S_4) firing angle is retarded to produce less negative voltage.

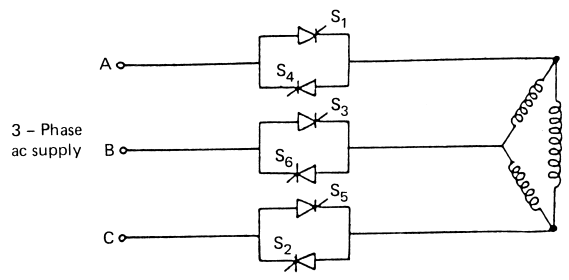


FIGURE 34 Three-phase ac regulator.

With respect to N, the midpoint connection of the transformer, the voltage e_{AN} is in a direction opposite to e_{BN}. As shown in Fig. 35b at time t_0, the load current is negative and flows through S_4. This thyristor, S_4, remains conductive as long as there is a current flow. At t_2, thyristors S_1 and S_3 receive gating pulses, S_4 turns off because $V_{BA} > 0$, and S_3 takes over the load current. At t_3, the load current becomes positive and S_3 turns off while S_1 starts conducting, and so on.

The firing angle is varied with reference to a sine wave having the required output frequency. The amplitude $V_{d\alpha}$ of the mean output voltage in each cycle will vary in accordance with

$$V_{d\alpha} = V_{d0} \cos\alpha, \tag{51}$$

where V_{d0} is the ideal no-load voltage of the rectifier [see Eq. (30)]. The basic circuit can be extended to polyphase operation by simply providing additional controlled elements. The basic three-phase power circuit is shown in Fig. 36.

The capacity of the cycloconverter to handle power flow in either direction together with the capability of reverse polarity (reverse direction of motor speed) provides an induction motor drive capable of four-quadrant operation.

Inverters are used to convert dc to ac power. The dc power can be provided either by a dc source (batteries or dc generators) or through a dc link line. Usually, a

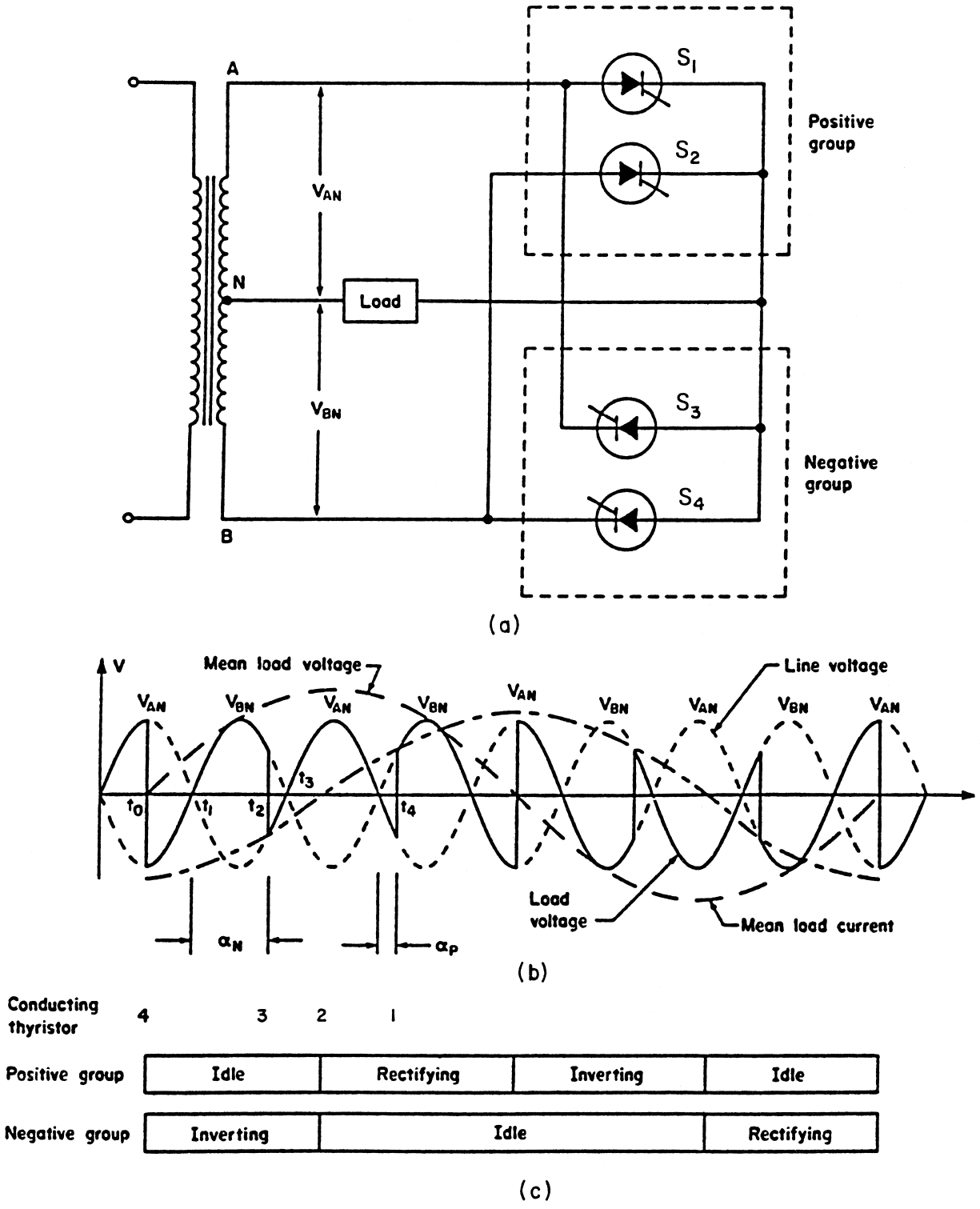


FIGURE 35 Single-phase cycloconverter: (a) basic circuit, (b) load voltage and current, (c) thyristor conducting sequence.

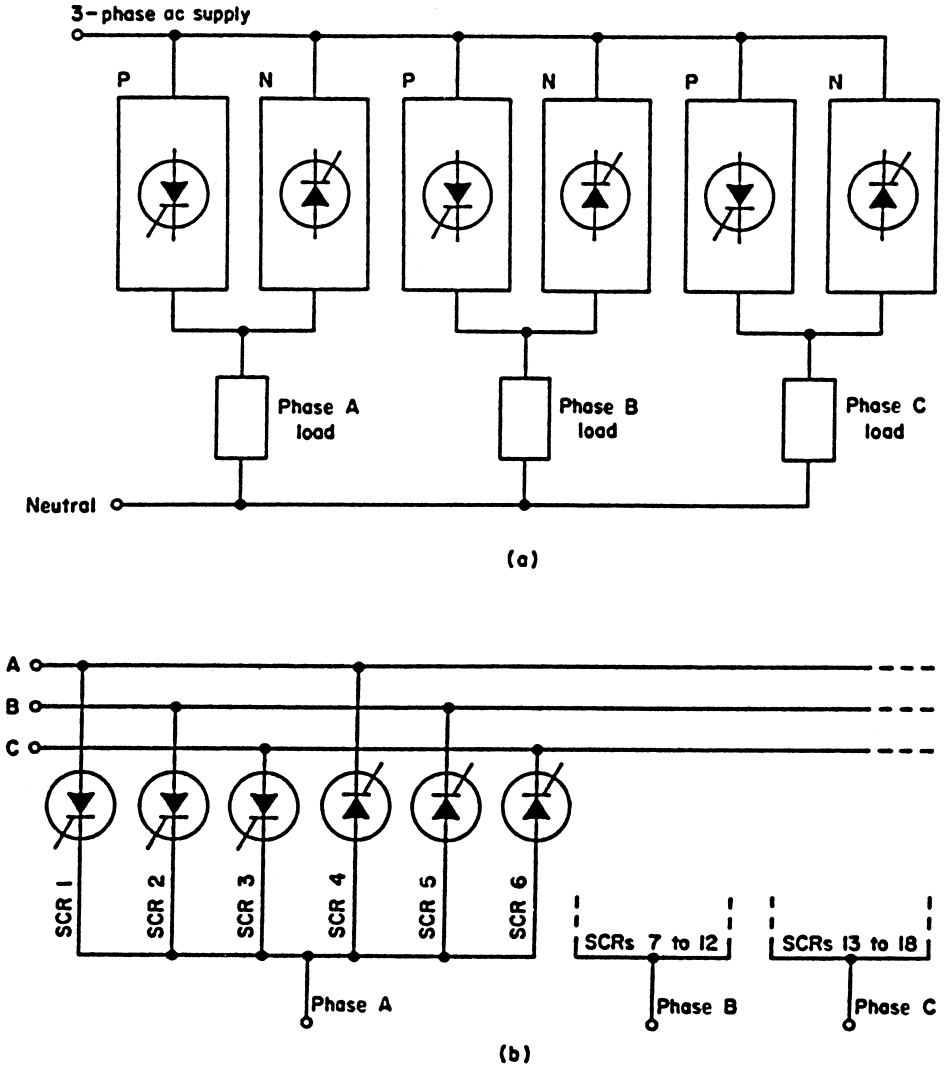


FIGURE 36 Three-phase to three-phase, three-pulse cycloconverter: (a) schematic, (b) basic circuit.

controlled rectifier generates the dc link voltage. Voltage control to obtain a constant volts/hertz relationship can be obtained either by phase control of the rectifier or by internal voltage control of the inverter. The three most commonly used inverter power circuits for ac motor drives are as follows:

1. Voltage control inverter (VCI)
2. Current control inverter (CCI)
3. Pulse-width modulation (PWM) inverter

Figure 37 illustrates the principle of a single-phase inverter. Switching elements S_1 and S_3 are fired at the same

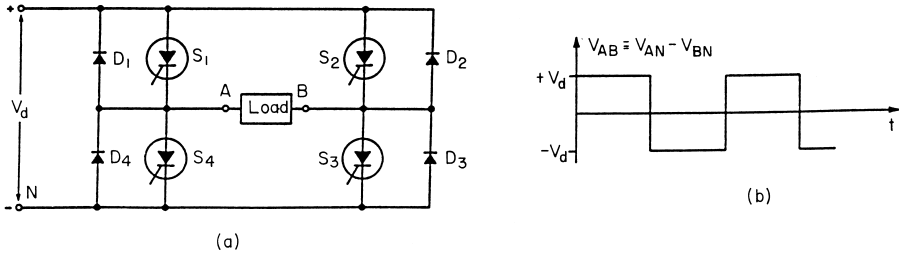


FIGURE 37 Single-phase bridge inverter: (a) basic circuit, (b) load voltage.

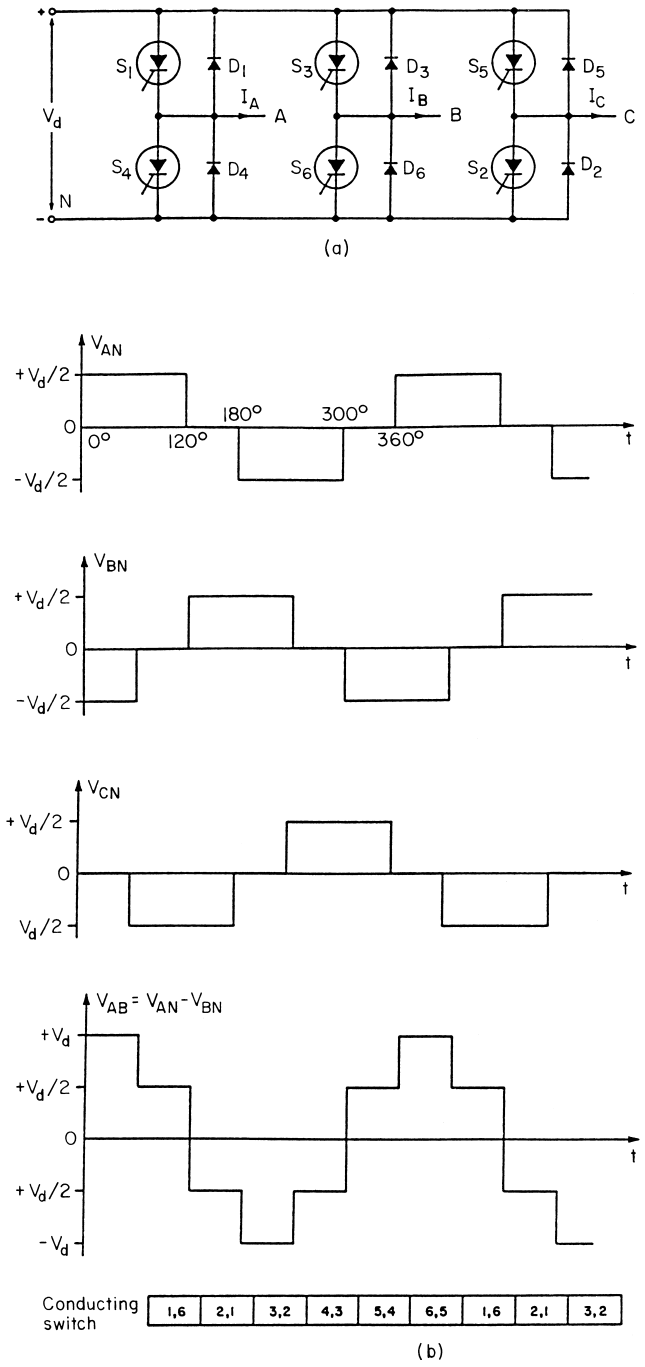


FIGURE 38 Basic three-phase inverter: (a) circuit, (b) voltage waveforms with two switches conducting at the same time.

time, while the other two are turned off. After half a cycle, $t = T/2$, switching elements S_2 and S_4 turn on, and the other two turn off. The voltage across the load is an alternating square wave of peak amplitude V_d. With an inductive load (such as a motor), the load current lags the inverter output voltage. Therefore, freewheeling diodes D_1, D_2, D_3, and D_4 are required to return the reactive energy from the load back to the dc source.

Figure 38a illustrates the simplest form of a three-phase inverter configuration. The switching devices are numbered in the sequence in which they are fired to produce ABC sequence voltages. The output voltage waveforms are given in Fig. 38b.

The first two control techniques, VCI and CCI, are achieved by regulating the voltage or current at the dc link side (input voltage or current). The block diagram

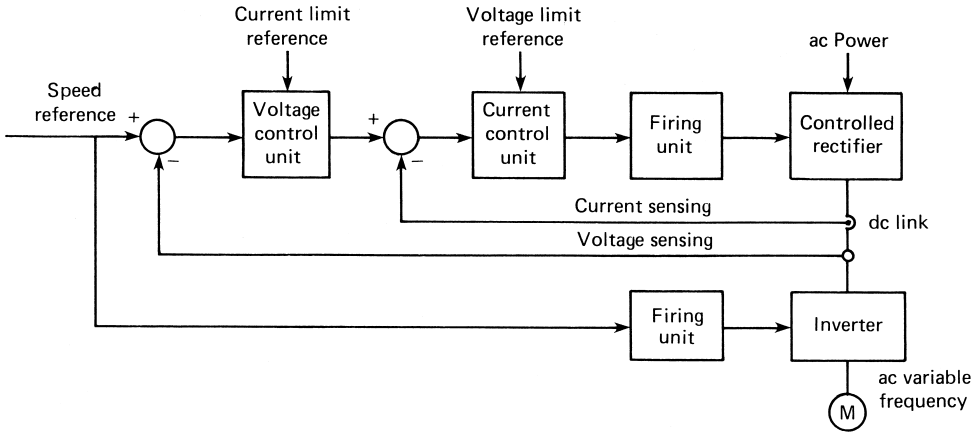


FIGURE 39 Voltage control inverter and current control inverter schematic.

in Fig. 39 illustrates the control technique. The speed reference signal is supplied to the controlled rectifier as well as to the inverter in order to maintain constant the volt/hertz ratio. The system operates as a VCI until the current reaches a limiting reference value and the voltage control unit is overridden. The output voltage is adjusted according to the motor speed. When the current limit is reached, the current will remain constant independent of the speed, so that the system operates as a CCI.

The PWM system controls the ac output voltage, as well as the frequency, within the inverter unit. Therefore, a constant-voltage source can be utilized. This is either a battery or a rectifying unit employing simple diodes. Each power switching device of the inverter unit is a dc chopper (GTOs, power transistors, etc.). By commutating one side of the bridge (Fig. 37) several times during a half-cycle, a train of constant-amplitude pulses is produced as shown in Fig. 40.

The amplitude of the individual pulses in the inverter is the constant input voltage V_d. The magnitude of the fundamental wave of the output voltage is controlled by varying the ratio of the total on time of the pulses to total off time. This ratio is controlled by varying the pulse width for a constant number of pulses per half-cycle. The repetition rate of the pulses is known as the carrier frequency. The pulse-width variation is accomplished in a cyclic manner by comparing a sine wave of the required output frequency against a synchronizing triangle carrier. The ratio of the carrier frequency to the fundamental frequency should be maintained as high as possible in order to reduce the total harmonic content.

4. Position Control

Position is expressed in terms of a displacement, which may be linear or angular, depending on the kind of motion. All types of actuators or motors can be used for position control. However, the most popular motor type is the stepper motor described in Section V of "Electric Machines, Rotating, Construction and Theory," because of the ease with which it can be controlled digitally.[1]

Since all mechanical systems involve some intertia, the velocity profile, or duty cycle, of a positioning operation includes acceleration and deceleration phases, as well as an interval of time at constant speed. It is not surprising that the velocity profile should resemble that of a rapid-transit car shown in Fig. 2 of the article "Electric Machines, Design," because a transportation system can be viewed as a positioning drive. It is noted, however, that in stepper motor parlance the term "slew" replaces "cruise" in defining the constant velocity interval of Fig. 2.

The main task of a positioning system is to produce the desired displacement in the shortest possible time. The functional dependence of the displacement on time is obtained by integrating twice the equation of motion, Eq. (3). Complications arise, because the torque developed by the motor is a function not only of the excitation current i, but also of the angular displacement θ of the rotor according to V Eq. (88). Moreover, the resistive torque includes not only the load torque, whose character varies with the application, but also two friction terms, one of which is proportional to the speed and the other (Coulomb friction) which is independent of speed and always opposes the velocity.

For this reason, the dynamical performance of the drive can be predicted accurately only by using numerical computations. However, some understanding can be gained by examining the quasi-steady-state operation, with various loads. Referring to V Fig. 39, this can be plotted as in Fig. 41. In all cases, the shaded areas above and below

[1]All references to Section V in that article will be denoted by the symbol V followed by the equation number, when needed.

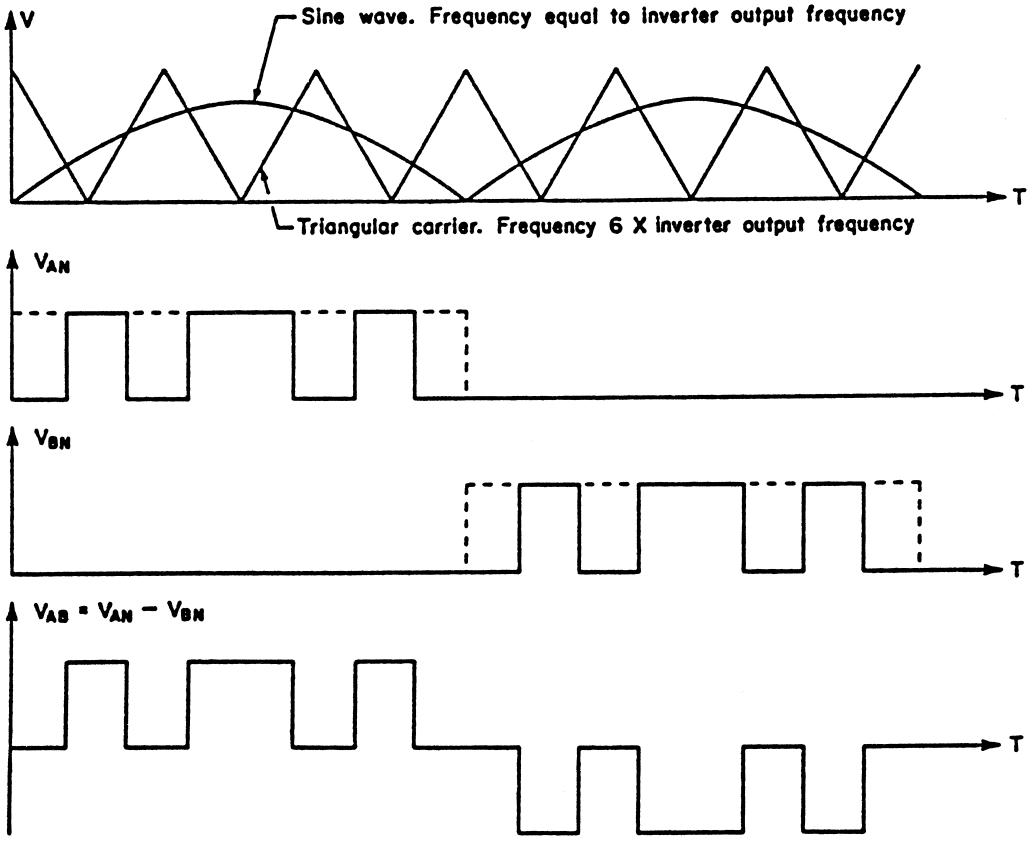


FIGURE 40 Pulse-width modulation by comparing a triangle carrier with a sine wave.

the resistance torque must be equal in order to yield a constant average velocity. Instantaneously, however, the drive undergoes a periodic cycle of accelerations and decelerations, so that the velocity has a ripple superimposed on its average value.

The no-load situation is depicted in Fig. 41a. In this case, the machine operates as a motor in the positive torque-position interval and as a generator in the negative one.

The limiting case in which the torque never becomes negative is shown in Fig. 41b. The equal area criterion yields an average value of the developed torque equal to 63.7% of its peak. The opposite case of pure regenerative braking is shown in Fig. 41c. The angle at which the developed torque attains its average value is called lag, torque, or power angle. It is referred to the stable zero torque position and denoted by θ_{ave} when measured in mechanical degrees and by $\delta = N\theta_{ave}$, where N is the number of electrical cycles per mechanical revolution, when measured in electrical degrees. It is equivalent to the power angle in synchronous machines, as defined in Section V.C of "Electric Machines, Rotating, Construction and Theory." Accordingly, the average torque can be expressed as

$$T_{ave} = -T_{max} \sin N\theta_{ave} = -T_{max} \sin \delta. \qquad (52)$$

In the cases depicted in Figs. 41b and 41c, $\delta = \pm 39.58^{\circ}$, respectively. The case of maximum, or pull-out, torque operation is shown in Fig. 42. According to V Eq. (88) the torque and, therefore, its pull-out value are proportional to the square of the current. With constant voltage this is a function of the supply frequency which, according to V Eq. (86), is, in turn, proportional to the speed. It follows that, with a constant voltage supply, the pull-out torque varies with the speed. In general, the pull-out torque vs speed is determined by both the motor and the power conditioner. A typical curve is shown in Fig. 42b. Knowledge of such curves is needed in order to design the drive, because they define the boundary beyond which the step motor definitely looses synchronism.

As was already mentioned, a positioning cycle includes an acceleration and a deceleration phase, besides the constant or slew velocity interval. During acceleration, the average developed torque lies above the resistive torque, and during deceleration, it lies below the resistive torque. The general trend of such a cycle of torque vs position is shown in Fig. 43.

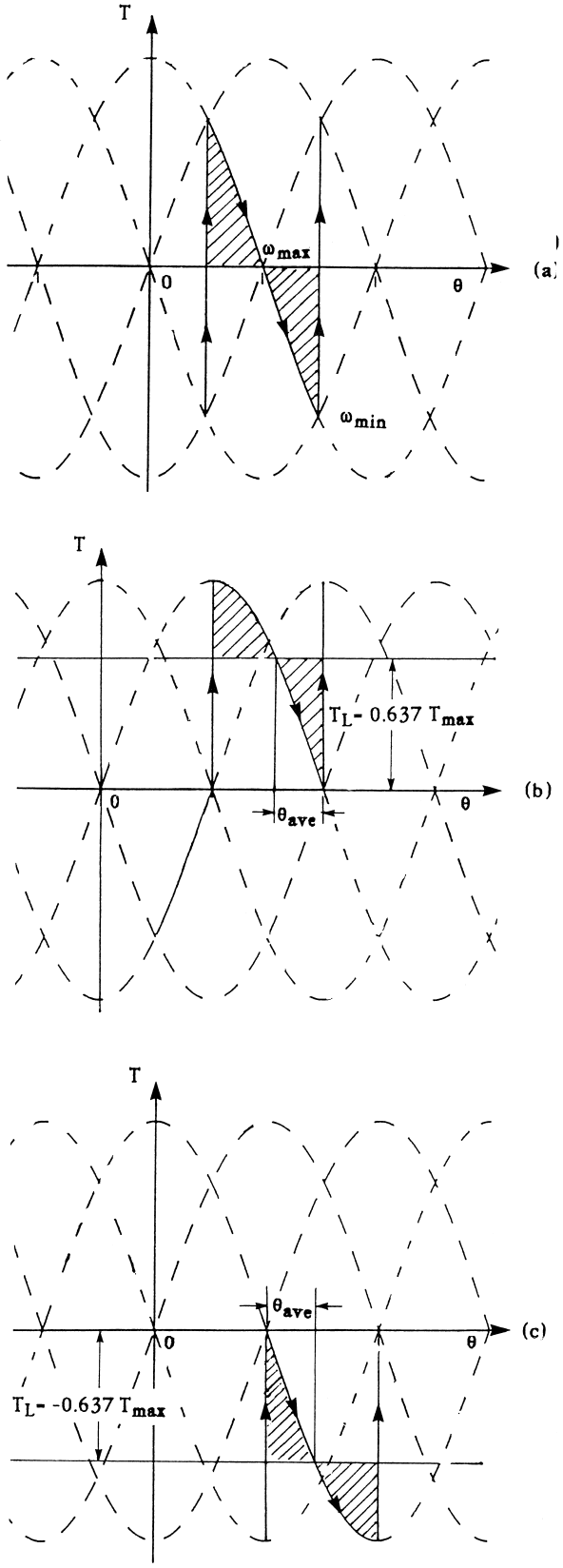


FIGURE 41 Constant speed operation: (a) no load, (b) motoring, (c) regenerative braking.

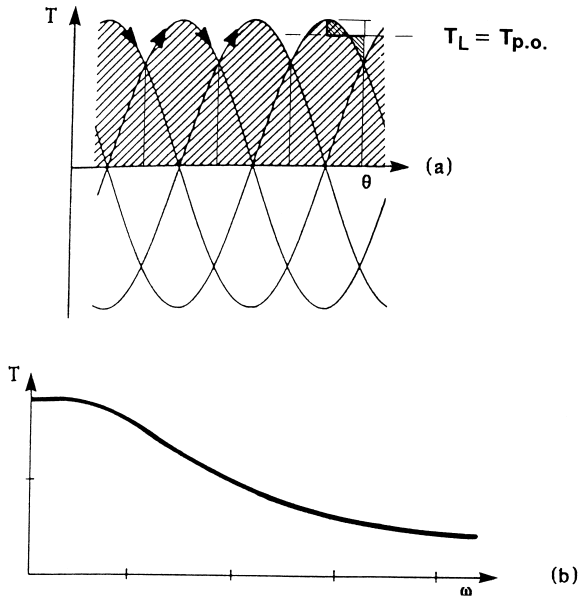


FIGURE 42 Pull-out torque operation: (a) torque vs displacement angle, (b) torque vs speed.

In all these considerations, electrical and mechanical transients have been ignored. However, mechanical and electrical resonances and inadequate damping can cause noise, vibration, loss of synchronism, or missed steps. Therefore, in the past it was recommended that positioning drives should be designed so that in normal operation the torque angle did not exceed 30°. However, the use of sophisticated power conditioning circuits, the addition of internal or external dampers, closed-loop control by means of feedback loops, and the speed by which logic functions can be performed have brought about significant improvements in performance in terms of accuracy and torque output. An example of a control technique that results in finer positioning resolution is *microstepping*. By increasing and decreasing the current in each phase winding in small steps, instead of using a simple on/off control, one may create intermediate steps between the motor's "cardinal" steps and reduce the velocity ripple.

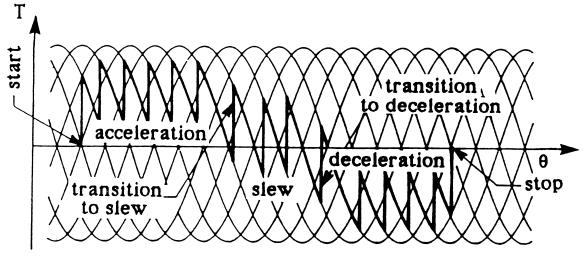


FIGURE 43 Torque vs displacement angle in a complete positioning operation.

SEE ALSO THE FOLLOWING ARTICLES

ELECTRIC MACHINES, DESIGN • ELECTRIC MACHINES, ROTATING, CONSTRUCTION AND THEORY • MACHINE DESIGN • POWER ELECTRONICS

BIBLIOGRAPHY

Chilikin, M. (1978). "Electric Drive," 3rd ed., Mir Publishers, Moscow; translated by J. Feinberg.

Datta, S. K. (1985). "Power Electronics and Control," Prentice Hall, Reston, VA.

Krause, P. C., and Wasynczuk, O. (1989). "Electromechanical Motion Devices," McGraw-Hill, New York.

Lechouts, A. C. (1987). "The Art and Practice of Step Motor Control," Intertec Communications Inc., Ventura, CA.

Leonhard, W. (1996). "Control of Electrical Drives," 2nd ed., Springer-Verlag, New York.

Mohan, N. (2000) "Electric Drives; An Integrative Approach," MNPERE, Minneapolis.

Pillai, S. K. (1982). "A First Course on Electrical Drives," Wiley Eastern Limited, New Delhi, India.

Vithayathil, J. (1995). "Power Electronics," McGraw-Hill, New York.

Wildi, T. (2000). "Electrical Machines, Drives, and Power Systems," 4th ed., Prentice Hall, New York.



Fiber-Optic-Based Smart Structures

Raymond M. Measures
University of Toronto Institute for Aerospace Studies

I. Introduction
II. The Need for Built-In Fiber-Optic Sensor Systems
III. Structurally Integrated Fiber-Optic Sensors
IV. Single-Mode Optical Fiber Sensor Theory
V. Structural Measurements Possible with Fiber-Optic Sensors
VI. Damage Assessment with Structurally Integrated Optical Fibers
VII. Future Prospects

GLOSSARY

Composite material A fiber-reinforced polymer generally divided into thermosets and thermoplastics. Graphite of Kevlar fibers in an epoxy matrix are examples of a thermoset; one of the most popular thermoplastics involves graphite fibers in a PEEK matrix.

Fiber-optic sensor A length of optical fiber in which the properties of light are influenced by some transduction mechanism. For example, if an interferometric fiber-optic sensor is subjected to changes in strain or temperature, there is a corresponding change in the phase of light propagating through the optical fiber.

Fly-by-light The use of light signals transmitted through optical fibers to convey sensor and control information between the cockpit and the various systems of an aircraft.

Intelligent structure A smart structure that is capable of adaptive learning.

Optical fiber A transparent fiber comprising a central core and an outer cladding. Light launched into the core at one end will be confined to propagate through the core with very little attenuation.

Passive smart structure A structure that possesses a built-in optical microsensor system for evaluating its state in real time.

Reactive (or adaptive) smart structure A structure that possesses a built-in optical microsensor system for evaluating its state in real time and an actuator control system for modifying that state when necessary.

Structurally integrated fiber-optic sensor A fiber-optic sensor that is either bonded to the surface of a structure or, in the case of a composite material, embedded within the structure.

A SMART STRUCTURE possesses a structurally integrated optical microsensor system for determining its state. This built-in sensor system can be thought of as an optical nervous system which can, in real time, evaluate the strain or deformation of the structure, monitor if it is vibrating or subject to some external pressure, check its temperature, and warn of the appearance of any hot spots. Most important of all, a smart structure should maintain a constant vigilance over its structural integrity and indicate if damage is inflicted or fatigue is developing. In more advanced smart structures the information provided by the sensor system could be used for controlling some aspect of the structure, such as its stiffness, shape, position, orientation, rotation, or acceleration; such systems might be called "reactive smart structures" to distinguish them for the simpler "passive smart structures," which only sense their state. Eventually, smart structures will be developed that will be capable of adaptive learning, and these can be termed "intelligent structures." Indeed, if we consider the confluence of the four fields, materials and structures, sensor systems, actuator control systems, and neural network systems, we can appreciate that there is the potential for a broad class of structures (Fig. 1). According to this designation, *sensing structure* would replace the term *passive smart structure*.

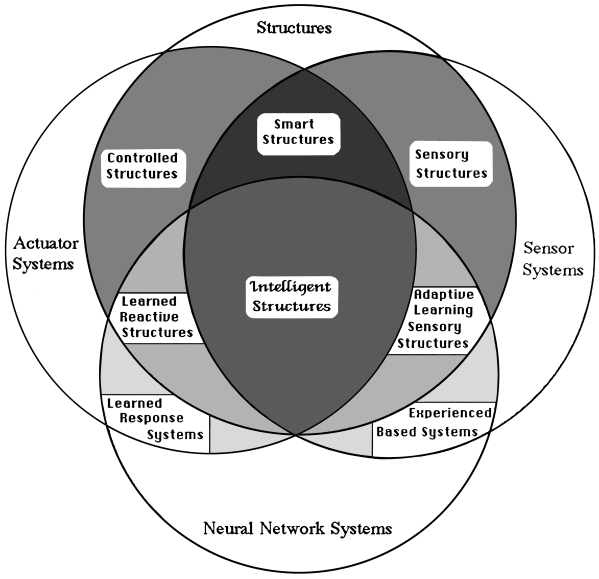


FIGURE 1 Types of structures possible by the confluence of structures, sensor systems, actuator control systems, and neural networks.

The successful development of smart structures technology could lead to aircraft that are safer, lighter, more efficient, easier to maintain, and easier to service; pipelines, pressure vessels, and storage tanks that constantly monitor their structural integrity and immediately issue an alert if any problem is detected; space platforms that check for pressure leaks, unwanted vibration, excess thermal buildup, and deviation from some preassigned shape. In many situations this optical sensing network would be part of a control system that could take some appropriate action. Such structures might also learn by means of a neural network the correct response for achieving a desired movement. In the 21st century this could lead to aircraft without flaps or ailerons that would instantly adjust their shape to conform with the flight conditions and to large flexible space structures that would actively damp out unwanted vibrations and maintain some preassigned shape. The development of structurally integrated optical microsensors combined with distributed actuator systems could also lead to the manufacture of flexible structures that are more adaptable and provide much finer motion control than their more rigid counterparts. Flexible manipulators built on these principles would endow robotic systems with human-like control capabilities.

The considerable potential of smart structure technology in a broad range of industrial sectors, from aircraft to bridges and spacecraft to submarines, is gaining recognition and, as consequence, is being taken seriously by industry and governments in the United States, the European Community, and Japan. Several substantial research programs have been initiated within the past few years, and the Japanese have identified "intelligent materials" as one of their national goals for the 21st century.

I. INTRODUCTION

Optical fibers can either be bonded to the surface of metal, plastic, ceramic, or other structures or embedded within composite material structures during their manufacture. These sensors could be used to detect and locate impacts, anomalous hot spots, damage, and structural deformation. In the case of composite materials, internal damage generated by impacts, manufacturing flaws, excessive loading, or fatigue could be detected and assessed, and, most important, the growth in the extent of these damage zones could also be monitored. Structurally embedded optical fibers sensors could undertake "realtime" monitoring of the cure state in thermosets or the degree of consolidation in thermoplastics. The residual stress during manufacture could be determined, and, subsequently, the structure could be examined for internal defects both prior to handling and

TABLE I Types of Fiber-Optic Sensors

Category	Single optical fiber	Localized sensing	Single ended	Sensitivity
Interferometric				
Mach–Zehnder	No	Yes	No	High
Michelson	No	Yes	Yes	High
Fabry–Perot	Yes	Yes	Yes	High
Bragg grating	Yes	Yes	Yes	High
Polarimetric				
Low birefringence	Yes	No	Yes	Low
High birefringence with 45° splices	Yes	Yes	Yes	Low
Intensiometric				
Bistable	Yes	No	No	Poor
Microbend	Yes	Yes	No	Low
Modified cladding/ core	Yes	Yes	Yes	Low
Modalmetric				
Few mode	Yes	Yes	No	Low
Elliptic core	Yes	Yes	No	Low
Twin core	No	Yes	No	Low

after installation (see Table I). This could lead to improvements in the quality control of composite products.

Two of the most important considerations for aircraft are system weight and life cycle costs. In-service monitoring of structural loads, especially in high load regions like wing roots or aircraft, could be of considerable benefit in helping to avoid structural overdesign and in evaluating the true fatigue life based on the actual load history of each component. Monitoring the dynamic response of structural components during operation might also provide insight into the structural integrity of the system.

Protection against lightning and various forms of electromagnetic interference also figure prominently in these deliberations. Unfortunately, where composite materials are being used to replace metal, fly-by-wire systems require extensive shielding. Even in the absence of a direct lightning strike, magnetic field coupling can generate voltages that can burn out various components and fragile circuits used in many electronic systems. Although these problems can be overcome, even in the hostile environment in which a military aircraft might find itself, a considerable weight penalty is associated with the appropriate sheilding. The use of optical fibers for the transmission of sensor, encoder, and flight control signals would alleviate this problem. Indeed, in many instances optical fibers can serve as the sensors, encoders, and the data and control signal conduits (Fig. 2).

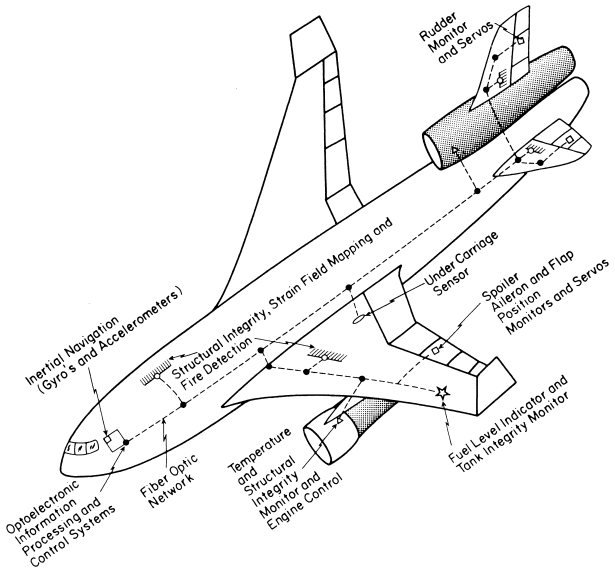


FIGURE 2 Potential fiber-optic systems within future aircraft.

II. THE NEED FOR BUILT-IN FIBER-OPTIC SENSOR SYSTEMS

In the 1970s new approaches to complex engineering systems like aircraft were adopted. These assume that each major component contains one or more critical flaws at a level not detected initially. This led to new engineering strategies, such as (1) lifetime system monitoring, (2) retirement for cause, and (3) designed-in inspectability. The earlier approach of retiring an entire batch of structural components once a potentially critical flaw (crack) was discovered in one component was enormously costly. Retirement for cause involves periodic inspection of each component with expensive and elaborate NDE systems. Once a critical flaw is detected, that component is retired. Although this aircraft structural integrity program (ASIP) has been successful in preventing loss of aircraft from structural failures, it is expensive and not always accurate. Furthermore, it leads to greatly increased downtime and a commencerate increase in life cycle costs. The introduction of components with built-in inspectability should significantly reduce costs, improve maintenance, and shorten downtime. In addition, monitoring the loads and deformation of a component on a continuous basis will permit its true fatigue life to be estimated at any time.

A. Composite Materials in the Aerospace Field

The special properties of composite materials, such as their high strength and stiffness-to-weight ratio, are leading to their increasing use by the aerospace industry. These attributes not only provide a direct incentive, but have a

"multiplier" effect that operates to reduce fuel consumption and engine weight. In addition, these materials possess other desirable properties, such as low thermal expansion and resistance to fatigue and corrosion. Furthermore, their elastic and thermal properties can be tailored. Although initially composite materials were used only for secondary structures and control surfaces, recently some military and civilian aircraft have used them for primary structures.

Unfortunately, the very nature of these fiber-reinforced polymers makes them prone to certain kinds of damage. In particular, fiber-reinforced composite materials are very susceptible to impact loading, because the epoxy matrix is at least an order of magnitude weaker than the embedded fibers of carbon or Kevlar. The damage characteristics of these materials when subjected to subperforation impact can be classified as indentation, fiber breakage, matrix cracking, fiber-matrix debonding, and interply disbonding, better known as "delamination." Among these damage modes, *delamination* has been found to be the most serious. This form of damage can be induced by relatively low energy impacts and is particularly insidious, as it can reduce both the tensile and the compressive strength with very little visible surface damage. This has given rise to the commonly used acronym BVID, for barely visible impact damage. Such impacts can result from a carelessly dropped tool, runway stones, hail, or a collision with a bird, and once started the region of damage can grow with load cycling incurred in normal operation.

Although rigorous structural testing is used to ensure that any aircraft structure put into service is safe and can withstand the service loads and environments, degradation of composites does occur due to moisture, chemical attack, thermal spikes, fatigue, overloading, erosion, impact damage, and lightning strikes. Accidental damage can also arise during transportation, installation, and operation. Consequently, a large proportion of aircraft maintenance time is currently spent inspecting and testing composite parts for manufacturing and service-induced defects and cracks. Because visual inspection is of such limited value in these materials, a variety of nondestructive evaluation techniques have been developed for the purpose of assessing damage to composite structural components.

B. Virtues of a Built-In Structural Integrity System

Unfortunately, all conventional damage assessment techniques are limited to ground use and require the structure to be taken out of service. Most techniques, including the more reliable ultrasound C-scan and X-ray techniques, are not practical in the field and often require disassembly and isolation of the component involved. When hand-held scanning systems are employed, the inspections are time consuming and require highly skilled technicians.

Clearly, the development of a *built-in* damage assessment system that could constantly check the structural integrity of each important or critical component could greatly alleviate concerns over the introduction of composite materials, as well as substantially reduce the cost of their use in aircraft due to savings in downtime for inspection. For composite materials, a built-in monitoring system could be based on a grid of optical fibers embedded within the structure at the time of manufacture. The simplest form of such a system would involve the interrogation of the optical fibers by a source of light, and disruption in their transmission would be used to determine the location and extent of damage. This could certainly be used in the case of damage resulting from impacts, but it could also be used more generally. Growth of a damage zone resulting from fatigue, mishandling, excessive load, or flaws created during manufacture could also be detected by this system. Structural integrity monitored by an embedded optical fiber array could also be of considerable benefit to many other structures that rely heavily on composite materials, such as storage tanks, pressure vessels, submarines, and pipelines.

C. Benefits of Smart Adaptive Structures

Control over the shape, deformation, motion, and vibration of a structure is required in a wide range of applications. The aerospace and robotic fields are particularly rich in the need for structural control. The most graphic examples are aircraft wings, flexible space structures, and robotic manipulators. If an external system of measurement (such as a camera) is used to control the shape or motion of a structure, then according to the Venn representation portrayed in Fig. 1 we have a "controlled structure." A smart adaptive structure, on the other hand, includes both structurally integrated sensing and actuation control systems.

Actuation control can be achieved by means, of a central or distributed system. This can involve either large, high authority central actuators or a distributed set of relatively small actuators. In the latter case the concept of centralized and distributed control can also be considered in terms of physical and algorithmic attributes. A physically centralized control system involves a central location for the control logic circuitry and mechanisms and implies that the control electronics are remote from the sensor–actuator location. A control structure where each actuator is driven by information from a single or local set of sensors exclusive to that actuator represents the limiting case of distributed control architecture. Human beings use both central and distributed control. The former relies on

conscious control of the muscles; the latter (termed "reflex action") involves direct control of muscles by local nerves through the spinal column. The operationally robust nature and potentially fast response of distributed control system represent two of their strongest virtues. Graceful degradation is assured by the fact that the failure of one or more components does not lead to complete loss of control.

Many current and proposed space operations involve the use of relatively large structures, including antenna reflectors, solar arrays, instrument booms, and trusses. These structures are lightweight and are often deployable to accomodate launch vehicle constraints. The lightweight nature of these structures make them highly flexible, with low resonant frequencies that are easily excited. This requires vibration control or suppression in many instances. In the case of large antenna reflectors, there are stringent shape requirements which will require active control. In some instances deployment to wavelength order of magnitude tolerances is required.

In the case of space station *Freedom*, a central truss beam of about 50 m in length supports two flexible solar arrays attached to each end. Excitation of this structure is inevitable, with continual low level disturbances caused by the activities of the crew and occasional events such as shuttle dockings and orbit adjustments. Shape control is also of paramount importance for future satellites such as space-based radar and radarsat. These spacecraft have planar arrays between 20 and 200 m^2 in area yet may require surface control to a tolerance of 1 mm or less.

There is growing interest in the possibility that future aircraft flight control systems could be based on smart adaptive wings. This would allow optimized aerodynamic profiling that would provide gust alleviation, flutter suppression, and stability augmentation for unstable airframes. Smart adaptive structures might also facilitate redundant flight control, wherein sudden failure or impact damage to some fraction of the control system can be compensated for by temporary structural changes. Smart adaptive wings could also lead to a concomitant reduction in maintenance, repair, and downtime due to their greatly reduced mechanical (articulating) complexity.

Military aircraft are forever pushing at the boundaries of high performance, and today that comes at the cost of aerodynamic stability. Some of these aircraft are almost impossible to fly without sophisticated flight controls that continuously sense the flight environment and act to keep it stable.

Another attractive feature of fiber-optic sensors that are embedded within the structure (or skin) of an aircraft is that they can perform their sensing function without affecting the external aerodynamic shape. This would be particularly important at the supersonic and hypersonic regimes currently being seriously considered even for civilian vehicles and probably represents an extreme example of where vigilant monitoring will be required.

In the case of future aircraft, smart structure technology will eventually

- Provide a preflight status report on integrity and airworthiness
- Monitor and record flight loads and the environment
- Warn of structural integrity problems during the flight
- Reduce vibration and flutter
- Alleviate gust response
- Provide real-time corrective action in response to mechanical failure through a mission adaptive system
- Schedule postflight repair and maintenance

III. STRUCTURALLY INTEGRATED FIBER-OPTIC SENSORS

Optical fibers are usually a little thicker than a human hair and basically comprise three cylindrical regions: a *core*, where the refractive index is a little higher than in the surrounding *cladding*, and a *buffer*, or protective jacket (Fig. 3). Light launched into the core of an optical fiber is confined and guided over considerable distances. This has led the communication industry to gradually replace electrical cables with optical fibers, with the substitution of photons for electrons as the information media. Multimode optical fibers have a core radius that is large enough for light to propagate in many different patterns or modes, and an appreciable amount of power can be transmitted. Single-mode optical fibers, on the other hand, have a very small core radius (comparable to the wavelength of light) and transmit only the lowest order transverse mode.

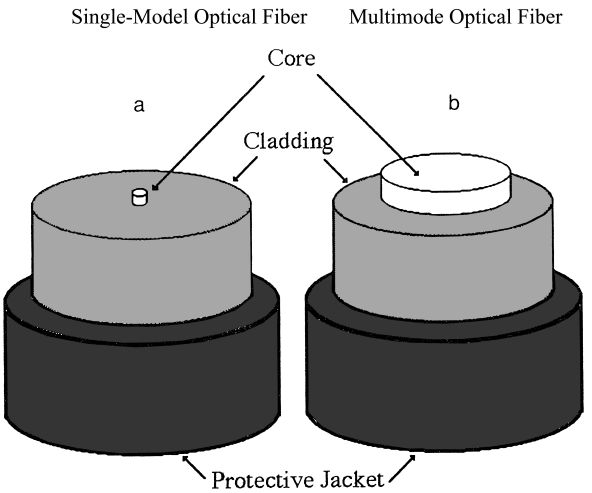


FIGURE 3 Single- and multimode optical fibers.

A. Merits of Fiber-Optic Sensors

Optical fibers possess many attractive features that make them eminently suitable to serve as the sensory system for a smart structure. In essence, they are

- Light in weight and very small
- Capable of high bandwidth
- Immune to electromagnetic interference
- Inert and corrosion resistant
- Safe and unlikely to initiate fires or explosions
- Capable of making very sensitive sensors
- Embeddable within composite materials
- Relatively nonperturbing in regard to structural properties
- Multifunctional
- Compatible with optical bus or network
- Amenable to multiplexing and signal processing with integrated optics

Structurally integrated fiber-optic sensor (SIFORS) technology is illustrated in the generic view presented as Fig. 4. Light is launched into a grid or array of optical fiber sensors that are either adhered to the surface of a metal structure or embedded within a composite material structure, through some input interface (connector). Information about the state of the structure is impressed upon the transmitted (or in many instances the reflected) light by a number of mechanisms. These include interactions that change the intensity, phase, frequency, polarization, wavelength, or modal distribution of the radiation propagating along the optical fiber. What makes the fibers particularly attractive for embedding within composite structures is that for the most part they serve the dual role of sensor and pathway for the signal, and since they are dielectric in nature, they are compatible with the composite materials, avoiding the creation of electrical pathways within the structure. This will become very important as composites are used increasingly in aircraft that have to fly through thunderstorms.

B. Types of Fiber-Optic Sensor

Fiber-optic sensors can be classified into four major categories: intensiometric, interferometric, polarimetric, and

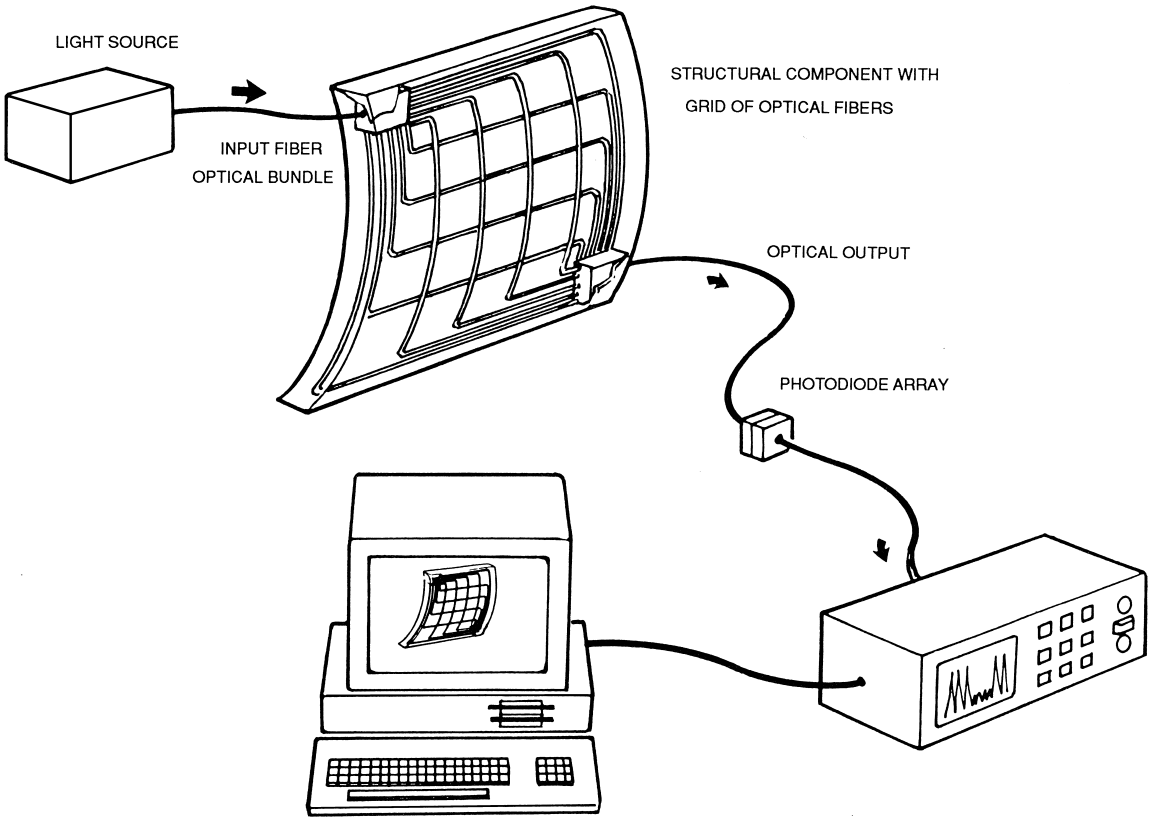


FIGURE 4 Generic fiber-optic smart structure involving a light source, an input optical fiber (or bundle), the optical grid or sensor array, the photodetection system, and the interrogation system with a computer and monitor.

modalmetric. The type of fiber-optic sensor selected will primarily depend on the application. If more than one type of optical fiber sensor could be used for a given measurement, then secondary considerations, such as the ease of integrating the sensor into the structure, might serve to determine the selection. Three questions pertinent to this selection are whether only a single optical fiber is required per sensor; whether the sensing region can be localized (i.e., made with insensitive lead-in and lead-out optical fibers); and whether the sensor can be single ended. The sensitivity is also very important. Some idea of the relative sensitivity of the various sensors can be gauged by judging them in terms of their sensitivity to strain. Table I provides a finer subdivision of fiber-optic sensors and a comparison of their relative attributes.

Intensiometric sensors depend on a variation of the radiant power transmitted through a multimode optical fiber. The simplest form of this sensor involves the presence or absence of light (a bistable approach). For example, fracture of an optical fiber can serve as the basis of a damage-sensing system. It has long been recognized that small bends in an optical fiber lead to a loss of its light, as the curvature permits core radiation to leak into the cladding, from which it escapes the fiber. The smaller the radius of curvature, the greater is the loss. Microbend fiber-optic sensors rely on this principle, using a small device to convert an applied force into a microbend in the optical fiber and thereby reduce its transmission of light. Optical time domain reflectometry (OTDR) uses a time-of-flight measurement to not only detect but also to locate breaks in an optical fiber by means of Fresnel reflection from the fracture. Other types of intensiometric fiber-optic sensors are based on fluorescence, Raman scattering, and blackbody emission. Optical fibers with modified segments of either the core or the cladding can also serve as sensors. These are used often to measure temperature, and Afromowitz and Lam (1988) demonstrated that an index-matching effect can be used for composite cure monitoring.

Interferometric sensors represent a large class of extremely sensitive optical fiber sensors. These primarily function by sensing the phase change induced in light that has propagated along a single-mode optical fiber. There are five basic types of interferometric fiber-optic sensors: Mach–Zehnder, Michelson, Fabry–Perot, Bragg, and the Sagnac. The latter is primarily used in the development of fiber-optic gyros and for this reason is not included in Table I. The Michelson uses two closely spaced, single-mode optical fibers, where one optical fiber serves as a reference. The sensing region is localized between the mirrored ends of the two optical fibers (Fig. 5), and interference associated with changes in the strain or temperature of the structure modulates the intensity of light incident on the detector. The coupler acts to mix the light from the two optical fibers. The differential Mach–Zehnder fiber-optic sensor is primarily a transmissive-based interferometer that also involves two closely spaced optical fibers, where localization is achieved by the introduction of an additional element of length in the sensing optical fiber. If both the input and the output interfaces are on the same side of a structure, the optical fibers form two closely spaced loops and the sensing region becomes the small difference in the length between these two loops.

A Fabry–Perot fiber-optic sensor involves a single monomode optical fiber with a sensing region defined by a cavity comprising two mirror surfaces that are parallel to each other and perpendicular to the axis of the optical fiber. A change in the optical path length between the mirrors leads to a shift in the frequencies of the cavity modes. In some ways the Fabry–Perot cavity represents the most attractive interferometric sensor and has several very attractive features: the reference and sensing optical fiber are one and the same, up to the first mirror that constitutes the start of the sensing region; it can form a single-ended configuration (see Fig. 6); it can be wavelength multiplexed; there is no need for phase preservation in the connector to a structure; it is a very sensitive sensor and is capable

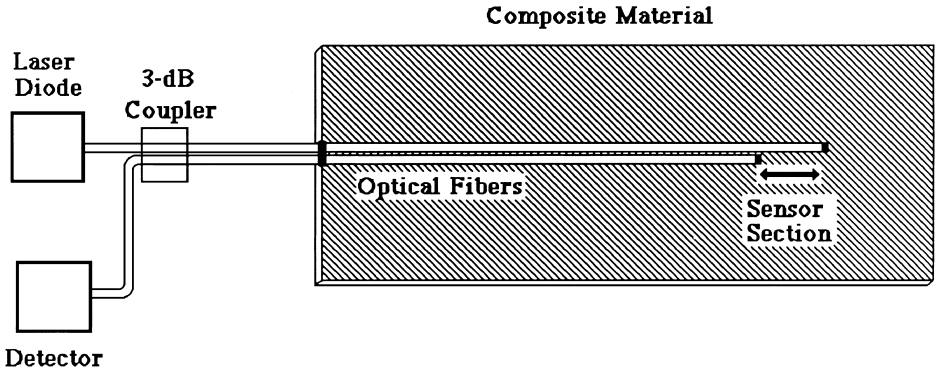


FIGURE 5 Michelson interferometric fiber-optic sensor embedded within a composite panel.

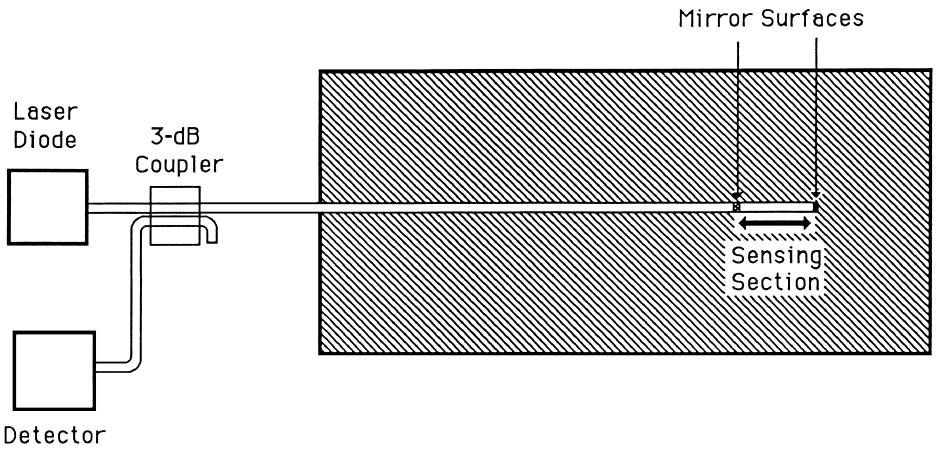


FIGURE 6 Fiber-optic sensor based on a Fabry–Perot interferometer.

of excellent spatial resolution with a well-defined sensing region.

Bragg reflection from a region of periodic variation in the core index of refraction can also form the basis of a fiber-optic sensor. In this system the spectrum of light reflected from the Bragg grating is in the form of a narrow spike with a center wavelength that is linearly dependent on the product of the wavelength of the periodic variation and the mean core index of refraction. Consequently, changes in strain or temperature to which the optical fiber is subjected will shift the reflected spectrum center wavelength and constitute the mode of measurement. Measures *et al.* (1992) have developed a single method of evaluating the wavelength of the Bragg grating sensor.

Modalmetric sensors involve changes in the distribution of light within the optical fiber. These involve single optical fibers in which modal interferometry is exploited in circular or elliptic core, single-mode optical fibers excited with light below the cutoff wavelength (the wavelength above which light propagates in only a single spatial mode). In general, the sensitivity of this type of sensor is far less than the interferometric sensors discussed above, and localization of the sensing region is more difficult. Fiber-optic sensors that rely on evanescent coupling of light can also be included under this heading, since these devices may also be thought to involve a redistribution of modal energy. In a twin-core fiber the evanescent field in the cladding permits the light to be exchanged between the two cores in concert with changes in the strain of temperature of the optical fiber.

Polarimetric sensors rely on evaluating the change in the phase of the two orthogonal polarization eigenmodes of a single-mode optical fiber. Consequently, a measurement of the change in the state of polarization of the light emanating from the optical fiber can be employed to determine the change in the strain or temperature to which the optical fiber is subjected. The polarimetric fiber-optic sensor can also be single ended by employing a mirrored end, and the sensing region can be localized by fusion splicing it to the lead-in/-out optical fiber with its axes rotated through 45° (Fig. 7).

Unfortunately, the nonlinear (cosine) relationship between the signals and the measurand factors, such as strain and temperature, in a structure is common to all of the sensors (see the theoretical section), with the exception of intensiometric and Bragg sensors. This gives rise to serious problems of signal fading and ambiguity in regard to the direction of phase change. There are, however, a number of quadrature detection schemes for overcoming this difficulty.

For sensitive fiber-optic interferometric sensors, the circular symmetry of the optical fiber may not be adequate

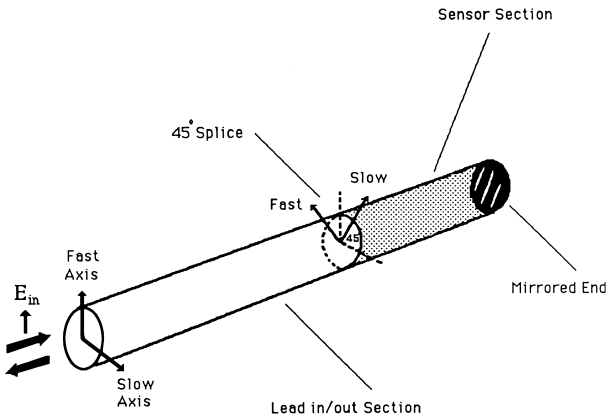


FIGURE 7 Polarimetric fiber-optic sensor in which the sensing section is *localized* between the mirrored end and the 45° rotation of the polarization eigenaxes. Light launched into the lead-in section of the optical fiber is linearly polarized and aligned with one of the fiber's polarization eigenaxes.

to stabilize the two orthogonal linear polarization eigenmodes, and this can lead to *polarization fading* and the possibility of cross coupling between the modes. These effects represent a source of noise in interferometric measurements based on small changes (of the order of a few fringes). One of the most common methods of avoiding this problem is to use polarization-maintaining, single-mode optical fiber. The high birefringence designed into this kind of optical fiber stabilizes the polarization because the two eigenmodes are forced to have an appreciable difference in their respective propagation constants. This limits the exchange of energy between the two modes. This high birefringence is created in one of two ways. The optical fiber is fabricated with a deliberate asymmetry, such as an elliptic core, or by stress loading the core. Several designs have been introduced to this end (Fig. 8).

In order to develop fiber-optic sensors that are well suited for smart structures, we have built and tested several different kinds, including Michelson, polarimetric, and Fabry–Perot sensors. This has allowed us to compare their relative strengths and weaknesses. A set of criteria can serve as a guide for judging any fiber-optic sensor's suitability for use with smart structures. An ideal sensor would be:

1. Intrinsic in nature for minimum perturbation and stability
2. Localized, so it can operate remotely with insensitive leads
3. Able to discern the direction of measurand field change
4. Well behaved with reproducible response
5. Able to determine measurand field directly or with additional measurement
6. All fiber for operational stability
7. Able to provide a linear response—most interferometric sensors require quadrature detection
8. A single optical fiber for minimum perturbation and common mode rejection
9. Single ended for ease of installation and connection
10. Adequate strain resolution with sufficient measurement range
11. Insensitive to phase interruption at the structural interface
12. Nonperturbation to the structure and robust for installation
13. Interrupt immune and capable of absolute measurement
14. Amenable to multiplexing to form sensing networks within structures
15. Easily manufactured and adaptable to mass production

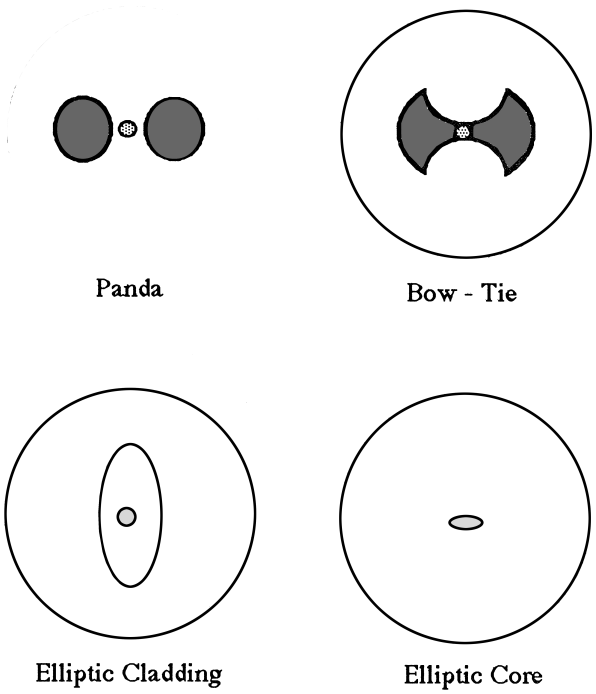


FIGURE 8 Four types of polarization-maintaining optical fibers.

If a single sensor could be developed for all of the measurements required by smart structures, the cost would be minimized. Such a sensor would have to be sufficiently sensitive to detect acoustic emission, while exhibiting a measurement range adequate for monitoring the general loading of the structure. This requires an interferometric fiber-optic sensor, and of the various types the choice seems to lie between the "in-fiber" intrinsic Fabry–Perot and the in-fiber Bragg grating sensor. A careful comparison will have to be undertaken to determine which has the greatest advantage in practice. The Fabry–Perot may have the edge when it comes to sensitivity, but the Bragg offers absolute measurement capability with simple demodulation.

C. System Considerations

1. Configuration and Structural Considerations

Designing a smart structure will require questions of *configuration* to be addressed. For example, should the fiber-optic sensors be embedded within the structure or bonded to its surface; can the fiber-optic sensors be double ended or are they required to be single ended (i.e., the signal exits through the same optical fiber as the exciting light entered the structure); are the fiber-optic sensors restricted to a single plane (which would be the case if surface bonded, since optical fibers should not cross each other), or can an orthogonal grid be incorporated in the design; should

the fiber-optic sensors make point, line, or distributed measurements.

Other important system design considerations would include the density of the optical fiber sensors, their arrangement, and, if embedded, their depth and orientation with respect to the reinforcing fibers. Extensive studies at the University of Toronto Institute for Aerospace Studies on damage assessment for composite materials using embedded optical fibers have provided answers to some of these questions. This work demonstrated that where fracture of an embedded optical fiber is used to assess damage the optimum orientation is perpendicular to the adjacent reinforcing fibers. Conversely, if the optical fibers are required to have the greatest chance of surviving in the presence of damage (as would be the case if they were serving as data conduits or sensors), they should be oriented parallel to the adjacent reinforcing fibers (see Fig. 9) and positioned as far from the surface of maximum tensile strain as possible.

A list of some of the fundamentals of composite damage assessment based on the fracture of embedded optical fibers is presented in Table II. This includes identification of through-the-thickness matrix cracks as the primary mechanisms for fracture of optical fibers embedded within composite materials. In the case of damage assessment systems based on fracture of an embedded grid of optical fibers, three basic configurations are possible. These are illustrated in Fig. 10.

Other factors that need to be addressed when designing fiber-optic sensors to be embedded within a composite structure include the material of the optical fibers in terms of its temperature survivability during the cure and compatibility with the material of the structure; the size of the optical fibers in order to minimize their perturbation; and the type of coating needed to protect the optical fibers and

TABLE II Fundamentals of Composite Damage Assessment Based on the Fracture of Embedded Optical Fibers

- *Fracture mechanism* for embedded optical fibers is through-the-thickness *matrix cracks*.
- *Damage sensitivity* of embedded optical fibers can be *tailored* by an intermittent etching process.
- *Maximum damage sensitivity* is attained when the optical fibers are embedded close to the surface of maximum tensile load.
- *Threshold* impact energy for fracture of embedded optical fibers depends on composite layup and orientation and placement of the optical fibers.
- Damage sensitivity and reliability are *optimized* when the optical fibers are embedded between *collinear* plies and oriented *orthogonal* to ply direction.
- Optical fibers embedded between *noncollinear* plies should be oriented *orthogonal to the ply* separating them from the region of damage.
- If this aligns the optical fibers with the other ply, then the orientation should bisect the larger angle between the plies.

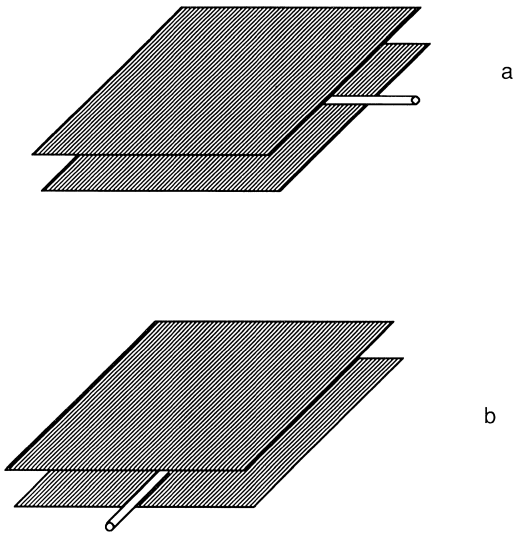


FIGURE 9 Optimum orientation for optical fibers embedded between collinear plies within composite materials for the purpose of (a) evaluating damage based on fracture of the optical fibers and (b) serving as data links or sensors.

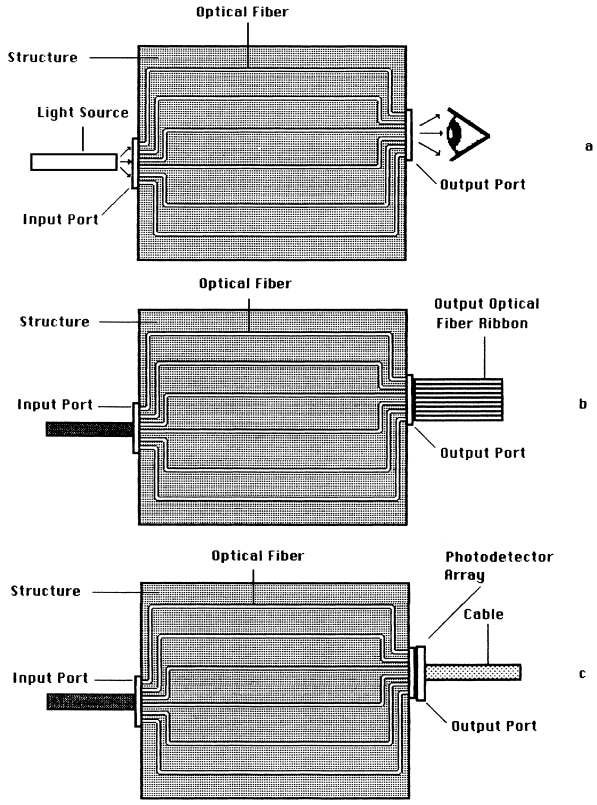


FIGURE 10 Three possible optical interface arrangements for a transmissive-based, optical fiber sensor system.

to provide the appropriate coupling to the structure for a given application.

2. Influence of Embedded Optical Fibers on Material Properties

A fundamental issue that must be addressed if optical fibers are to be embedded within practical composite material structures concerns their potential to compromise the tensile or compressive strength, increase the damage vulnerability of these materials by reducing their interlaminar fracture toughness, or reduce their fatigue life.

Some specific micromechanic questions related to the presence of imbedded optical fibers are:

- Do they initiate matrix cracks or delaminations?
- Do they stimulate the growth of an established (through impact, fabrication flaw, or fatigue) region of delamination?

Some work has already been undertaken in an attempt to address these concerns, and preliminary results suggest that no noticeable deterioration of the mechanical properties occurs for small optical fibers. In particular, it has been shown that neither the tensile nor the compressive strengths of Kevlar/epoxy panels are compromised by the presence of embedded optical fibers of 125 μm diameter (see Fig. 11). Blagojevic *et al.* (1989) have also demonstrated that the interlamina fracture toughness of the composite material is not reduced by the presence of embedded optical fibers; if anything, they marginally improve the damage resistance of the material (see Table III). A study of the influence of embedded optical fibers on graphite bismaleide specimens has been undertaken. Unfortunately, the conclusions are limited, as the buffer on the optical fibers was not removed and at the high temperature of the cure it probably melted, possibly creating voids. A recent extensive series of experiments by Loken (1990) has shown that embedded optical fibers with a diameter

TABLE III Critical Energy Release Rate, G_c (100 mm) for Kevlar/Epoxy Panel Samples with and without Embedded Optical Fibers

Composite sample configuration	Critical energy release rate, G_c
Reference sample with no optical fibers	
$[O_8/O_8]$	1.09 (kJ m^{-2})
$[O_8\{0\}O_8]$	1.12 (kJ m^{-2})
$[O_8\{45\}O_8]$	1.14 (kJ m^{-2})
$[O_8\{90\}O_8]$	1.24 (kJ m^{-2})
Reference sample with no optical fibers	
$[O_2/90/O_5/90/O_5/90/O_2]$	1.43 (kJ m^{-2})
$[O_2/90/O_5\{0\}90/O_5/90/O_2]$	1.44 (kJ m^{-2})
$[O_2/90/O_5\{45\}90/O_5/90/O_2]$	1.46 (kJ m^{-2})
$[O_2/90/O_5\{90\}90/O_5/90/O_2]$	1.48 (kJ m^{-2})

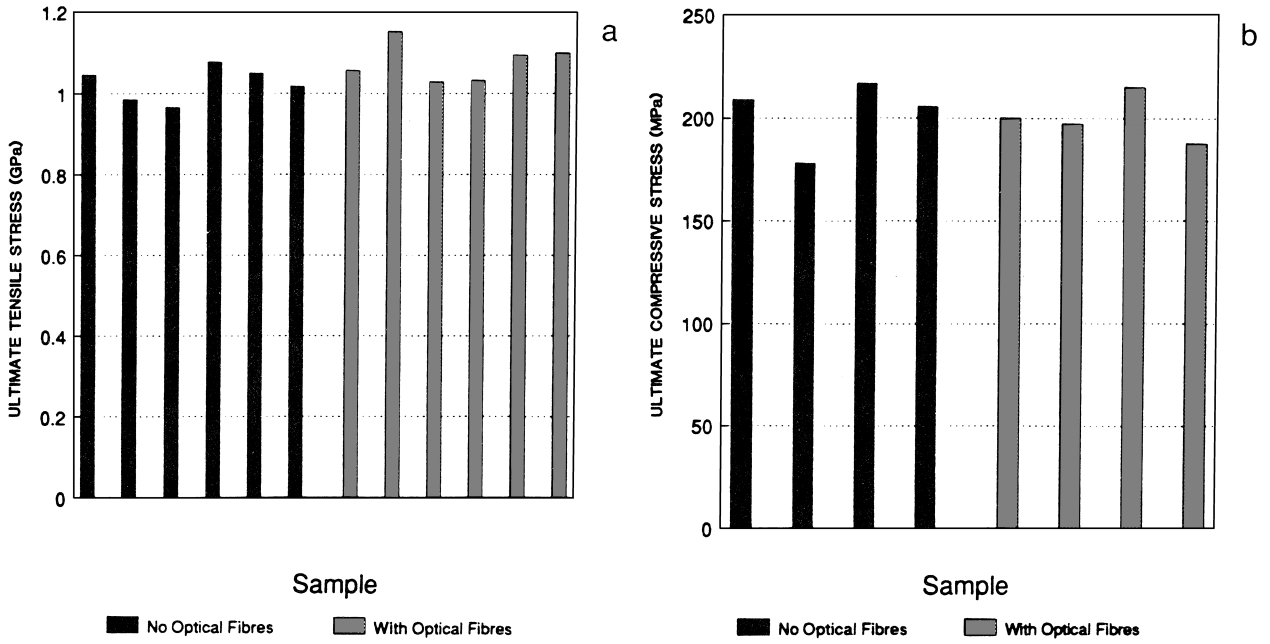


FIGURE 11 (a) Ultimate tensile strength of instrumented [0,0{90}0,0] and uninstrumented [0_4] 4-ply Kevlar/epoxy laminates. (b) Ultimate compressive strength of instrumented [$0_4\{90\}0_4$] and [0_8] 8-ply Kevlar/epoxy laminates.

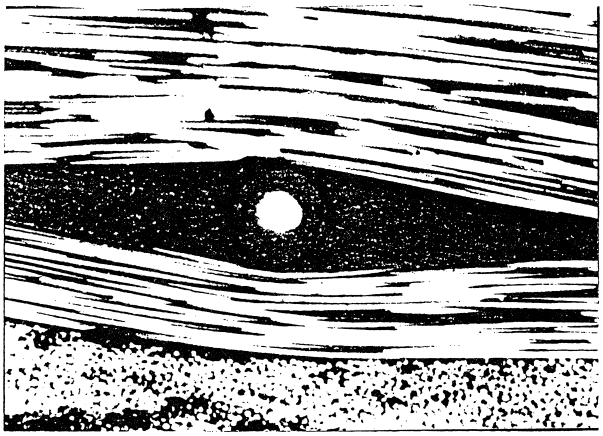


FIGURE 12 Resin-eye associated with an 125-μm diameter optical fiber embedded orthogonal to adjacent graphite/epoxy plies.

of less than 140 μm do not appear to reduce the ultimate tensile strength or the fatigue life of the graphite/epoxy specimens.

Fractographic studies reveal that optical fibers embedded at an angle to the adjacent ply directions of the composite material create resin cavities (termed "resin-eyes" because of their shape) (Fig. 12). The aspect ratio of such an eye is larger the greater the angle and is a maximum when the optical fiber is embedded orthogonal to the adjacent plies. Optical fibers embedded parallel to reinforcing fibers do not create resin-eyes and therefore minimally perturb the material. Dasgupta *et al.* (1990) have recently modeled the formation of resin-eyes and have predicated stress concentration factors in excess of 10 in the vicinity of the embedded optical fibers, a result suggested earlier by Tay, Wilson, and Wood (1989) and demonstrated by Czarnek *et al.* (1988) using Móire interferometry. Sirkis and Dasgupta (1990) have suggested that these high stress concentrations may be diminished with coatings of appropriate size and stiffness. Clearly, careful consideration will have to be given to the diameter of optical fibers and their type of coating if they are to be embedded within composite structures and function correctly for the useful life of the structure. The presence of optical fibers within laminated composite structures could increase the complexity of fabrication, design, and analysis. Since the influence optical fibers have on their host composites is not fully understood, more definitive research will be needed before they can be imbedded with confidence within composite structures intended to have a 20 (plus)-year working life.

3. Demultiplexing/Multiplexing

Multiplexing is the merging of data from a number of different channels into one channel, while demultiplexing is the inverse (Fig. 13). Both are expected to play an important role in the development of smart structures. The primary parameters used in multiplexing schemes are wavelengths, time, frequency, phase, and space. Wavelength division multiplexing (WDM) involves each sensor operating at a different wavelength and looks the most promising for certain types of fiber-optic sensors to be used in smart structures. The relevant wavelength encoding can be performed by the light source, spectral filters, resonant cavities, or dispersive elements. Demultiplexing is achieved by means of a spectral analyzer.

If the sensors have to be single ended, then, as shown in Fig. 14a, couplers would be required to divert some fraction of the returning light signal to the photodetection system. It has been assumed that the interface to the structure would include a device for optical demultiplexing/multiplexing so that only one (armored) optical fiber is needed to convey the interrogating light in and the signals out of the structure. If this is not the case and each optical fiber sensor has its own input fiber, it might also require its own coupler, light source, and detector, and the system would be bulky, heavy, expensive, and less practical.

In the case of a double-ended system, there are three possible detection modes. The output light from the network of optical fiber sensors could be communicated to an external photodetection system by means of an optical ribbon or fiber-optic bundle; if spatial coherency is not maintained, it would be necessary to identify the detector viewing each optical fiber sensor. An alternative arrangement would involve mounting the photodetection system directly on the output port; this may be practical in many situations because of the small size of photodiode arrays. The third possibility would rely on a small multiplexing output device to recombine the optical signals and send them all down one armored optical fiber to the photodetection system (see Fig. 14b). We have assumed in each of these double-ended arrangements that the interrogating light is delivered to the structure through one armored optical fiber and is then distributed to each of the fiber-optic sensors by means of an optical demultiplexer. For many situations it may be just as convenient to use a fiber-optic bundle, or an optical ribbon, to deliver the light to each sensor on a one-to-one basis.

In terms of system components, the factors that would determine the choice of light source include wavelength, brightness, size, whether pigtailed to an optical fiber, and cost. The first decision in regard to the choice of photodetector would hinge on whether a single detector or an array of detectors is needed; then the questions of wavelength response, sensitivity, and size would have to be addressed. If couplers are required, it would then be necessary to decide between 3-dB (50:50) couplers or asymmetric couplers

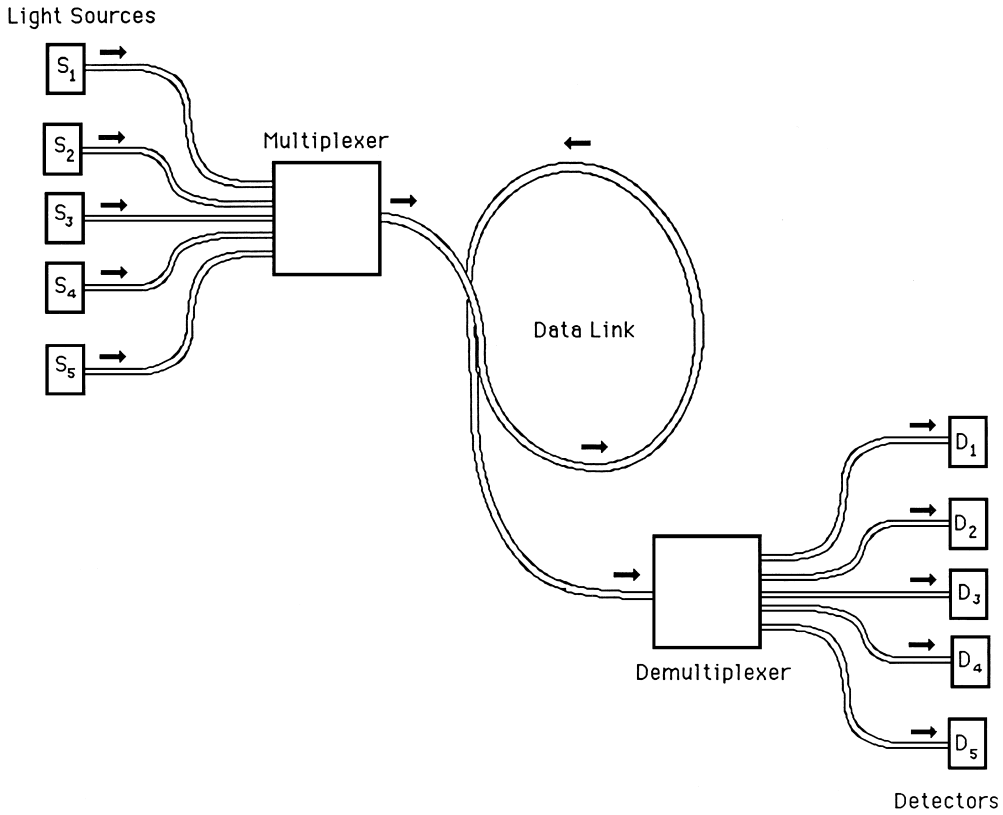


FIGURE 13 Illustration of "multiplexing" (MUX) the light from several channels into one channel and "demultiplexing" (DEMUX) the light from one channel into several channels.

and to consider if they need to be polarization-splitting couplers.

IV. SINGLE-MODE OPTICAL FIBER SENSOR THEORY

Strain, deformation, and structural loading can be evaluated by phase or polarization transduction. Figure 15 represents a generic system that can be used to discuss either interferometric or polarimetric fiber-optic sensors. In principle, the *signal*, as influenced by the measurand field through the transduction mechanism, is mixed with a *reference* wave and the resulting interference signal is detected. If the signal and reference wave have the same optical frequency, we refer to "homodyne" detection; if their frequencies are different, we speak of "heterodyne" detection. We assume no attenuation.

A. Interferometric Response Function

In both a Michelson and a Mach–Zehnder interferometer, the signal and reference paths are spatially separate, and we assume that none of the components are birefringent. In the case of polarimetry, the two waves are orthogonal polarization states that are spatially equivalent. The couplers in Fig. 15 are described by matrices K_1 and K_2, while the two paths (a) and (b) are together described by the transduction matrix T. In general, the detected (output) field $\{E_7, E_8\}$ is related to the incident (input) field $\{E_1, E_2\}$ by the relation

$$E_{\text{out}} = K_2 T K_1 E_{\text{in}}. \tag{1}$$

If we assume 3-dB couplers so that the power is shared equally between the two outputs, then

$$K_1 = K_2 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix}, \tag{2}$$

and the transduction matrix can be expressed in the form

$$T = \begin{bmatrix} \exp(i\phi_{\text{a}}) & 0 \\ 0 & \exp(i\phi_{\text{b}}) \end{bmatrix}, \tag{3}$$

where ϕ_{a} and ϕ_{b} represent the phase change experienced by the waves propagating along paths (a) and (b), respectively.

Substitution of Eqs. (2) and (3) into Eq. (1) yields

$$E_7 = E_1\{\exp(i\phi_{\text{a}}) - \exp(i\phi_{\text{b}})\}/2 \tag{4}$$

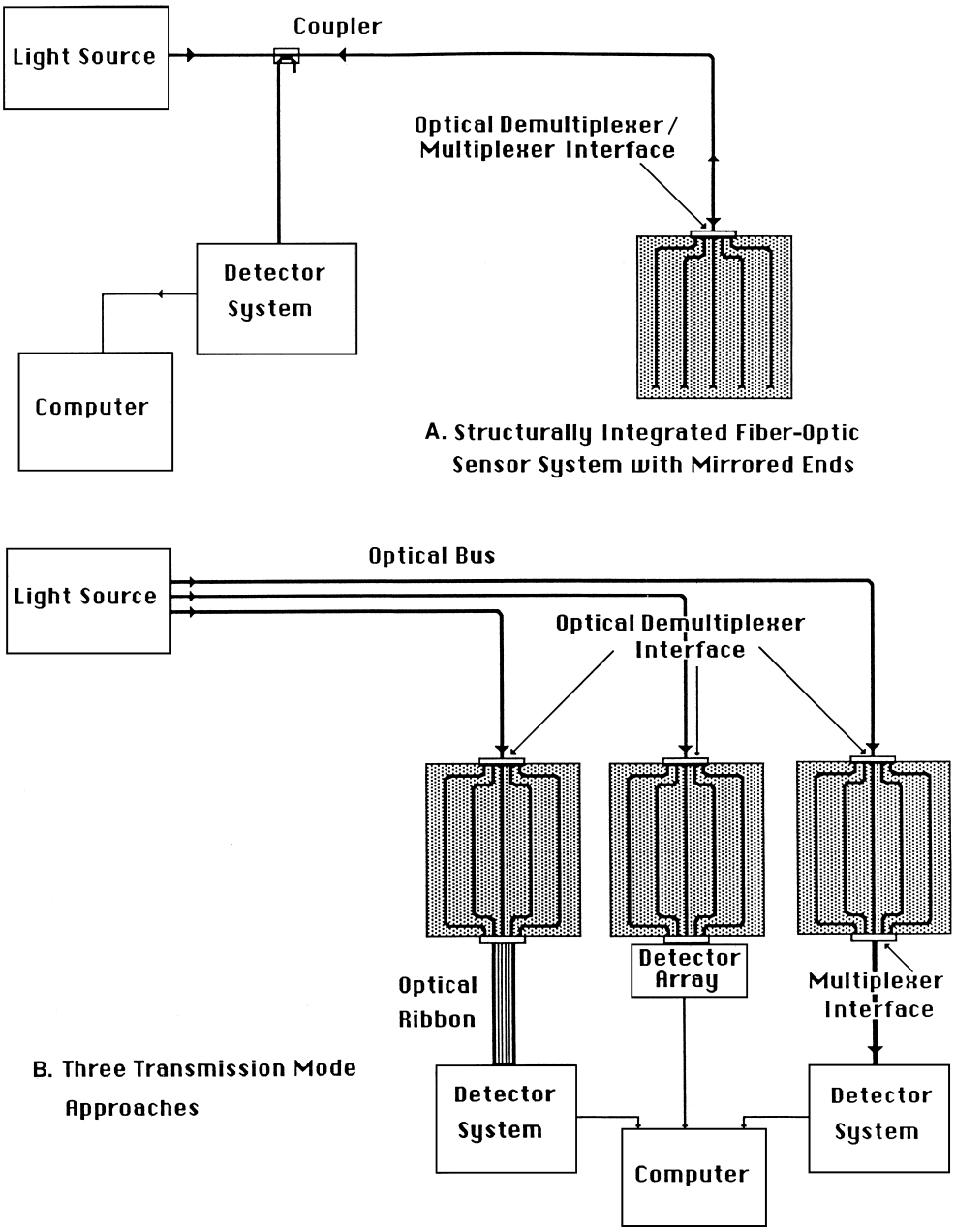


FIGURE 14 (a) Illustrates how the combination of a coupler and a DEMUX/MUX interface would permit an array of structurally integrated, single-ended, fiber-optic sensors to be interrogated by a single input/output optical fiber. (B) Illustrates three detection modes possible with double-ended, transmission-based, structurally integrated fiber-optic sensor arrays. Excitation of the array of optical fibers in each structure is assumed to be accomplished by only one input optical fiber and an appropriate demultiplexer interface.

and

$$E_8 = iE_1\{\exp(i\phi_a) + \exp(i\phi_b)\}/2 \tag{5}$$

assuming $E_2 = 0$. The corresponding photodetector signals for output ports 7 and 8 are

$$I_7 = I_1[1 - \cos(\phi_a - \phi_b)]/2 \tag{6}$$

and

$$I_8 = I_1[1 + \cos(\phi_a - \phi_b)]/2. \tag{7}$$

In general, we can write

$$\phi = \phi_a - \phi_b. \tag{8}$$

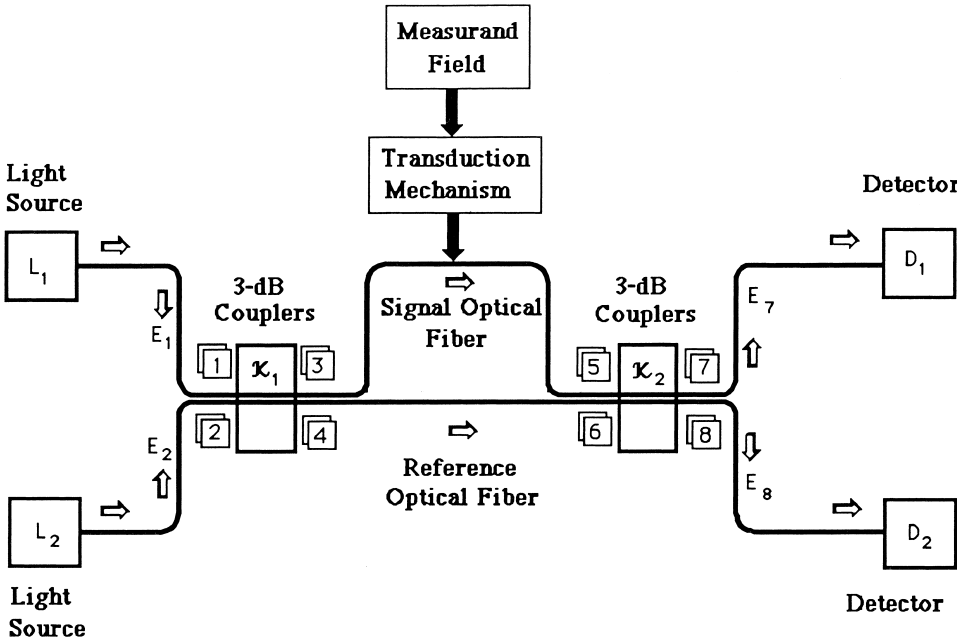


FIGURE 15 Generic theoretical model of a single-mode optical fiber sensor.

In the case of a single optical fiber sensor, ϕ represents the phase retardation between the reference and sensing fields propagating along the optical fiber for a path difference L. Under these conditions, we may write

$$\phi = \beta L. \tag{9}$$

In general, ϕ depends upon the length, temperature, and operating wavelength of a sensor; i.e., $\phi = \phi(L, T, \lambda)$, where T is taken as the temperature change from some reference temperature. The variation in ϕ due to incremental changes in these parameters is given by the expression

$$\Delta\phi = \left(L\frac{\partial\beta}{\partial L} + \beta\right)\Delta L + \left(L\frac{\partial\beta}{\partial T} + \beta\frac{\partial L}{\partial T}\right)\Delta T + L\frac{\partial\beta}{\partial\lambda}\Delta\lambda. \tag{10}$$

For the simplest case of strain sensing when the temperature is kept constant (and equal to the reference value) and the wavelength is fixed, then

$$\Delta\phi = \left(L\frac{\partial\beta}{\partial L} + \beta\right)\Delta L, \tag{11}$$

or if we introduce the *path integrated strain*

$$\varepsilon \equiv \Delta L/L, \tag{12}$$

then

$$\Delta\phi = \beta L\varepsilon + L\frac{\partial\beta}{\partial L}\Delta L. \tag{13}$$

It might be noted that variations in the wave-length can lead to methods of signal recovery. For example, alternating laser pulses with a wavelength difference corresponding to a multiple of $\pi/2$ can lead to quadature signals.

B. Strain-Optic Theory

For an *interferometric sensor*, $\beta = kn$, where k is the free-space propagation constant and n is the effective index of refraction of the guided mode. Thus,

$$\Delta\phi = knL\varepsilon + kL\frac{\partial n}{\partial L}\Delta L. \tag{14}$$

However,

$$\Delta n = \frac{\partial n}{\partial L}\Delta L, \tag{15}$$

and since

$$\Delta\left(\frac{1}{n^2}\right) = -2\frac{\Delta n}{n^3}, \tag{16}$$

we can rewrite Eq. (14) in the form

$$\Delta\phi = knL\varepsilon - \frac{kn^3L}{2}\Delta\left(\frac{1}{n^2}\right). \tag{17}$$

From strain-optic theory, changes in the optical indicatrix $(1/n^2)_\alpha$ due to strain applied to the material is described by the relation

$$\Delta\left(\frac{1}{n^2}\right)_\alpha = \sum_{\gamma=1}^{6} P_{\alpha\gamma}\varepsilon_\gamma, \tag{18}$$

where ε_γ is the *strain tensor* and $P_{\alpha\gamma}$ are the strain-optic coefficients of the medium. If a uniaxial longitudinal stress σ_z is applied to an isotropic, elastic optical fiber oriented in the z-direction, the resulting strain from first order elastic theory will be

$$\varepsilon_\gamma = \begin{bmatrix} -\nu\varepsilon_z \\ -\nu\varepsilon_z \\ \varepsilon_z \end{bmatrix} \tag{19}$$

where we have assumed zero shear strain.

$\varepsilon_z = \sigma_z/E$ is the longitudinal strain, while $-\nu\varepsilon_z$ is the corresponding transverse strain, and E is Young's Modulus (70 GPa for fused silica). Poisson's ratio, ν, is typically 0.17 for fused silica.

Since the optical fiber is assumed to be homogeneous and isotropic:

$$P_{\alpha\gamma} = \begin{bmatrix} P_{11} & P_{12} & P_{12} & 0 & 0 & 0 \\ P_{12} & P_{11} & P_{12} & 0 & 0 & 0 \\ P_{12} & P_{12} & P_{11} & 0 & 0 & 0 \\ 0 & 0 & 0 & P_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & P_{44} & 0 \\ 0 & 0 & 0 & 0 & 0 & P_{44} \end{bmatrix}, \tag{20}$$

with $P_{44} = (P_{11} - P_{12})/2$, as only the P_{11} and P_{12} strain-optic coefficients are nonzero, in which case using Eqs. (16)–(18) we have for an interferometric sensor

$$\Delta n = \frac{n^3\varepsilon_z}{2}[(P_{11} + P_{12})\nu - P_{12}] \tag{21}$$

and

$$\Delta\phi = knL\varepsilon_z\left[1 - \frac{n^2}{2}\{P_{12} - \nu(P_{11} + P_{12})\}\right] \tag{22}$$

which can be expressed in terms of the interferometric phase-strain sensitivity:

$$S \equiv kn\left[1 - \frac{n^2}{2}\{P_{12} - \nu(P_{11} + P_{12})\}\right]. \tag{23}$$

Thus, we can write

$$\Delta\phi = SL\varepsilon_z, \tag{24}$$

where L is the length of the sensing region (in essence the separation of the mirrors). Recent values of these strain-optic coefficients, P_{11} and P_{12}, have been reported by Bertholds and Dandliker (1988) for single-mode optical fibers. They indicate that for fibers with a pure silica core and B_2O_3-doped cladding, $P_{11} = 0.113$ and $P_{12} = 0.252$, values that are about 7% lower than for bulk silica. In the case of silica optical fibers (with $n = 1.458$),

$$S = 11.3 \text{ rad } \mu \text{ strain}^{-1} \text{ m}^{-1}$$

(i.e., $S = 1.13 \times 10^7$ rad strain^{-1} m^{-1}).

Recently, Sirkis and Haslach (1990) considerably extended the Butter and Hocker (1978) phase-strain model and have shown that their results are closer to those observed in transverse loading experiments (Mathews and Sirkis, 1990). According to Sirkis and Haslach (1990), the change in the phase retardation experienced by light propagating along an optical fiber subject to a strain field is given by the relation

$$\Delta\phi = nk\int_{s_1}^{s_2}\left[\varepsilon_1 - \frac{n^2}{4}\{P_{2j} + P_{3j}\}\varepsilon_j\right]ds, \tag{25}$$

where ε_1 represents the strain component taken tangential to the fiber and ε_j represents the strain components of the strain tensor. Mathews and Sirkis (1990) have shown that large deviations from the Butter and Hocker free-fiber uniaxial-stress model are only expected *when the stiffness of the optical fiber is less than ten times that of the host matrix,* i.e., $E_f/E_h < 10$. The differences between the theories is, however, expected to arise primarily for conditions of transverse loading. For the simplified case of a "small" sensing region, the Sirkis and Haslach model leads to a phase change for one of the eigen polarization states:

$$\Delta\phi = nkL\left[\varepsilon_1 - \frac{n^2}{2}\{P_{11}\varepsilon_2 + P_{12}(\varepsilon_1 + \varepsilon_3)\}\right]. \tag{26}$$

We see that if there is substantial transverse loading the phase change suffered can no longer uniquely determine the strain state as it depends on each of the three strain components. Clearly, under these circumstances it is essential to measure each of the strain components.

When optical fibers are embedded within composite materials, due consideration must be given to the effects of the cure, or consolidation, process upon their coating. For example, acrylic coatings have been observed to melt during the cure process in an autoclave, and this can lead to poor bonding, asymmetrical loading, and the creation of large resin-rich perturbations of the composite around the optical fiber. Indeed, even the much larger diameter of the optical fiber with its coating will, in general, represent a significant perturbation. Improved coatings that are thin; capable of surviving the cure, or consolidation, process; and able to provide excellent bonding remain to be developed.

C. Thermal Sensitivity and Apparent Strain

Apparent strain is another issue that has to be addressed when considering the use of optical fiber sensors for practical smart structures, especially in the aerospace field, where they are likely to be subjected to considerable temperature excursions. Strain measurements made with optical fibers adhered to or embedded within such structures would thus be complicated by the appearance of *apparent strain* due to the appreciable thermal sensitivity of the index of refraction and thermal expansion of the optical fiber, and the host.

In the case of the interferometric sensors, the incremental change in the phase associated with an incremental change in both the local stress, $\Delta\sigma$, and the temperature, ΔT, is given by

$$\Delta\phi = \left[\frac{\partial\phi}{\partial\sigma}\right]_T \Delta\sigma + \left[\frac{\partial\phi}{\partial T}\right]_\sigma \Delta T. \tag{27}$$

This can be expanded in the form

$$\Delta\phi = nkL\left[\left\{\left[\frac{\partial\varepsilon}{\partial\sigma}\right]_T + \frac{1}{n}\left[\frac{\partial n}{\partial\varepsilon}\right]_T\left[\frac{\partial\varepsilon}{\partial\sigma}\right]_T\right\}\Delta\sigma + \left\{\frac{1}{n}\left[\frac{\partial n}{\partial T}\right]_\sigma + \frac{1}{L}\left[\frac{\partial L}{\partial T}\right]_\sigma\right\}\Delta T\right], \tag{28}$$

which can be expressed in terms of *Young's Modulus*, E, and the *coefficient of thermal expansion* for the optical fiber, α_F:

$$\Delta\phi = nkL\left[\left\{1 + \frac{1}{n}\left[\frac{\partial n}{\partial\varepsilon}\right]_T\right\}\frac{\Delta\sigma}{E} + \left\{\frac{1}{n}\left[\frac{\partial n}{\partial T}\right]_\sigma + \alpha_F\right\}\Delta T\right], \tag{29}$$

where the second term in Eq. (29) corresponds to the *strain-optic effect* and the third term in Eq. (29) accounts for the *thermooptic effect*.

If we introduce the *bonded fiber phase-strain coefficient*,

$$g = nkL\left\{1 + \frac{1}{n}\frac{\partial n}{\partial\varepsilon}\right\}_T. \tag{30}$$

If we introduce the *free fiber thermal phase coefficient*,

$$f = nkL\left\{\frac{1}{n}\frac{\partial n}{\partial T} + \alpha_F\right\}_\sigma. \tag{31}$$

In the case of an optical fiber bonded to a host material, under conditions of no applied load other than the incremental stress arising from the difference in the coefficient of thermal expansion between the host material and the optical fiber, we can write:

$$\Delta\sigma = E(\alpha_H - \alpha_F)\Delta T, \tag{32}$$

where α_H is the *coefficient of thermal expansion* for the host material. Under these circumstances the incremental change in the phase associated with an increase of temperature, ΔT, can be expressed in the form

$$\Delta\phi = g\{\alpha_H - \alpha_F\}\Delta T + f\Delta T. \tag{33}$$

If we rewrite Eq. (33) in terms of an *apparent strain*, ε_{app}, then we have

$$\Delta\phi = g\varepsilon_{app}, \tag{34}$$

and we see that the apparent strain is given by

$$\varepsilon_{app} = \left[\frac{f}{g} + \alpha_H - \alpha_F\right]\Delta T. \tag{35}$$

Some researchers have suggested that this problem could be overcome by simply measuring the temperature at the same time as the strain is monitored. This may not be true for sensors embedded within composite materials, as their anisotropic properties cause the apparent strain to depend quite strongly on the orientation of the optical fiber relative to the reinforcing fibers. In this case point 5 in the criteria for sensors may require the development of special thermally compensated optical fibers. Conventional strain gauges expected to work over large temperature ranges also have to be tailored to the material to which they are adhered.

It should be noted that apparent strain needs to always be taken into account and is not restricted to interferometric sensors. Indeed, a comparison between polarimetric and interferometric fiber-optic sensors has found that the former has an "apparent strain sensitivity" of 20 μ strain/°C compared to 7 μ strain/°C for the Michelson fiber-optic sensor.

Farahi *et al.* (1990) have also considered strain and temperature cross sensitivity terms and have shown that these terms may need to be considered for large strain or temperature excursions.

D. Polarimetric Fiber-Optic Sensors

In polarimetry the waves travel the same physical path, but the two orthogonal polarization states travel with different velocities due to the birefringent properties of the optical fiber. Under these circumstances,

$$\Delta\phi = kLB, \tag{36}$$

where k is the free-space propagation constant, L is the length of the optical fiber, and

$$B = [n_f - n_s] \tag{37}$$

is the birefringence of the optical fiber, with n_f and n_s representing the index of refraction for the fast and slow

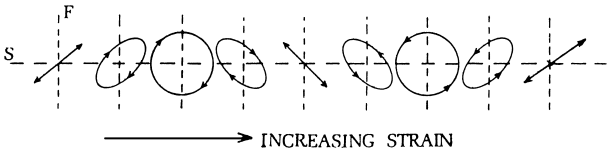


FIGURE 16 Variation in the state of polarization of light emerging from a birefringent medium subject to increasing strain.

eigenmodes, respectively. The beat length, L_p, of an optical fiber is given by the simple relation

$$L_p = 2\pi/kB, \tag{38}$$

and we see that

$$\Delta\phi = 2\pi L/L_p, \tag{39}$$

which means that the beat length corresponds to the length of optical fiber needed for the two orthogonal polarization modes to change their relative phase by 2π.

Polarimetric-based fiber-optic sensors can be viewed as a form of differential interferometer in which radiation in the two polarization eigenmodes propagate at different speeds within a single-mode optical fiber. The state of polarization (SOP) consequently evolves along the optical fiber in the manner indicated in Fig. 16. Optimized results are obtained if the optical fiber is highly birefringent and both polarization eigenmodes are equally excited. This latter condition can be achieved if the light launched into the optical fiber is either circularly polarized or linearly polarized at 45° to the polarization eigenaxis of the fiber. In general, the sensitivity of these sensors is significantly (about 100 times) less than that of interferometric fiber-optic sensors.

The usual method of localizing the sensing region of a polarimetric-based fiber-optic sensor is to orient the polarization eigenaxes of this section at 45° to that of the lead-in and lead-out sections of the optical fiber (Fig. 7). It is also important to ensure that the light launched into the lead-in optical fiber is linearly polarized and aligned with one of the polarization eigenaxes. This is most easily accomplished in an all-fiber system by using polarizing optical fiber in the lead-in and lead-out sections. An alternative, and more readily available, method of launching light correctly into the optical fiber would use a small, thin film polarizer in the connector to the structure. The transmission axis of this polarizer would have to be accurately aligned with one of the eigenaxes of the polarization-maintaining lead-in optical fiber. This type of sensor can be operated in either a transmission or a reflection mode. The latter approach, in which the sensing section is terminated by a mirrored end, is more attractive from the smart structures perspective in that it constitutes a single-ended device is twice as sensitive for the same sensing length and requires only one 45° splice. It has been shown that it is possible to achieve a 45° rotation by means of local heating and twisting of the optical fiber. This eliminates the need for the fusion splices.

The state of polarization of linearly polarized light launched into one of the polarization eigenmodes of the lead-in optical fiber is unchanged by environmental effects as it propagates along this section of optical fiber. On arriving at the 45° splice, it will equally populate the two polarization eigenstates of the sensing section of the optical fiber. After traversing the sensing region (once for the transmission-based system or twice for the case of the mirrored end), these orthogonally polarized beams are incident upon the lead-out optical fiber, which in effect samples the new state of polarization, determined by the birefringence of the sensing fiber. Clearly, as the birefringence of the sensing region responds to the changes in the environment (e.g., strain or temperature), this is reflected in the change in the state of polarization of the light leaving the sensing region. A photodectector viewing the intensity of light in one of the polarization eigenmodes through the use of a polarizing element (Fig. 7) of the lead-out optical fiber will respond to this variation in the state of polarization of the light the exits the sensing region. The output intensity for the polarization eigenmodes can, in general, be expressed in the form

$$I_1 = I_o[1 + \cos(\phi_f - \phi_s)] \tag{40}$$

and

$$I_2 = I_o[1 - \cos(\phi_f - \phi_s)], \tag{41}$$

where $\phi_f = kn_fL$ and $\phi_s = kn_sL$, with k again as the free-space propagation constant, and L as the sensing length of the optical fiber.

E. Fiber-Optic Sensor Signal Recovery

We see that for each of the sensors the outputs are cosine in nature, which means that they are multivalued and clearly nonlinear. Furthermore, the signals cannot be used to determine the direction of change, i.e., $\cos(\Delta\phi) = \cos(-\Delta\phi)$, and the response becomes very insensitive whenever $\Delta\phi = n\pi$, $\{n = 0, 1, 2, \ldots\}$. This latter condition is termed "*signal fading*."

These problems can be avoided if we have two quadrature outputs, i.e., two signals that differ by $\pi/2$. These can be expressed in the form

$$X = A\cos\{\Delta\phi\} \tag{42}$$

and

$$Y = B\sin\{\Delta\phi\}, \tag{43}$$

for then one would always be at the point of maximum sensitivity when the other was at the point of minimum sensitivity. In fact, if the following operation is performed:

$$\int \{X\, dY/dt - Y\, dX/dt\}\, dt = -AB\,\Delta\phi, \qquad (44)$$

we see that a signal directly proportional to $\Delta\phi$ can be obtained. An overview of signal processing for interferometric fiber-optic sensors has been provided by Liu and Measures (1992). Unfortunately, we see from Eqs. (40) and (41) that the outputs from the interferometer are antiphase (i.e., different by π). If the typical measurements involve many cycles (or fringes), the signal fading is not an issue; however, the *ambiguity* in regard to the direction of change remains a problem.

An alternative approach mixes the (non-dc) quadrature outputs with a mutually orthogonal local oscillator for signals $[\sin(\omega t), \cos(\omega t)]$. The sum of the resultant mixed signals,

$$\sin(\omega t)\cos\Delta\phi + \cos(\omega t)\sin\Delta\phi = \sin(\omega t + \Delta\phi), \quad (45)$$

is then compared with the local oscillator $\sin(\omega t)$ in a "phase sensitive detector" and the output is "low pass filtered" to extract the phase retardation signal, $\Delta\phi$.

In the presence of attenuation, the signals can be expressed in the form

$$I_7 = I_1[1 - U\cos(\Delta\phi)]/2 \qquad (46)$$

and

$$I_8 = I_1[1 + U\cos(\Delta\phi)]/2, \qquad (47)$$

where $0 < U < 1$ is termed the visibility constant.

Two techniques that can be used to produce quadrature signals and are suitable for smart structures involve

1. Replacing the 4 port (2×2) output direction coupler of Fig. 15 with a (4×4) direction coupler; (3×3) direction couplers may also be used.
2. Switching the absolute frequency of the laser (in a time that is short compared to the times of interest) between two frequencies f_1 and f_2 which correspond to an effective phase difference for the unbalanced interferometer of $\pi/2$, i.e.,

$$2\pi nL\{f_1 - f_2\}/c = \pi/2. \qquad (48)$$

The problems of signal fading and fringe ambiguity for polarimetric fiber-optic sensors can be avoided by also operating at the quadrature point. Two methods of producing quadrature signals with such sensors that are also suitable for use with smart structures are:

1. Using a 3×3 coupler and necessary electrons for the required signal processing
2. Switching between two lasers in order to shift the wavelength by an amount that corresponds to an effective phase difference for the two polarization eigenmodes of $\pi/2$

It should be noted that the wavelength shift corresponding to a $\pi/2$ phase difference in the case of polarimetry is much larger than normally required for interferometry. For that reason, it may be necessary to switch between two lasers rather than shifting the output wavelength of one laser (by switching the injection current) (Fig. 17).

V. STRUCTURAL MEASUREMENTS POSSIBLE WITH FIBER-OPTIC SENSORS

The potential breadth of applications that can be undertaken with optical fiber sensors embedded within composites can be gauged by reference to Table IV. These can be divided into four groups: fabrication control; structural integrity; loading, shape, and vibration; and thermal state. Structural integrity monitoring (or damage assessment) will be considered in detail in the next section. Here, we shall review the more general kinds of structural measurements possible with fiber-optic sensors.

Spatial resolution with optical fiber sensors can be attained by using

- Distributed sensors
- A grid of line integrated sensors
- A grid of point sensors

Distributed sensors are, in principle, very attractive for they would permit the use of fewer sensors and represent a more effective use of the optical fibers, in that each element of the optical fiber would be used for both measurement and data transmission. In practice, the spatial resolution attainable is somewhat limited. This is particularly true if the measurements have to be made in real time. Furthermore, the need for high reliability and graceful degradation tends to detract somewhat from the advantages of distributed sensors, since a single break in the optical fiber effectively takes out of commission a large number of sensors.

A grid of line-integrated fiber-optic sensors can be employed to map a scalar field, such as pressure or temperature, by use of the inverse Radon transform. Indeed, line-integrated measurements of position or angular orientation may even have specific value in certain forms of structural control situations. However, a grid of line-integrated optical fiber sensors cannot, in general, be used to map a vector or tensor field such as strain. Exceptions

can arise where the field is constrained in such a way that it is not truly two dimensional in nature. The recently proposed spatially weighted fiber-optic sensor of Murphy *et al.* (1991) could be useful where structural vibration analysis is required for applications like smart adaptive structures.

Where the field is two dimensional, three independent measurements are required at each point to uniquely specify the field. This can be accomplished by using an optical strain rosette sensor, comprising three small identical fiber-optic sensors set at a different orientation.

A. Strain Measurements

In terms of smart structures, strain and deformation represent two of the most important measurements.

TABLE IV Three Groups of Applications for Fiber-Optic Sensors Embedded in Advanced Composite Materials

Manufacturing	Installation	In-service operation
Cure mapping	Detect bad handling	Damage assessment
Residual stress measurement	Sense excessive force	Structural integrity monitoring
Defect detection	Warn of incorrect riveting	Real-time load evaluation
	Check bond integrity	Structural dynamic response monitoring
	Detect impacts	Opto-control monitoring
	Damage evaluation	Thermal mapping
		Strain and deformation mapping
		Remaining sevice lifetime evaluation

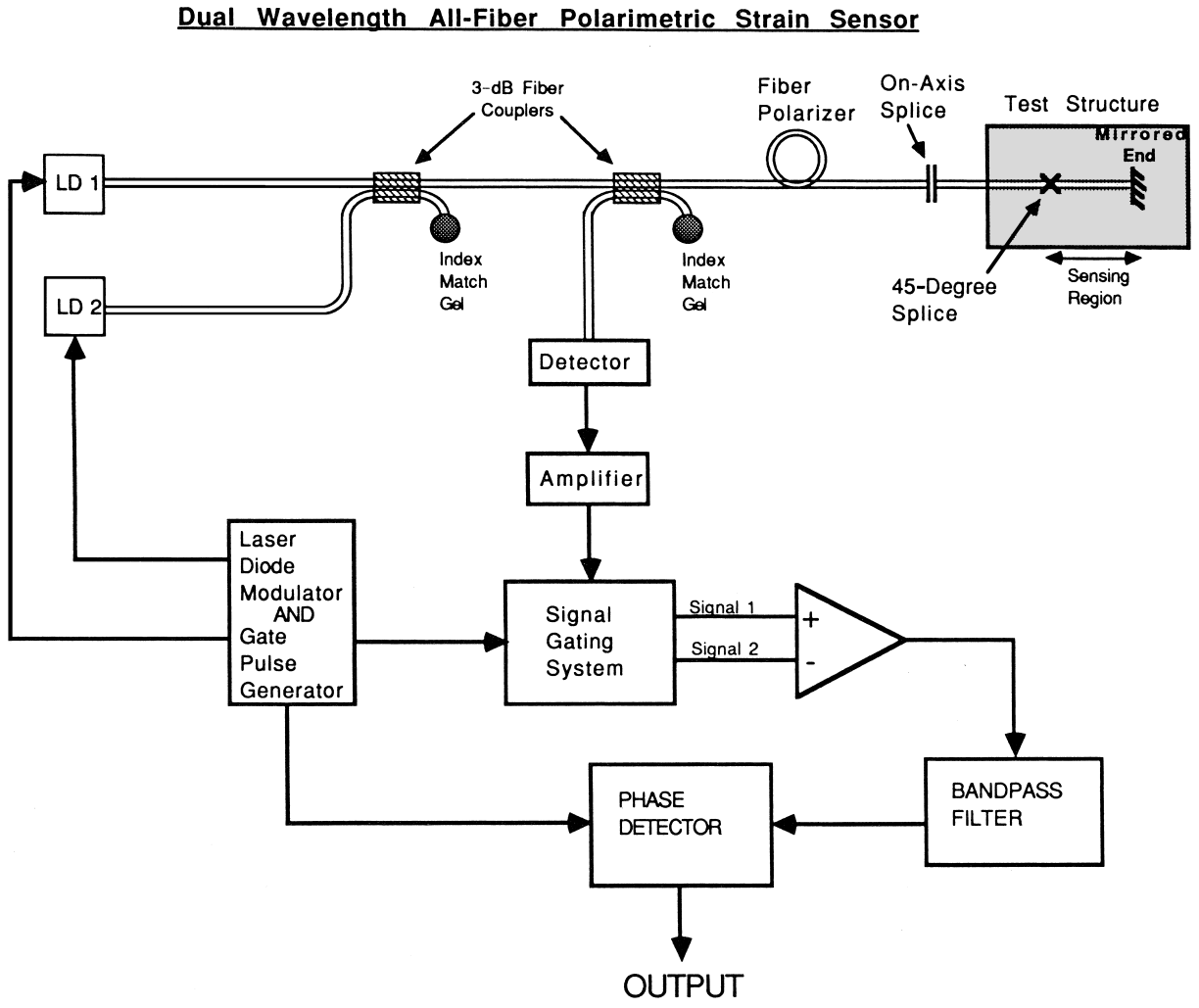


FIGURE 17 A schematic of a dual wavelength, single-ended, all-fiber polarimetric strain fiber-optic sensor.

Furthermore, real-time measurements of the strain in a structure normally permit its state of vibration to also be evaluated. Although a variety of fiber-optic sensing techniques (intensiometric, interferometric, polarimetric, and modalmetric) have been proposed for determining strain, in general, the interferometric-based approaches offer the highest sensitivity. Furthermore, it is expected that for most smart structure applications local measurements of strain will be required. Thus, in considering any sensing technique, the ease of ensuring lead-in, lead-out insensitivity will constitute an important factor.

The potential properties of an interferometric fiber-optic strain gauge (the term "gauge" is used to mean localized measurement) can be as good as, if not better than, a conventional electrical strain gauge in respect to temperature range, strain resolution, and maximum strain. However, the optical fiber strain gauge is likely to have superior frequency response and fatigue life and be more conducive to embedding within composite materials due to its small size and absence of the electrical leads and solder joints required by conventional resistive strain gauges.

Although it is possible to build a Mach–Zehnder fiber-optic strain sensor, a gauge based on a Michelson fiber-optic interferometer offers several advantages, including a better defined sensing region; a single-ended configuration; twice the sensitivity (as it is a double-pass system); and a much higher spatial resolution, because the Michelson is not limited by the minimum radius of curvature of the optical fiber, as is the Mach–Zehnder interferometric sensor. Unfortunately, both of these sensors require two closely spaced optical fibers. This represents a perturbation to any structure within which they are embedded, and the requirement of exact phase cancellation in their common lengths may not easily be achieved. Furthermore, application of this type of dual fiber sensor to smart structures could be hindered by the difficulty of producing a connector that could provide phase preservation between both optical fibers at all times and under all conditions of vibration and temperature.

We have undertaken the first strain measurements with a Michelson optical strain gauge embedded within various composite specimens, including a cantilever beam made of thermoplastic (carbon/PEEK). We have found an excellent linear relation between the number of fringes and the strain produced by end loading of this cantilever beam. Although considerable care was taken to position the two optical fibers very close together in order to ensure lead-in and lead-out insensitivity, some movement during the consolidation led to a primary source of phase noise. The phase-strain sensitivity for this double-pass device when embedded within graphite/PEEK was determined to be 12.3° ($\pm 1.6^\circ$) per μstrain per cm {or $2 \times 1.07 \times 10^7$ rad per strain per m}. Use of a 3×3 output coupler permits this all-fiber Michelson optical strain gauge to have passive quadrature detection.

An important consideration when considering any fiber-optic sensor system for a smart structure is the relative ease of developing a practical interface. In the case of interferometric fiber-optic strain sensors, this involves a reliable alignment of a number of single-mode optical fibers. This is no minor problem since the core diameter of each optical fiber is on the order of 1 μm. Although this task has been accomplished in the telecommunications field, smart structures impose a number of different constraints on the connector: it must be small (so as not to perturb the structure), it must not introduce any arbitrary phase even when the structure is subject to various forms of vibration and motion, and it must accurately maintain the alignment of the optical fibers over a wide range of temperatures.

The interface problems would be somewhat relaxed if the sensor involved only one optical fiber, and common mode rejection is far more effective, as both the sensing and the reference paths coincide in the lead-in and lead-out sections of the sensor. A fiber-optic strain sensor based on a Fabry–Perot interferometer operated in the reflection mode involves only one optical fiber, has a well defined sensing region, is a single-ended system with double-pass-enhanced sensitivity, and is capable of high spatial resolution. Furthermore, if operated at low finesse (mean reflectivity of cavity $<50\%$), the Fabry–Perot output is also sinusoidal, as it behaves under these conditions like a two-beam interferometer.

The difficulty of creating mirror surfaces within an optical fiber (Fig. 18) has represented a formidable barrier to the development of this sensor for smart structures. Lee *et al*. (1989) first demonstrated that Fabry–Perot-based fiber-optic sensors can be fabricated with reflective fusion splices. They formed the internal mirrors by means of titanium oxide spluttering. This constitutes an important advance, as it permits these sensors to be embedded within composite materials. We have been able to fabricate Fabry–Perot sensors using a reflective fusion splice based on a metal evaporation technique. The advantage of this approach is that we can create internal mirrors of relative high reflectivity (10–50%). This is important from a power budget point of view. We have measured the strain sensitivity of embedded Fabry–Perot strain gauges and shown that it is the same as that of the Michelson sensor. We have also shown that the strain response is linear when embedded within several different composite materials, Kevlar/epoxy, graphite/epoxy, and graphite/PEEK, and that it is free of hysteresis during load and unload cycles (Fig. 19). We have also shown that the strain sensitivity

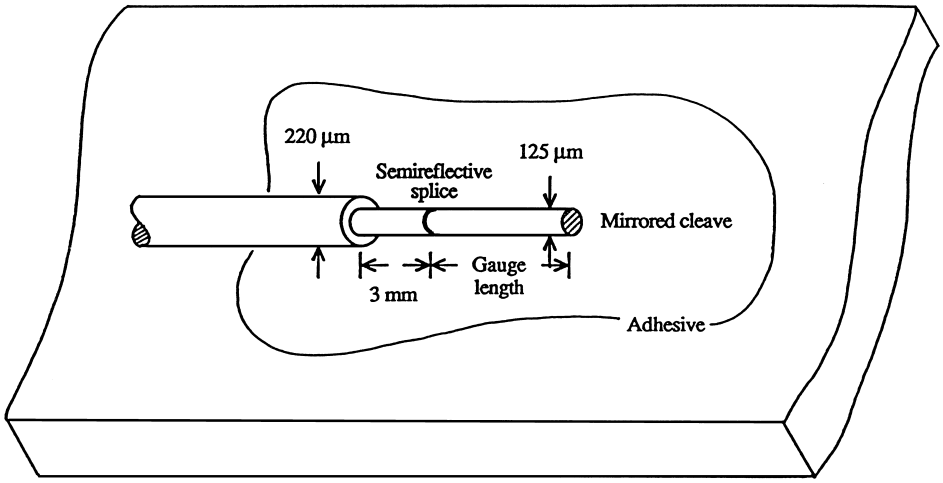
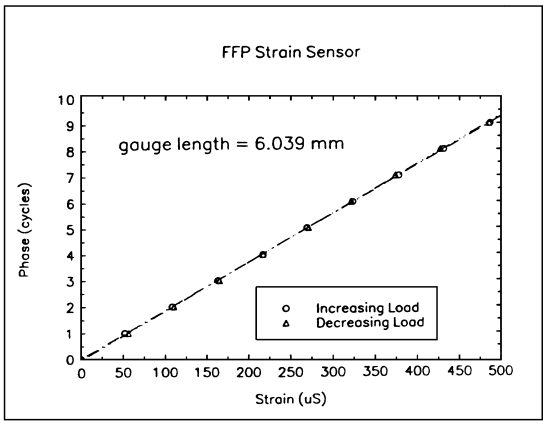
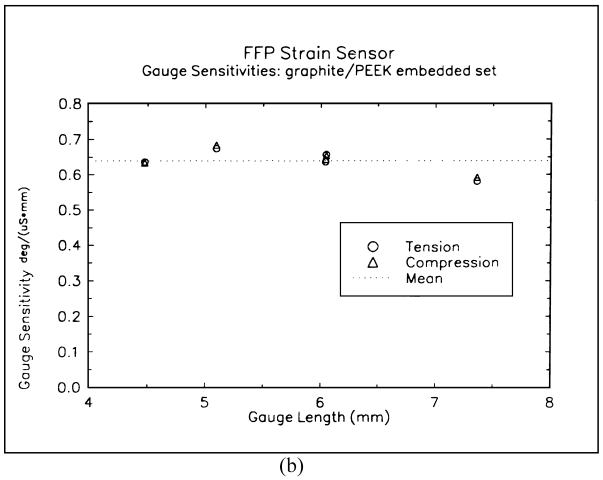


FIGURE 18 Detail of a localized, intrinsic Fabry–Perot fiber-optic strain gauge, with its first mirror in the form of a reflective fusion splice adhered to the surface of a cantilever beam.



(a)



(b)

FIGURE 19 (a) Load and unload phase-strain response for a Fabry–Perot fiber-optic strain gauge embedded within a graphite/PEEK cantilever beam. (b) Fabry–Perot fiber-optic sensor phase-strain sensitivity for a number of sensors of different gauge length embedded within graphite/PEEK cantilever beams. The error bars are compared in size to the symbols used.

of the Fabry–Perot sensor is not a function of its gauge length.

We embedded uniaxial Fabry–Perot sensors in graphite/PEEK and Kevlar/epoxy to assess sensor survival and response. In both cases, we used standard temperature and pressure profiles. The performance of these devices indicates that they may match, or eventually surpass, the performance of resistive-foil electrical strain gauges (see Table V).

Unfortunately, the semireflective fusion splice represents a source of structural weakness for the Fabry–Perot sensors. It is susceptible to failure in bending and tension at loads significantly less than observed with virgin fibers. The reason for this weakness is that at present the entire endface of each optical fiber is coated with a metallic film prior to fusion. This approach is unwarranted, as only the core needs to be reflective, and has the effect of preventing complete fusion of the two fibers. Ideally, the cladding regions of the endfaces should be glass-to-glass fused. In order to improve the strength of these sensors, we have developed a method of creating localized semireflective mirrors, an example of which is displayed in Fig. 20.

Another form of single optical fiber sensor that is under consideration is based on the creation of a Bragg

TABLE V Fabry–Perot Fiber-Optic Strain Gauge Characterization

Typical gauge length	3–10	(mm)
Optical fiber sensor diameter	125	(μm)
Operating wavelength	632.8	(nm)
Strain sensitivity	13.3	(degrees μstrain^{-1} cm^{-1})
Temperature sensitivity	55	(degrees K^{-1} cm^{-1})
Sensor strength: adhered or embedded (tension)	2000	(μstrain)

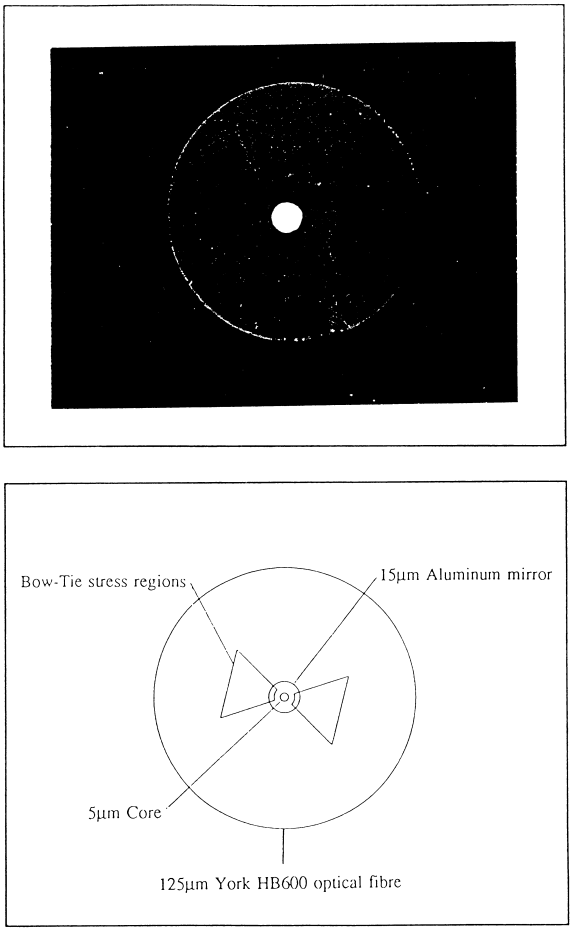


FIGURE 20 (a) Micrograph of a 15-μm diameter localized-mirror evaporated onto a 125-μm diameter York HB 600 optical fiber. (b) Schematic to illustrate a localized mirror with respect to the core and the bow ties.

grating along a designated length of an optical fiber with a germanosilicate (GeO_2-doped silica) core by means of photorefraction. Steplike changes in the core refractive index, $\Delta n/n$, of about 0.01% have been produced by exposure to intense laser radiation at around 245 nm. The resulting periodic index perturbation acts as a narrow band reflection filter (typically 25 GHz), with a center line frequency that shifts with a change of temperature or strain. Reflectivities in excess of 90% have been achieved with a 1-cm long grating that has a sensitivity of about 0.05 nm per Ksi. In principle, this approach could be adapted to produce a quasi-distributed fiber-optic sensor by impressing a number of such gratings, each with a different spacing and therefore different center line frequency, along the length of an optical fiber. Scanning the laser frequency would allow each of the sensing elements to be interrogated in turn. Additional virtues of this sensing concept are that it is single ended, it does not require quadrature detection, and it lends itself to wave-length division demultiplexing/multiplexing and simple demodulation.

1. Fiber-Optic Strain Rosette

In the event that an arbitrary state of strain in two dimensions needs to be determined, three independent measurements of the strain field have to be undertaken. This requires the development of the optical analog of the conventional foil strain rosette. We have fabricated the first fiber-optic strain rosette based on three Michelson fiber-optic sensors (Fig. 21). This prototype device has been tested for use in smart structures by embedding it within a graphite/PEEK cantilever beam and comparing its measurement of the strain with that of a conventional resistive strain rosette adhered to the surface of the beam. More recently, we have fabricated superior optical strain rosettes based on three Fabry–Perot fiber-optic sensors. Figure 22 shows a schematic of such an optical strain rosette. These have been tested within composite structures, and a comparison of their measurements to that of an electrical strain rosette have been undertaken. A photograph of both a small electrical strain rosette and the fiber-optic equivalent, illuminated with a helium neon laser, is presented in (Fig. 23). An appreciation of why it is much more practical to think of embedding the optical strain rosette within composites is immediately apparent, even with the overexposed optical fiber strain rosette. A representative example of such a comparison is displayed in Fig. 24. The agreement is excellent for a prototype device and clearly indicates the viability of using embedded optical strain rosettes for mapping strain fields within composite materials.

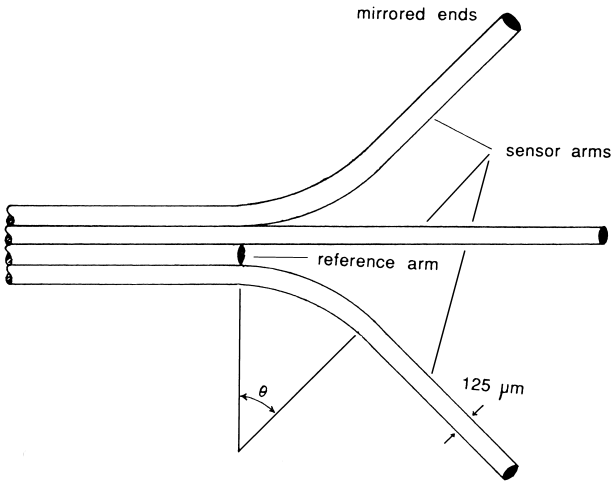


FIGURE 21 Physical arrangement of an optical strain rosette based on three Michelson fiber-optic sensors.

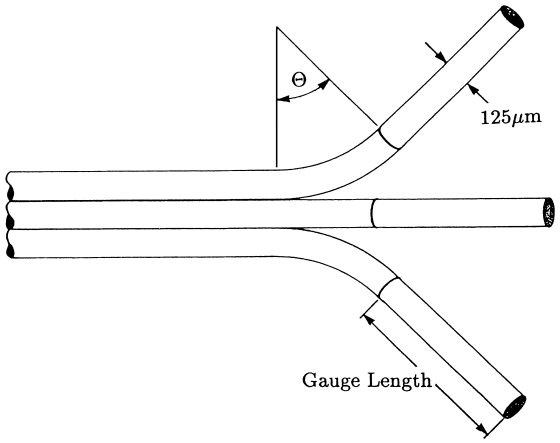


FIGURE 22 Physical arrangement of a practical optical strain rosette based on three Fabry–Perot fiber-optic sensors.

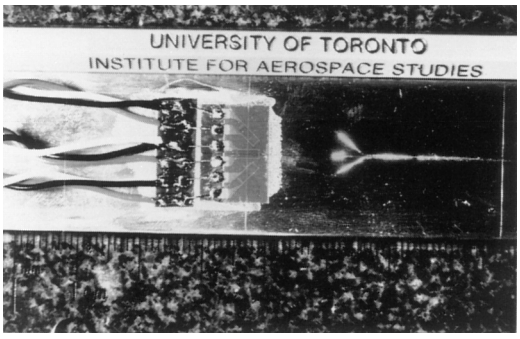


FIGURE 23 Photographic comparison of a conventional foil electrical strain rosette and a (somewhat overexposed) fiber-optic strain rosette.

2. Acoustic Emission Detection within Composites

based fiber-optic sensor permits it to detect acoustic energy released as a result of matrix cracking, fiber breakage, or interlamina debonding (delamination) within composite materials. Because of the ease of fabricating this kind of sensor and its adequacy for this type of measurement, we have used it within various multilaminated Kevlar/epoxy and graphite/epoxy panels for the purpose of studying damage-induced acoustic emission. A schematic of the Michelson fiber-optic sensor system used for this purpose is displayed as Fig. 25. The sensor comprises a pair of unbuffered optical fibers with mirrored ends embedded within the composite material. In this work only signals with a frequency greater than 70 kHz were permitted to be detected, with quadrature detection being ensured by a low frequency piezoelectric (PZT) phase modulation feedback system that also eliminated drifts and slowly varying strains.

The panels were subjected to out-of-plane loading, and the subsequent high frequency signals from the fiber-optic sensor were recorded. We have been able to demonstrate that embedded Michelson fiber-optic strain gauges can detect acoustic emission associated with the generation of threshold delaminations. A representative pair of acoustic emission signals and associated frequency spectra is represented as Fig. 26a. The corresponding enhanced backlight images revealing the delaminations associated with the detected acoustic signals are displayed below each frequency spectrum. We have noticed that if the ratio of acoustic energy in the frequency range 100–300 and 300–600 kHz is plotted as a function of delamination size (Fig. 26b), a trend is suggested that might lead to a method of assessing the extent of delamination detected by means of acoustic emission. However, the complexity of composite materials may frustrate this approach, and in the end use of *neural networks* may be required in order to locate and assess the extent and type of damage from acoustic emission signals.

By operating the Michelson sensor with our damage-sensitized optical fibers, we have been able to correlate,

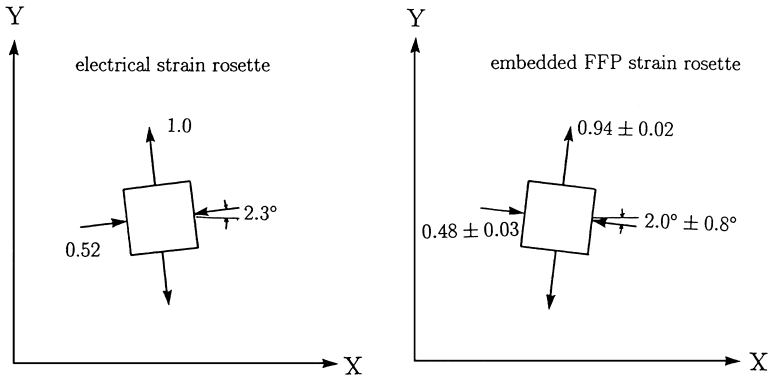


FIGURE 24 A comparison of the (projected) surface strain tensor (measured with a surface-adhered conventional foil electrical strain rosette) with the results from a Fabry–Perot fiber-optic strain rosette embedded with the Kevlar/epoxy cantilever beam.

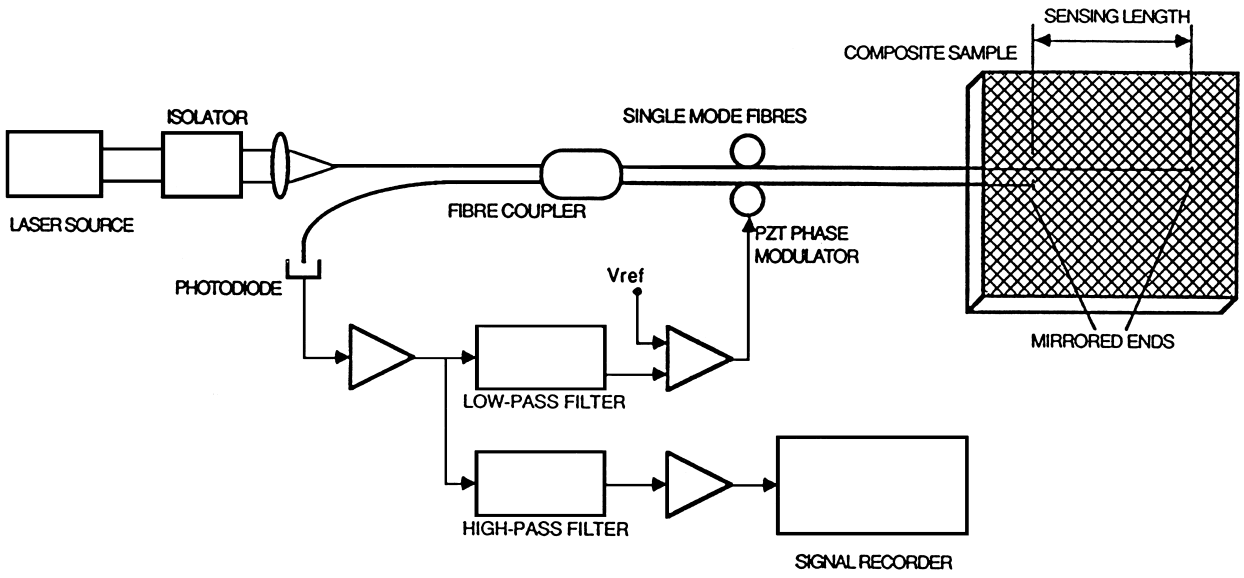


FIGURE 25 Schematic of the embedded Michelson fiber-optic sensor used to detect acoustic emission through a high pass filter. Also shown is the phase modulation technique to employed ensure quadrature detection.

for the first time, the occurrence of an acoustic signal with its concomittant matrix crack. More recently, we have demonstrated that the acoustic emission signals associated with the incremental growth of a region of delamination can be identified. This correlation of acoustic signals with their associated growth in the region of delamination could lead to an improved understanding of acoustic emission and how to use it for damage assessment. This growth in the area of delamination resulted from a compressive load applied to the specimen. This work serves to indicate that a combination of these techniques could be used to study, for the first time, the dynamics of delamination within translucent composites. This could enhance our understanding of such damage events and improve our ability to model composite materials response to excess loading.

Recently, we have also addressed the problem of acoustic wave-generated birefringence on the embedded fiber-optic sensors and have shown that the amplitude of the acoustic emission signals carries little useful information unless polarization-maintaining optical fibers are used.

B. Fabrication and Installation Monitoring

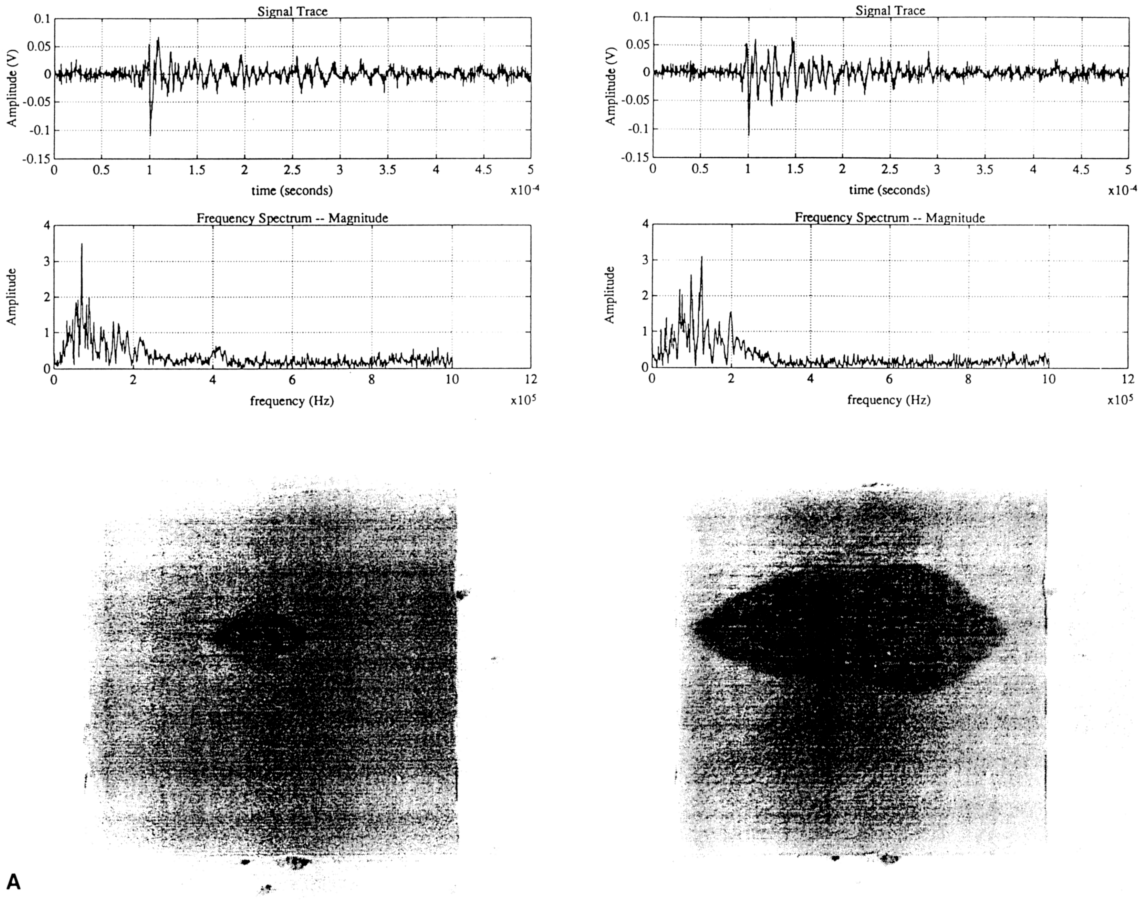
Fabrication monitoring of composite material structures could lead to both optimization and quality control of these products. Inspecting (or, more accurately, interrogating) the structure after installation would represent a bonus that could be very important with future aircraft and space structures, where errors leading to premature failures could be very costly.

Three examples will be considered:

1. Cure monitoring of thermoset composite materials
2. Consolidation monitoring of thermoplastic composite materials
3. Rivit and screw installation examination

Thermoset composite materials comprise strength-reinforcing fibers (such as carbon, aramid, or glass) embedded within a resin matrix (like epoxy). The optimum mechanical properties of these materials are only achieved if their fabrication includes the correct curing cycle at elevated temperatures and pressures. Unfortunately, the degree of cure during fabrication can vary with position for a large structure. Quality control, especially for aerospace structures, and cost demand a more controlled cure, and this can only be achieved if the cure state can be mapped by an *in situ* sensor.

Localized cure monitoring has been demonstrated by making a small section of an optical fiber from a sample of the fully cured resin used in the composite material. When this optical fiber is embedded within the uncured structure and illuminated, the segment made of the fully cured resin will guide the light, as its index of refraction is greater than that of the surrounding uncured material. As the structure cures (i.e., polymerizes), its index of refraction increases and the guiding properties of this sensing section of the optical fiber diminish. When the surrounding material is fully cured, the transmissions of the optical fiber falls to zero. The cure point is, as a consequence, unequivocally determined by a complete loss of light; this approach is seen to be indifferent to the temperature or pressure of the



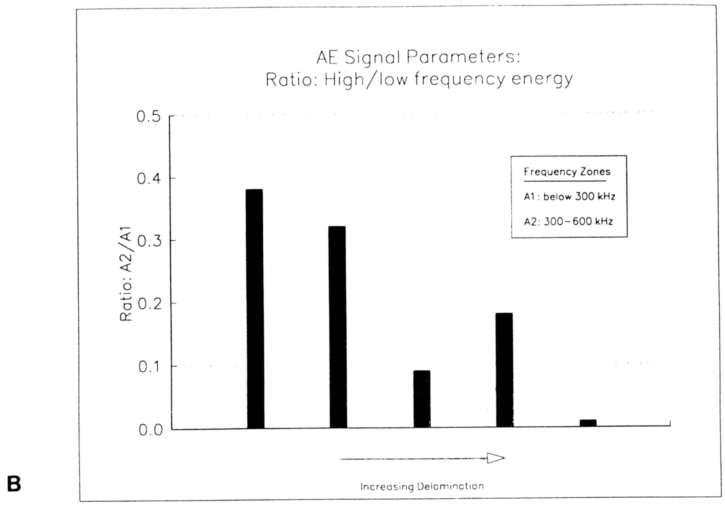


FIGURE 26 (A) A representative pair of acoustic emission signals and associated frequency spectra. The corresponding enhanced backlight images revealing the delaminations associated with the detected acoustic signals are displayed below each frequency spectrum. (B) The ratio of acoustic energy in the frequency range 100–300 and 300–600 kHz is plotted as a function of delamination size for Kevlar/epoxy specimens.

material and requires no calibration. Its major drawbacks are that it requires the sensing section of the optical fiber to be made of the same resin as the structure, and when cured these optical fibers can no longer be part of the ongoing optical neurosystem.

A totally different approach toward cure monitoring is based on observations of viscosity-dependent fluoresence of nonreactive probe dye molecules (a fluorophor) added to the resin. For example, diglycidolether of bisphenol-A has been shown to experience a 16-fold increase in its fluoresence during a room-temperature cure. The eventual outcome of this approach might be the development of a nonreactive fluorophor that can be used in conjunction with optical fibers, embedded to monitor the structural properties, to reliably monitor the cure state of the structure.

It might be possible to assess the state of cure of thermosets by measuring the velocity of acoustic waves transmitted through the material. The high sensitivity of interferometric fiber-optic sensors has led us to consider the prospect of measuring the velocity of optically generated acoustic waves within a thermoset and thereby monitoring the cure state with the same sensor that would later be used to determine the in-service strain and vibration state of the structure. Clearly, the more functions an embedded sensor can perform, the greater its cost effectiveness. We have recently initiated work on this optoacoustic approach.

The presence of embedded fiber-optic sensors capable of measuring both the pressure and the temperature distributions within a composite structure during the cure process might provide enough information to ensure optimization and quality control. The same may also be said for *thermoplastic*, where it is necessary to monitor the degree of consolidation. Indeed, we have demonstrated that a Michelson interferometric optical fiber sensor can be embedded within carbon/PEEK specimens and can be used to measure both temperature and strain within such structures.

Another important diagnostic role for which optical fibers may be quite suitable is checking various installation procedures. For example, the compressive load applied to a structure by bolts or rivets could be measured. In this context, either interferometric, polarimetric, modalmetric, or even microbend fiber-optic sensors may be suitable for checking the compressive loads associated with installation procedures. This is an important application because overtightening can lead to surface cracks and excessive subsurface stress concentrations, which can initiate matrix damage that would greatly shorten the fatigue life or, worse, initiate a major failure.

Although in the laboratory it is often easy to demonstrate the ability of a given fiber-optic sensor to monitor a specific event, this is a far cry from performing the same function in the very noisy world of operational systems. Localization is essential for fiber-optic sensors operating in an operational environment if they are to discern the desired signal from any extraneous noise.

C. Temperature Measurements

Temperature measurements can be required in their own right or to correct strain measurements. Most of the interferometric, polarimetric, and modalmetric strain sensors discussed above could also be used to evaluate the change in temperature, if it is known *a priori* that the applied strain remains constant. The temperature of a structure can also be evaluated by optical fibers in a variety of other ways. These include blackbody emission, differential absorption, fluorescence decay time, and Raman backscattering.

Localized hot spots can be detected by the blackbody emission arising from the section of optical fiber with an elevated temperature or by special optical fibers possessing sections of modified cladding for which the index of refraction becomes equal to or greater than that of the core at temperatures above a critical value. The optical fiber consequently becomes very lossy at temperatures above this value. Differential absorption has been used to measure temperatures with rare earth (Nd_2O_3-doped aluminosilicate glass core) optical fibers up to about 1000 K or with a chip of semiconductor (GaAs or CdTe) spliced into an optical fiber from 260 to 570 K. Thermooptical fiber-optic sensors can also be employed for the low temperature range (77–150 K) using Europium (Eu^{3+})-doped fluoride glass.

The dependence of the fluoresence decay time of neodymium (Nd^{3+})-doped glass with temperature when excited with infrared radiation has been shown to be capable of evaluating the temperature over the range 270–620 K. A great virtue of this approach is its independence from variations in the intensity or wavelength of the light-emitting diode source or the sensitivity of the detector. Although a distributed fiber-optic temperature sensor has been developed based on the ratio of Raman Stokes to anti-Stokes backscattering, its poor spatial resolution would limit its use to very large structures.

VI. DAMAGE ASSESSMENT WITH STRUCTURALLY INTEGRATED OPTICAL FIBERS

In the short term one of the simplest smart structure concepts to be implemented involves assessing damage within composite components with embedded optic fibers that fracture when a critical load is exceeded. This "built-in" sensing system will permit the structural

integrity of any composite structure to be monitored at all times. A network of suitable optical fibers could be used for a variety of damage assessment tasks. Current methods of aircraft structural integrity monitoring often involve both disassembling and component-by-component (sometimes rivet by rivet) inspection. This represents a very time-consuming and labor-intensive procedure that is often prone to human error. Considerable savings in time, cost, and human resources could be achieved if real-time, automatic, integrity-monitoring systems could be developed. One of the current approaches involves bonding crack-detecting wires onto components. Cracks encountering these wires interrupt their current and reveal their presence. Unfortunately, this technique has a number of serious drawbacks and limitations:

- Corrosion can affect wire crack-sensor performance.
- The sensitivity for fine cracks is poor.
- The system is subject to, and can generate, electrical interference.
- It cannot reliably be used within composites.

A. Damage Assessment for Composite Materials Based on Optical Fiber Fracture

Aircraft wing leading edges, submarine sonar domes, and propellers are examples of composite structures with complex geometries involving compound curvatures and a thickness that varies in two or more directions. These structures are very difficult to diagnose for damage by conventional techniques like ultrasound and X-radiography. Fracture of optical fibers, with the attendant disruption in their transmitted light, represents the simplest technique for damage detection. The first attempt at detecting damage within composite material with embedded optical fibers was based on the use of a regular buffered multimode optical fiber (Corning 1516 with a diameter of 250 μm) and was only capable of detecting quite severe damage. Another factor contributing to the relative insensitivity of the optical fibers to damage was the alignment of the optical fibers parallel to the reinforcing fibers.

Optical fibers have been shown to have particular relevance to the aircraft industry. They have been shown to be capable of detecting cracks in metal surfaces, bond failure in lap joints, airframe delaminations and fractures in the bonds between stringers and the airframe skin, and delaminations in composite structures.

1. Treatment, Orientation, and Depth Optimization

In working with Kevlar/epoxy we found it necessary to devise a special treatment for the optical fibers that enabled us to tailor their fracture strength to the point where they would detect barely visible impact damage (BVID). Control of the damage sensitivity was achieved by periodic etching of the sensing length of the optical fiber with a saturated solution of ammonium hydrogen difluoride. An example of the variation in the threshold impact energy for fracture of the optical fibers embedded in a specific manner within a given layup as a function of the etch time is presented as Fig. 27. The inset shows a section of intermittently etched optical fiber.

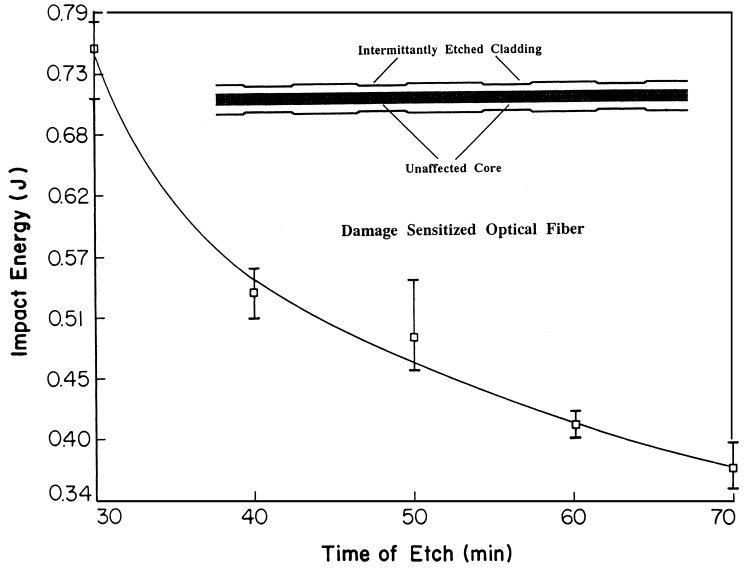


FIGURE 27 Variation with etch time of threshold impact energy for fracture (TIEF) of optical fibers embedded in the midplane of a 4-ply [0,90{0}90,0] Kevlar/epoxy laminate. The "0" in the curly bracket indicates that the optical fibers are oriented at 0° and are therefore orthogonal to the adjacent 90° plus. Inset: a schematic of a specially damage-sensitized (intermittently etched) segment of optical fiber.

An extensive analysis of the sensitivity to fracture of embedded, treated, optical fibers with respect to their orientation relative to the material-reinforcing fibers in the adjacent plies and their depth within a structure has been undertaken. The result of this work has allowed us to formulate the following three basic rules that are important in the design of composite damage assessment systems based on the fracture of embedded optical fibers:

1. The damage sensitivity of embedded optical fibers is increased by intermittant etching along the sensing length, and the longer the etch time is, the lower the threshold impact energy for their fracture in any given layup.
2. The closer the optical fibers are positioned to the surface of maximum tensile strain, the greater their damage sensitivity. For thin panels that can readily flex under impact, this surface coincides with the rear surface. For thicker laminates, or thin laminates that are rigidly reinforced, this is no longer the case due to the reduced flexure and high contact stresses. Under these circumstances, the optical fibers have to be embedded closer to the impact surface for optimum sensitivity.
3. The optimum placement of the embedded optical fibers is sandwiched orthogonally between collinear plies. In the event that optical fibers are positioned between plies that are not collinear, they should be oriented perpendicular to the ply closest to the surface of the maximum tensile strain (Fig. 28), unless this makes them collinear with the other ply, where-upon they should be oriented to bisect the largest angle between the plies.

We have recently shown that these rules, in regard to optimum orientation and location for damage sensitivity, established for Kevlar/epoxy appear to also apply to optical fibers embedded within graphite/epoxy. We have also demonstrated that fracture of embedded optical fibers can be used to map the boundaries of impact-induced delamination zones within fiber-glass shells.

2. Fracture Mechanism for Embedded Optical Fibers

We have established that each degree of treatment leads to a fairly well-defined *threshold impact energy for fracture* (TIEF) of the embedded optical fibers in any given composite layup and that this TIEF also depends on the orientation and location of the embedded optical fibers within the composite. Furthermore, these threshold energies can be closely matched to the threshold impact energy for delamination of the material by an appropriate choice of etch time for the optical fiber. The TIEF for optical fibers embedded collinear with the material-reinforcing fibers in the adjacent plies can be six times

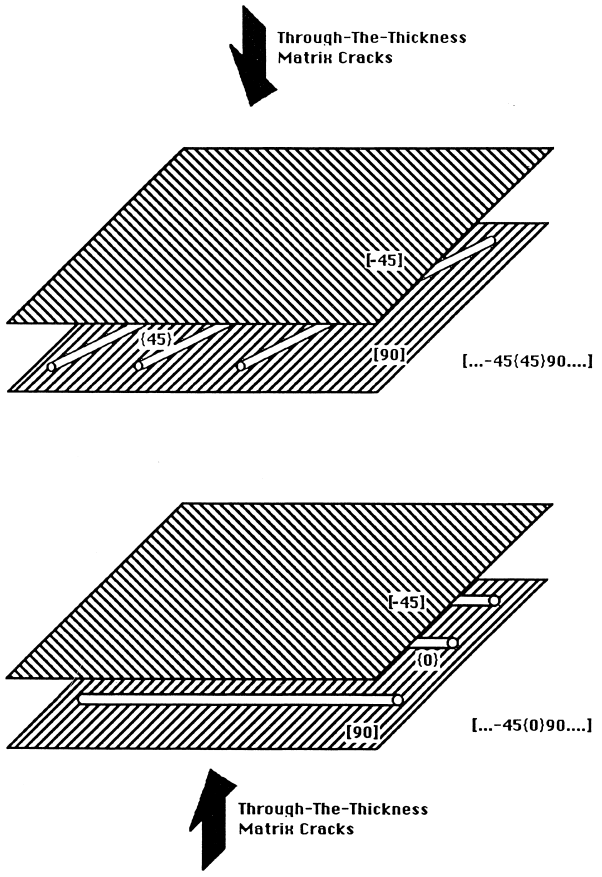


FIGURE 28 Illustration of the rule used to determine the best orientation for optical fibers embedded between two noncollinear plies of composite material for the purpose of evaluating damage based on fracture of the optical fibers. This rule states that the optical fibers should be oriented orthogonal to the ply (i.e., the direction of the reinforcing material fibers) that lies between the optical fibers and the surface of maximum tensile strain.

the TIEF of a similar optical fiber placed in the same location within the same composite but oriented perpendicular to the reinforcing fibers. Results of this nature and other observations have led us to suggest that the mechanism responsible for fracturing embedded optical fibers is primarily through-the-thickness matrix cracks (see Fig. 29). The above rules are consistent with our understanding that through-the-thickness matrix cracks constitute the principal fracture mechanism for embedded optical fibers. A list of some of the fundamentals of composite damage assessment based on the fracture of embedded optical fibers is presented in Table II.

3. Detection of Load-Induced Growth of a Damaged Area

In a series of experiments that combined light transmission measurements with light bleeding we have been able

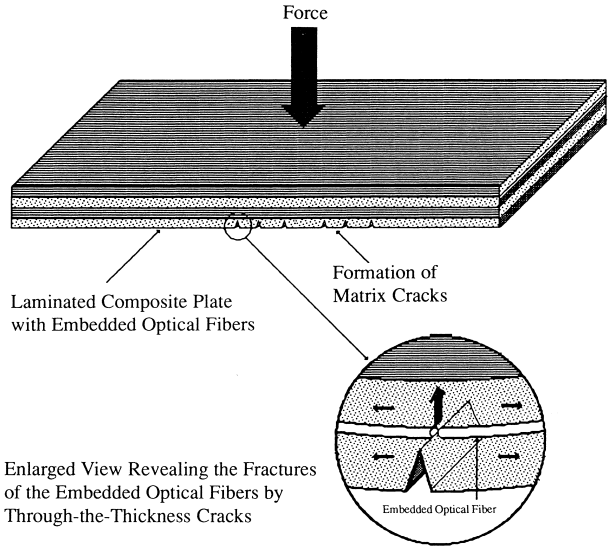


FIGURE 29 Fracture mechanism for optical fibers embedded within composite materials is primarily through-the-thickness matrix cracks.

to demonstrate that load-induced *growth* of a region of delamination within a composite panel could be detected by an embedded grid of suitably damage-sensitized optical fibers. These experiments involved out-of-plane loading of Kevlar/epoxy panels of various layups and thicknesses. The delamination damage produced by out-of-plane loading was determined with image-enhanced backlighting, even in the case of complex composite structures such as the thick panel, comprising 13 different plies and a honeycomb reinforcement, used to simulate a planar section of the leading edge of an aircraft wing.

Figure 30 provides a representative set of results that reveal (Fig. 30a) fracture of embedded optical fibers (as seen by light bleeding—the dark spots) within the region of delamination (grey shadow produced by image-enhanced backlighting); Figure 30b shows that as the load is increased and this region of delamination grows, more fractures of the optical fibers are observed. In most situations the changes in the tranmission characteristics of the imbedded optical fibers will be used for damage assessment as observation of light bleeding will be impractical. Our results to date are encouraging and strongly support the viability of using embedded optical fibers to assess the growth of a region of delamination under load conditions. The original cause of this delamination could have been an impact or a manufacturing flaw.

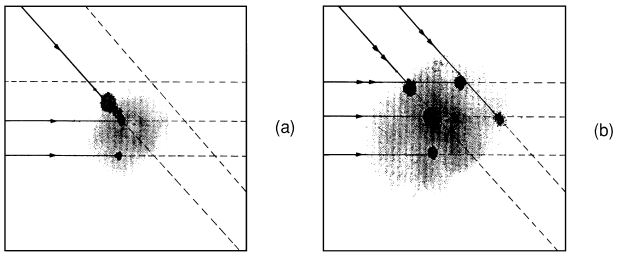


FIGURE 30 Growth in the area of delamination within a Kevlar/epoxy coupon with a $[0_W,0_2,45,0_2\{90\}0,-45\{45\}90,0_W$, Honeycomb, $0_W,90,0_W]$ layup subject to a central out-of-plane force is clearly seen by comparison of the two image-enhanced backlit video images. The curly brackets $\{x\}$ indicate the location and x indicates the orientation in degrees of the imbedded optical fibers. The shadow (grey area) reveals the region of delamination region, while the light bleeding (dark) spots indicate the fracture locations of the embedded optical fibers orientated at 90° and 45°.

B. Fiber-Optic Damage Assessment System for an Aircraft Wing Leading Edge

We collaborated with Boeing (De Havilland Division) of Canada to fabricate a prototype aircraft wing leading edge instrumented with a structurally integrated fiber-optic damage assessment system (Fig. 31). This system uses three arrays of damage-sensitized optical fibers embedded within three separate lamina (Fig. 32). These optical-fiber arrays were embedded within the leading edge after being mounted in the appropriate configuration on thin adhesive films. This is illustrated for the optical fibers oriented at 90° in Fig. 33. Although visual inspection of the optical fiber integrity is quite feasible, the human eye is not good at discerning modest changes (say a 30% drop) in intensity associated with a partial fracture. A photodiode array that is butt-fitted against the output port

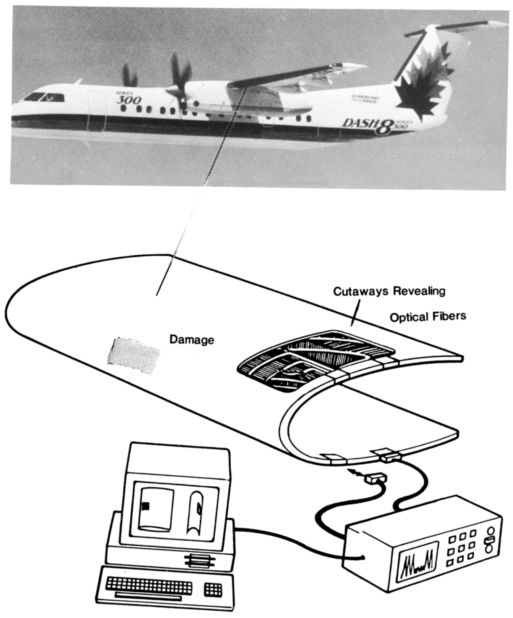


FIGURE 31 Schematic of aircraft wing leading edge instrumented with three arrays of damage-sensing optical fibers.

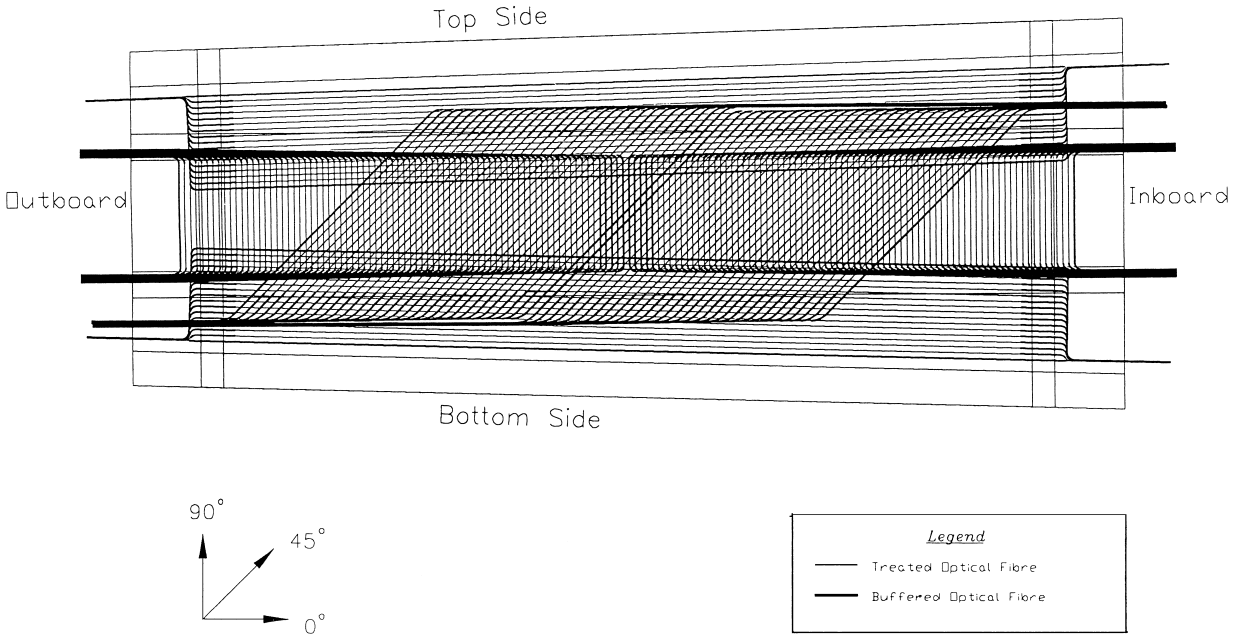


FIGURE 32 Arrangement of the three arrays of embedded optical fibers that constitute the damage-sensing system in the aircraft wing leading edge.

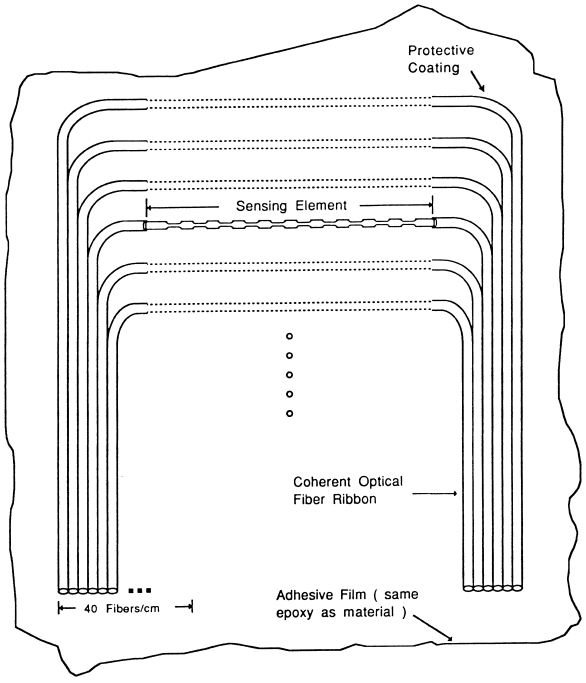


FIGURE 33 A schematic illustration of the damage-sensitized 90° optical fibers as laid in the resin film for mounting within the leading edge.

for the embedded optical fibers will thus be used in evaluating this system. The output from this photodiode array will be interpreted by a computer, and a display of the predicted region of damage will be presented on a monitor. This type of system could be used for periodic interrogation when the leading edge is dismounted from the wing or adapted for real-time monitoring during flight using an onboard computer. Figure 34 is a photograph of the first aircraft leading edge with a built-in fiber-optic damage detection system undergoing tests at the University of Toronto Institute for Aerospace Studies. Because of the translucent nature of the Kevlar/epoxy leading edge, the relatively weak leakage of laser light along the path of the embedded optical fibers can be seen, in addition to the relatively bright laser light "bleed spots" corresponding to the optical fiber fracture locations.

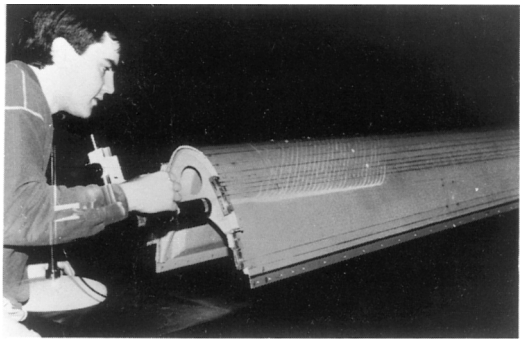


FIGURE 34 Photograph of the first aircraft leading edge with a built-in damage detection system undergoing tests at the University of Toronto Institute for Aerospace Studies. Due to the translucent nature of the Kevlar/epoxy leading edge, the relatively weak leakage of laser light along the path of the embedded optical fibers can be seen in addition to the relatively bright laser light bleed spots corresponding to the optical fiber fracture locations.

We have tested the reliability of the embedded optical fiber array to correctly assess the location and extent of damage inflicted by impacts. This was accomplished by subjecting the leading edge to controlled impacts and comparing the disruptions in the transmission through the optical fibers with the formation of new laser light bleed spots. These results were then compared with the damage measured by image-enhanced blacklighting. The remarkable diagnostic potential of image-enhanced backlighting can be appreciated when it is seen how well it detects delamination through the curved, complex, multilaminated structure (including a honeycomb core) that constitutes the leading edge.

A representative result is presented as Fig. 35. Figure 35a is a subtracted backlit image of a small section of the leading edge, the location of which is indicated in

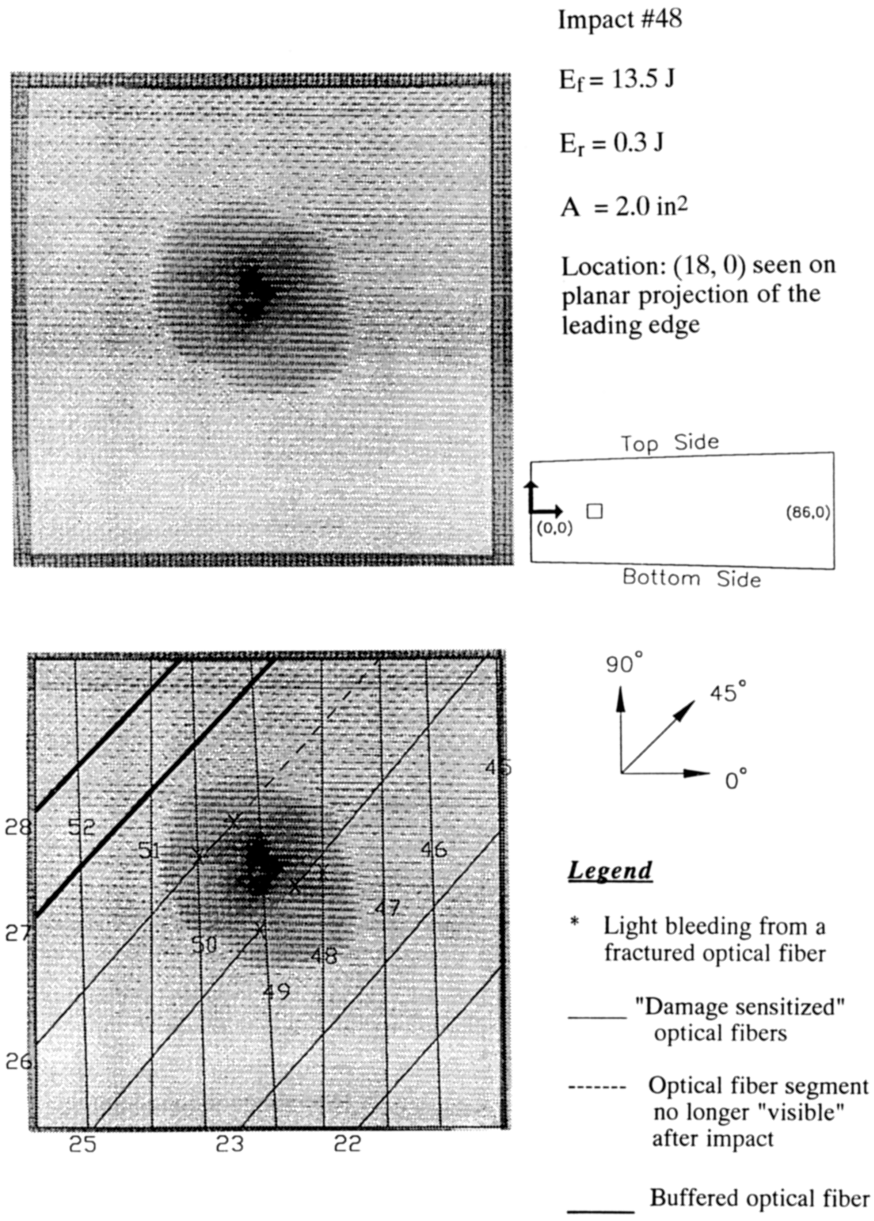


FIGURE 35 The upper picture is a subtracted backlit image of a small section of the leading edge, the location of which is indicated in the small sketch. It reveals the presence of a delamination produced by a 14.6-J impact. The lower picture shows the array of embedded optical fibers superimposed on this subtracted backlit image and indicates the points of fracture of these embedded optical fibers. The two buffered (and therefore untreated) optical fibers are seen to remain unscathed by the impact.

the small sketch. This picture reveals with surprising clarity the presence of a delamination produced by a 14.6-J impact. Figure 35b shows the array of embedded optical fibers superimposed on the subtracted backlit image. It can be seen that all of the damage-sensitized optical fibers within the damage shadow were fractured. It is also noteworthy that the two buffered (and therefore untreated) optical fibers were unscathed by the impact, providing strong evidence of the need to treat the optical fibers if they are to be used in this modality.

VII. FUTURE PROSPECTS

Smart structure technology could lead to a revolution in engineering concepts and ethics. In the 21st century it might be quite unacceptable engineering to build any structure or vehicle (that could endanger the lives of people or the environment) without including an integrated sensor system to warn of impending mechanical failure. For pipelines, pressure vessels, storage tanks, etc., simple built-in fiber-optic structural monitoring systems could provide an early warning of structural weaknesses and prevent many of the failures that today are responsible for a number of environmental problems. In major structures such as dams, nuclear reactors, or large buildings, this kind of system could warn of the development of structural faults or, after events like earthquakes, provide important information regarding structural soundness.

In aircraft, where composite materials are finding increasing application, structurally integrating fiber-optic sensor systems capable of detecting damage and assessing its growth in time could find acceptance as part of the new fly-by-light technology (Fig. 2). Through the use of artificial intelligence such a system might also determine the effect of this damage in terms of limiting flight maneuvers or evaluating the reduction in the useful life of duly instrumented components. This could also greatly improve maintenance efficiency and effectiveness.

Beyond structural integrity, the concept of using light in conjunction with optical fibers for sensing the shape, deformation, temperature, state of vibration, etc. also looks very promising. Indeed, it may eventually be possible to build aircraft with wings that sense their load distribution with arrays of integrated fiber-optic sensors and adapt their shape directly with distributed actuators to optimize flight control. Such wings would have no ailerons, flaps, motors, or cables. Everything would be integrated into the structure, effecting a considerable weight saving, a major simplification in the construction, a significant reduction in maintenance and downtime, and a commensurate improvement in reliability and safety. Furthermore, if this can be accomplished with opto-control systems, there would be additional improvements in safety. This technology might also find application to space platforms, especially those large, flexible space structures where active shape control is required, and to robotic systems where human-like dexterity could be of value.

Structures with a fiber-optic nervous system and a structurally integrated distribution array of actuators (or muscles) might best be controlled by neural networks that would be capable of adaptive learning. Such entities might learn from their own experiences and from their environment and evolve into truly intelligent structures.

SEE ALSO THE FOLLOWING ARTICLES

COMPOSITE MATERIALS • CONSTITUTIVE MODELS FOR ENGINEERING MATERIALS • FIBER-OPTIC CHEMICAL SENSORS • FRACTURE AND FATIGUE • LASERS, OPTICAL FIBER • MECHANICS OF STRUCTURES • MULTIPLEXING • OPTICAL FIBERS, FABRICATION AND APPLICATION

BIBLIOGRAPHY

Afromowitz, M. A., and Lam, K. Y. (1988). "Fiber optic cure sensor for thermoset composites," *Fiber Opt. Smart Struct. Skins* **986,** 135–139.

Ahmad, I., Crowson, A., Rogers, C. A., and Aizawa, M. (1990). "US-Japan Workshop on Smart/Intelligent Materials and Systems," Technomic Publishing.

Blagojevic, B., Tsaw, W., McEwen, K., and Measures, R. M. (1989). "The influence of embedded optical fibers on the interlamina fracture toughness of composite materials," Review of Progress in Quantitative NDE, Brunswick, ME, July 23–28.

Bertholds, D., and Dandliker, R. (1988). "Determination of the individual strain-optic coefficients in single-mode optical fibers," *J. Lightwave Technol.* **6,** 17–20.

Butter, C. D., and Hocker, G. P. (1978). "Fiber optics strain gauge," *Appl. Opt.* **17,** 2867.

Culshaw, B., and Dakin, J., eds. (1988). "Optical Fiber Sensors: Principles and Components," Artech House, Boston and London.

Culshaw, B., and Dakin, J., eds. (1989). "Optical Fiber Sensors: Systems and Applications," Artech House, Boston and London.

Czarnek, R., Guo, Y. F., Bennett, K. D., and Claus, R. O. (1988). "Interferemetric measurements of strain concentrations induced by an optical fiber embedded in a fiber reinforced composite," *Fiber Opt. Smart Struct. Skins I* **986,** 43–54.

Dakin, J. P., and Wade, C. A. (1984). "Compensated polarimetric sensor using polarization-maintaining optical fiber in a differential configuration," *Electron. Lett.* **20,** 51–53.

Dasgupta, A., Wan, Y., Sirkis, J. S., and Singh, H. (1990). "Micromechanical investigation of an optical fiber embedded in a laminated composite," *Fiber Opt. Smart Struct. Skins III* **1370,** 119–128.

Farahi, F., Webb, D. J., Jones, J. D. C., and Jackson, D. A. (1990). "Simultaneous measurement of temperature and strain: Cross-sensitivity considerations," *J. Lightwave Technol.* **8,** 138–142.

Freidah, J. T., Cahill, R. F., Turner, A. A., Holmes, S. A., and Wagoner, R. E. (1985). "Passive homodyne optical grating demodulator: Principles and performance," *Fiber Opt. Laser Sens. III* **566,** 114–121.

Giallorenzi, T. G., Bucaro, J. A., Dandridge, A., Sigel G. H., Jr., Cole, J. H., Rashleigh, S. C., and Priest, R. G. "Optical fiber sensor technology," *IEEE J. Quantum Electron.* **QU-18,** 626–665.

Glossop, N. D., Dubois, S., Tsaw, W., LeBlanc, M., Lymer, J., Measures, R. M., and Tennyson, R. C. (1990). "Optical fibre damage detection for an aircraft composite leading edge," *Compos. J.* **21,** 71–80.

Hofer, G. B. (1987). "Fiber optic damage detection in composite structures," *Composites* **18,** 309–316.

Hogg, W. D., Turner, R. D., and Measures, R. M. "Polarimetric fiber optic structural strain sensor characterisation," *Fiber Opt. Smart Struct. Skins* **1170,** 60.

Jackson, D. A., and Jones, J. D. C. (1986). "Fiber optic sensors," *Opt. Acta* **33,** 1469–1503.

Jensen, D. W., and Pascual, J. (1990). Degradation of graphite/bismaleimide laminates with multiple embedded fiber optic sensors. *Fiber Opt. Smart Struct. Skins III* **1370,** 19–21.

Koo, K. P., Tveten, A. B., and Dandridge, A. (1982). "Passive stabilization scheme for fiber interferometers using (3×3) fiber directional coupler," *Appl. Phys. Lett.* **41,** 616–618.

Le Blance, M., Dubois, S., McEwen, K., Hogg, D., Park, B., Tsaw, W., and Measures, R. M. (1989). "Development of a fiber optic damage detection system for an aircraft leading edge," *Fiber Opt. Smart Struct. Skins* **1170,** 25.

Liu, K., Ferguson, S. M., and Measures, R. M. (1989). "Damage detection in composites with embedded fiber optic interferometric sensors," *Fiber Opt. Smart Struct. Skins* **1170,** 22.

Liu, K., Ferguson, S. M., McEwen, K., and Measures, R. M. (1990). "Acoustic emission detection for composite damage assessment using ordinary single mode interferometric sensor," *Fiber Opt. Smart Struct. Skins III* **1370,** 316–323.

Liu, K., and Measures, R. M. (1992). "Signal processing techniques for interferometric strain sensors," Accepted for the *J. Intelligent Mate. Syst. Struct.*

Loken, D. (1990). "Effect of Fiber Straightness on Fatigue of Aligned Continuous CFRP Composites," M.A.Sc. Thesis, University of Toronto (Chemical Eng.).

Measures, R. M. (1989a). "Smart structures with nerves of glass," *Prog. Aerosp. Sci.* **26,** 289–351.

Measures, R. M. (1989b). "Fiber optics smart structures program at UTIAS," *Fiber Opt. Smart Struct. Skins* **1170,** 11.

Measures, R. M., Melle, S., and Liu, K. (1992). "Wavelength demodulated. Bragg grating fiber optic sensing systems for addressing smart structure critical issues," Accepted for the *J. Smart Mater. Struct.* **1** (1).

Morey, W. W., Meltz, G., and Glenn, W. H. (1989a). "Fiber optic Bragg grating sensors," *Fiber Opt. Laser Sens.* **1169,** 12.

Morey, W. W., Meltz, G., and Glenn, W. H. (1989b). "Bragg Grating Temperature and Strain Sensor," Springer Proceedings in Physics 44, Optical Fiber Sensors, Proceedings of the 6th International Conference, OFS'89, Paris, France, September, 18–20, pp. 526–531.

Murphy, K. A., Fogg, B. R., Vengsarkas, A. M., and Claus, R. O. (1991). "Spatially weighted fiber optic sensors for smart structure applications," Active Materials and Adaptive Structures, Alexandria, Nov 4–8.

Sirkis, J. S., and Dasgupta, A. (1990). "Optimal coatings for intelligent structure fiber optic sensors," *Fiber Opt. Smart Struct. Skins III* **1370,** 129–140.

Sirkis, J. S., and Haslach, H. W., Jr. (1990). "Full phase-strain relation for structurally embedded interferometric optical fiber sensors," *Fiber Opt. Smart Struct. Skins III* **1370,** 248–259.

Tay, A., Wilson, D. A., and Wood, L. (1989). "Strain analysis of optical fibers embedded in composite materials using finite modelling," *Fiber Opt. Smart Struct. Skins II* **1170,** 521–533.

Turner, R. D., Valis, T., Hogg, W. D., and Measures, R. M. (1989). "Fiber optic strain sensors for smart structures," *J. Intelligent Mater. Syst. Struct.* **1,** 26–49.

Udd, E., ed. (1988). *Fiber Opt. Smart Struct. Skins* **986**.

Udd, E., ed. (1990). "Fiber Optic Smart Structures and Skins III," San Jose, September 16–21, SPIE Vol. **1370**.

Udd, E., ed. (1991). "Fiber Optic Sensors," Wiley, New York.

Valis, T., Tapanes, E., and Measures, R. M. (1989). "Localized fiber optic strain sensors embedded in composite materials," *Fiber Opt. Smart Struct. Skins* **1170,** 53.

Valis, T., Hogg, D., and Measures, R. M. (1990). "Fiber Fabry–Perot strain gauge," *IEEE Photonics Technol. Lett.* **2,** 227–228.

Valis, T., Tapanes, E., Liu, K., and Measures, R. M. (1991). "Passive quadrature demodulated localized Michelson fiber optic, strain sensor embedded in composite materials," *J. Lighwave Technol.* **9,** 535–544.



Electric Machines, Rotating, Construction and Theory

Leo Birenbaum
Enrico Levi
Polytechnic University

I. Introduction
II. Direct Current Machine
III. Synchronous Machine
IV. Induction Machine
V. Stepper and Switched-Reluctance Motors
VI. Written-Pole Motor and Generator

GLOSSARY

Armature Portion of a rotating machine on which the active conductors are located; most often, the rotor of the dc machine or the stator of the synchronous machine.

Brushes Set of small, stationary graphite blocks mounted so that they press against the commutator (or slip rings). They make electrical contact with the moving rotor.

Commutator Cylindrically arranged set of copper segments, mounted on the rotor structure, insulated from one another and from the shaft, and to which the ends of the armature coils are connected. In a dc motor, the commutator reverses the current direction in the coils periodically so that the direction of the electromagnetic torque always remains the same.

Electromotive force (emf) Voltage induced in the active conductors of the armature by (relative) motion through the magnetic field.

Field Portion of a rotating machine that creates the magnetic field through which the active conductors move; most often, the stator of the dc machine and the rotor of the synchronous machine.

Induction motor Polyphase (or single-phase) machine with a stator winding like that of a synchronous machine and a rotor in which voltages are induced by its motion through the field created by the stator winding currents. Also called an ac machine. The rotor may be of the squirrel-cage type, or in the polyphase machine, it may also be wound, with slip rings for electrical accessibility.

Magnetomotive force (mmf) Product of current through a coil and the number of turns (ampere turns)—the quantity that creates the magnetic flux in a given region.

Rotor Rotating (central) part of a rotating machine.

Slip rings Set of two, three, or four copper rings mounted on and concentric with the rotor shaft, but insulated

from it, and connected to the windings of the rotor. Usually, they supply dc current to the field winding of a synchronous machine or connect to the terminals of the rotor winding on a wound-rotor induction motor.

Synchronous machine Rotating machine in which the speed is rigidly tied to the electrical frequency. Most electric power is generated by three-phase synchronous machines.

Torque Turning moment (force times distance from the center of rotation) that is applied to a generator rotor, as by a turbine, or that is electromagnetically generated in a motor to make it turn.

ROTATING ELECTRIC MACHINES are used mainly to convert energy between electrical and mechanical forms. They come in sizes ranging from milliwatts to more than 1 gigawatt (10^9 W), and they operate very efficiently, especially in the higher ratings. Their main types are direct current (dc), synchronous, and induction machines.

I. INTRODUCTION

This article presents some background on three basic rotating machines—dc, synchronous, and induction machines. The intention is to describe their operation in a qualitative way first and then to present a more detailed analytical quantitative picture of each machine.

It will become apparent that the difference between a motor and a generator is merely an operational one: When electrical power is fed to the terminals of a motor, it turns, and mechanical power can be taken from the shaft. On the other hand, if the same machine is mechanically driven, electrical power can be taken from its terminals. Hence, the same machine is either a motor or a generator, depending on how it is used.

The dc machine was chosen for discussion here because conceptually it is the simplest, the oldest, and the most easily controlled—that is, its speed can be varied very easily, whereas the others are, by their nature, devices of relatively fixed speed needing complex solid-state drives to achieve efficient speed control. This probably explains why the dc machine has refused to disappear, despite its greater maintenance requirements and the need to provide dc rather than ac power for its operation. The dc machine is heavily relied on to supply traction for locomotives and railroad cars on diesel-powered as well as on electrified railroad systems.

The synchronous machine was chosen because almost all electrical power is generated by synchronous machines driven mechanically by steam turbines in coal- or oil-fired plants or in nuclear plants; by water turbines in hydroelectric plants; by gas turbines in smaller generating units; and by diesel engines in standby or emergency generators.

The induction machine was chosen because it is by far the most commonly used motor in commerce, industry, and the home. The range of sizes is enormous, varying from a fraction of a horsepower to thousands of horsepower.

The fact that only three machines have been chosen for discussion does not mean that the enormous variety of engineering devices that have been conceived and built since Faraday's first motor necessarily consist of only these three types. However, the basic principles needed to explain their operation are fundamental to the understanding of all these devices.

II. DIRECT CURRENT MACHINE

A. Construction

Figure 1 shows the construction of a typical dc machine. The outer part (the stator) is stationary, and the inner part (the rotor) rotates.

The outer part consists of the frame, a yoke that fits within the frame, and sections (poles) protruding inward

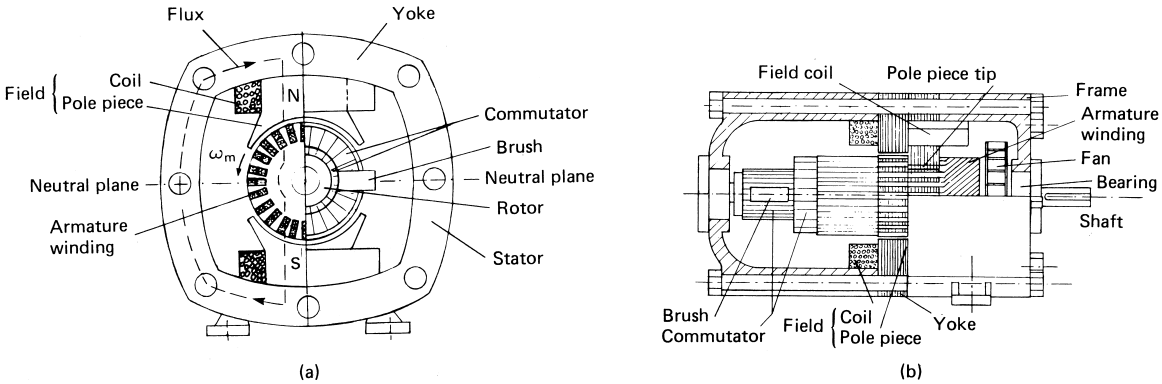


FIGURE 1 Two-pole dc machine. (a) Cross section; (b) cutaway side view.

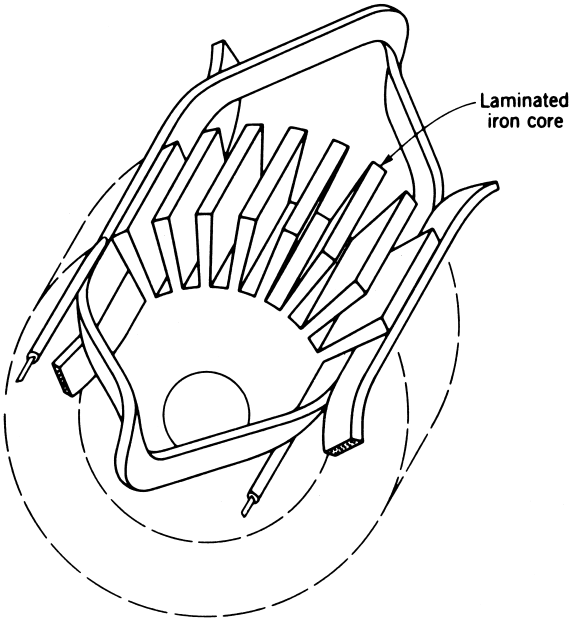


FIGURE 2 Typical two-layer arrangement of multiturn coils on the rotor. [Courtesy of Krieger Publishing Company, Melbourne, Florida.]

from the yoke. These are made of solid or laminated iron. On these poles, coils are wound. When dc current flows through this "field" winding, the poles become magnetized, as shown. Although Fig. 1 depicts only two poles, dc machines are often built with a number of pole pairs of alternating north–south polarities. The rotor (more often called the armature) is thus located in a magnetic field created by the field winding on the stator.

The inner part is more complex. It consists of a central shaft, a laminated iron core slotted at its periphery, a set of coils that fit in the slots, and a set of copper commutator bars, each of which is insulated from its neighbors and from the shaft. The ends of the armature coils are electrically connected to appropriate commutator bars. The coil sides are usually embedded in the slots in two layers (Fig. 2).

Pressing against the commutator structure is a set of stationary graphite "brushes" held in a brush rigging. Pressure is maintaind by brush springs. Brushes are devices for establishing continuous electrical contact with the rotating armature. The material most commonly used is graphite, because it is soft and does not degrade the copper commutator.

With the field due to the stator winding present, suppose that a source of dc voltage is connected to the brushes. Current flows into one brush, flows into the segments below, and leaves through the other brush after having passed through the armature winding in some fashion. The armature starts to turn, accelerates, and finally reaches a steady speed.

B. Principles of Operation

The armature winding is a closed one. It is wound so that all of the conductors lying in the slots above the neutral plane (Fig. 1a) carry current flowing in one direction (away from the commutator) and all the conductors lying in slots below the plane carry current flowing in the opposite direction. As soon as the voltage is applied between the brushes, the current starts to flow in the manner indicated in Fig. 3 for a single one-turn coil.

It appears that the commutator switches the connection between the external circuit and the coil in such a manner that the dc supply can provide the conductors of the coil with a current that alternates in sign as they pass from a north to a south pole.

There is a physical law that states the following. When a conductor of length L and carrying the current I_{co} is exposed to a perpendicular magnetic field of flux density B, it experiences an electromagnetic force of magnitude

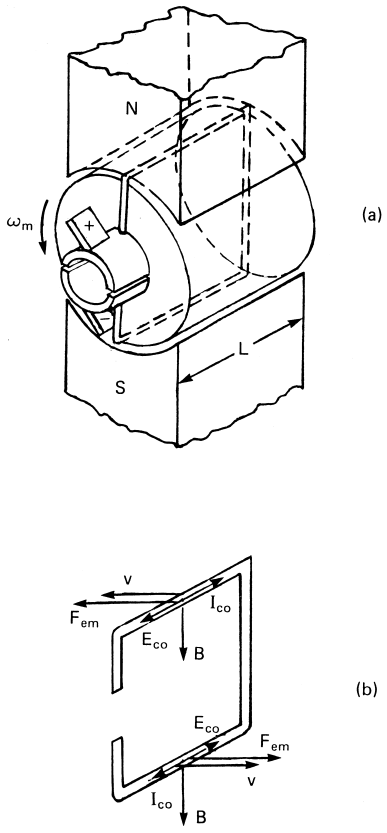


FIGURE 3 Isolated coil showing (a) operation of the commutator and (b) orientation of the electromagnetic force F_{em} and electromotive force E_{co}.

$$F_{em} = B I_{co} L \tag{1}$$

in the direction shown in Fig. 3b. Here F_{em} is measured in newtons (N), B in teslas (T), I_{co} in amperes (A), and L in meters (m).

It is seen that all of the conductors above the neutral plane experience a counterclockwise force. All of the conductors below the neutral plane experience a force in the counterclockwise direction as well. If this force is sufficiently large to overcome bearing friction and the resistance to turning presented by the shaft load, the armature will start to turn and will start to accelerate.

As the armature turns, a second physical law comes into play, which states that when a conductor moves in a magnetic field perpendicular to it with linear velocity v, an electromotive force (emf), or voltage, is induced in the conductor of magnitude

$$E_{co} = vBL \tag{2}$$

with the polarity shown in Fig. 3b. Here E_{co} is measured in volts (V) and v in meters per second (m/sec).

In a rotating machine the mechanical variables are the torque T_{em} and the angular velocity ω_m. The torque is related to the force as

$$T_{em} = F_{em} D/2, \tag{3}$$

where $D/2$ is the radius at the position of the conductor. Usually D, measured in meters, is identified as the diameter of the machine bore at the air gap.

The angular and linear velocities are related as

$$\omega_m = \frac{v}{D/2} = 2\pi \frac{\text{rpm}}{60}, \tag{4}$$

where rpm is the number of revolutions per minute.

The electromagnetic torque of the whole winding is the sum of the torques contributed by each one of the z conductors:

$$T_{em} = \sum_z F_{em} \frac{D}{2}. \tag{5}$$

Introducing, then, the current flowing at the terminals of the armature winding,

$$I_a = 2a I_{co}, \tag{6}$$

where $2a$ is the number of parallel branches, and the average lateral displacement between adjacent conductors,

$$\Delta x = \pi D/z, \tag{7}$$

into Eqs. (1) and (5), one obtains

$$T_{em} = \frac{z}{2\pi} \frac{I_a}{2a} L \sum_z B \Delta x. \tag{8}$$

Finally, if one lets

$$L \sum_z B \Delta x = 2p\Phi, \tag{9}$$

where $2p$ is the number of poles in the machine and Φ, expressed in webers (Wb), is the magnetic flux per pole, the torque is given by

$$T_{em} = z \frac{p}{a} \frac{\Phi}{2\pi} I_a = k \Phi I_a, \tag{10}$$

where $k = (zp/a)(1/2\pi)$.

Similarly, by summing up the voltages induced in each conductor, one obtains the total emf of the armature winding:

$$E_a = z \frac{p}{a} \frac{\Phi}{2\pi} \omega_m = k \Phi \omega_m. \tag{11}$$

It is seen that an essential factor in the generation of the electromagnetic torque T_{em} and of the electromotive force E_a is the magnetic flux per pole Φ. As suggested earlier, dc current flowing through the field windings generates the needed magnetic flux. For this matter to be discussed in detail, it is necessary to invoke a third law that relates the flux to the currents flowing in the windings. These give rise to a magnetomotive force (mmf), which is the dual of the emf. As the current in an electric circuit is given by the ratio of emf to resistance, the flux in a magnetic circuit is given by the ratio of the mmf to the equivalent resistance called the reluctance and denoted by $\Re$.

The magnetic circuit, or path of the magnetic flux, is interlinked with the coils that are formed by the conductors. It crosses the air gap between stator and rotor once under a north pole and once again under the south pole, closing in the iron structures behind the stator and rotor conductors (Fig. 1a). The major contribution to the circuit reluctance comes from the reluctance of the air gap. Therefore, the gap should be made as narrow as possible. Iron is used in the rest of the circuit because it exhibits high "magnetic conductivity," called permeability, and hence offers a low reluctance to the magnetic flux. However, this is true only up to a point. Beyond a certain value the flux ceases to increase in proportion to the mmf and tends to level off. This saturation effect is undesirable because it sets an upper limit to the flux and hence to the torque and emf that can be achieved in a given machine. However, at the other extreme, iron exhibits another peculiarity that is used to advantage in some dc machines. As the mmf is reduced to zero, the flux does not vanish but retains a certain value called remanence.

A sketch of the functional dependence of Φ on the mmf, or magnetization curve, is shown in Fig. 4. This characteristic is drawn by adding the mmf drop required to drive the flux Φ through the iron to the mmf needed to overcome the reluctance of the air gap, which is given by the

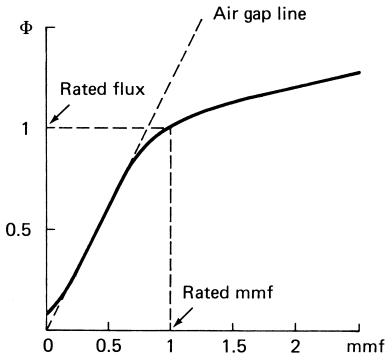


FIGURE 4 Typical magnetization curve (mmf, magnetomotive force).

inverse slope of the air-gap line. In this sketch and in many following ones the scales are expressed in per unit; that is, 1 corresponds to the value of the variable under rated conditions of operation.

C. Governing Equations and Equivalent Circuits

The equation governing the mechanical performance is Newton's law, which in a rotating machine is stated as

$$\Theta\, d\omega_m/dt = T_{em} + T_a - T_{F+W}, \tag{12}$$

where Θ is the moment of inertia measured in kilogram meter square (kg m^2), T_a the torque applied to the shaft by the attached mechanism, and T_{F+W} the opposing torque due to friction and windage inside the machine.

The equations describing the electrical performance can be derived by making use of Kirchhoff's law. In the field circuit the relation between voltage V_f and current I_f is

$$V_f = R_f I_f + L_f\, dI_f/dt, \tag{13}$$

where R_f, the resistance, is measured in ohms (Ω) and L_f, the inductance, is measured in henries (H).

In the armature circuit the relation between voltage V_a and current I_a is

$$V_a = R_a I_a + L_a(dI_a/dt) + E_a + V_{br}, \tag{14}$$

where V_{br} is the voltage drop at the brushes. This drop, which is kept as small as possible (below 1 V), is usually lumped with the resistive drop.

Normally, when conditions of operation change, currents vary much faster than speed. This makes it possible, when one is dealing with mechanical transients, to simplify the analysis by neglecting the inductances, that is, by considering the evolution in time as a succession of electrical quasi-steady states.

The information conveyed by the governing equations can be described conveniently in graphic form by the equivalent circuits shown in Fig. 5. The armature and

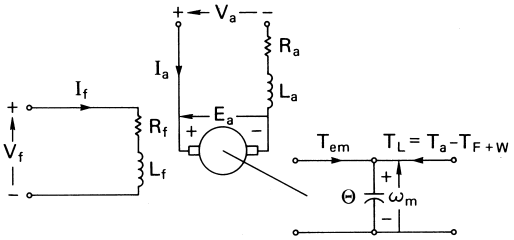


FIGURE 5 Equivalent circuits of dc machine.

mechanical circuits can be combined with the help of Eqs. (10) and (11) into the equivalent circuit of Fig. 6. In the absence of an externally applied torque T_a and if T_{F+W} is neglected, the series combination of an inductance and a capacitance suggests the possibility of an oscillatory motion with natural frequency

$$f_n = \frac{1}{2\pi}\sqrt{\frac{1}{L_a\Theta/(k\Phi)^2}}. \tag{15}$$

The equivalent circuits of Fig. 5 show an arrangement in which the field winding is fed by a separate supply. As a consequence the machine is said to be separately excited. In general, it is adequate to use a single supply. If the field and armature windings are connected in parallel and $V_f = V_a$, the machine is called shunt excited. If the field and armature windings are connected in series and $I_f = I_a$, the machine is series excited. It is also possible to provide the excitation with two field windings, one connected in parallel and one in series with the armature winding. This compound excitation may be cumulative, or positive, and differential, or negative, depending on whether the flux components produced by the two windings tend to add or subtract. The connection of the field affects the performance, as discussed later in Section II.E.

D. Power Balance and Operating Modes

Power relations are obtained by multiplying Eq. (12) by ω_m and Eqs. (13) and (14) by their respective currents. These are

$$(d/dt)\left(\tfrac{1}{2}\Theta\omega_m^2\right) = T_{em}\omega_m + T_a\omega_m - T_{F+W}\omega_m \tag{16}$$

for the mechanical port, that is, for the shaft;

$$V_f I_f = R_f I_f^2 + (d/dt)\left(\tfrac{1}{2}L_f I_f^2\right) \tag{17}$$

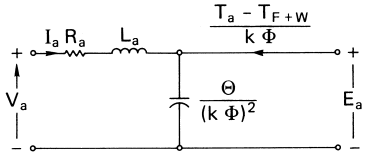


FIGURE 6 Combined electromechanical equivalent circuit.

for the field winding; and when the voltage drops in the brushes are lumped together with the resistive drop,

$$V_a I_a = R_a I_a^2 + (d/dt)\left(\tfrac{1}{2} L_a I_a^2\right) + E_a I_a \qquad (18)$$

for the armature winding.

Observing that, according to Eqs. (10) and (11),

$$T_{em}\,\omega_m = E_a I_a, \qquad (19)$$

one can identify this component coupling the mechanical relation [Eq. (16)] with the armature one [Eq. (18)] as the power converted between mechanical and electrical forms. It is then appropriately denoted by P_{conv} and is measured in watts (W). This term can be eliminated by subtracting Eq. (19) from Eq. (16) to yield a power balance equation:

$$(d/dt)\left(\tfrac{1}{2}\Theta\omega_m^2\right) + (d/dt)\left(\tfrac{1}{2} L_a I_a^2\right) = V_a I_a + T_a \omega_m - T_{F+W}\omega_m - R_a I_a^2, \qquad (20)$$

where the two terms on the left represent the rate of change of energy stored in kinetc form (W_{kin}) and magnetic form (W_{mag}), respectively. The first term to the right is the electric power P_{elec}, the second the mechanical power P_m, the third the power loss P_{F+W}, and the fourth the power dissipated $P_{a\,diss}$.

It should be noted that all power components can be positive or negative, depending on the signs of their time rates of change or constitutive factors, except for the loss terms P_{F+W} and $P_{a\,diss}$, which are always positive. The signs of P_{conv}, P_{elec}, and P_{mech} determine the mode of operation of the machine. Motoring prevails when P_{conv} and P_{elec} are positive and P_{mech} is negative. In this mode $V_a > E_a$ and I_a is positive, so that the machine absorbs electrical power. The torque applied to the shaft opposes the motion, so that the machine delivers mechanical power. Generating occurs when P_{conv} is negative and P_{mech} is positive; that is, the applied torque favors the motion and the machine absorbs mechanical power. On the electrical side, $E_a > V_a$ and I_a is negative, so that P_{elec} is negative and the machine delivers electric power. These are the main operating modes. A third mode is used to bring the machine to a quick stop and is called plugging. In this mode V_a and E_a have opposite signs. Current I_a is positive and limited only by the circuit resistance. The applied torque is in the direction of motion. Therefore, both P_{elec} and P_{mech} are positive. The combined electrical and mechanical power absorbed by the machine is dissipated in the resistance of the armature circuit.

E. Steady-State Performance Characteristics

In steady state the time derivatives vanish and the governing equations can be solved easily to yield the performance characteristics. Making use of Eq. (11) and lumping the voltage drop at the brushes with the ohmic drop in the winding, one can see that Eq. (14) gives

$$I_a = (V_a - k\Phi\omega_m)/R_a. \qquad (21)$$

Equation (10) then gives the electromagnetic torque as

$$T_{em} = k\Phi\frac{V_a}{R_a} - \frac{(k\Phi)^2}{R_a}\omega_m. \qquad (22)$$

It follows that when both Φ and V_a are maintained constant, the torque–speed characteristics are straight lines (Fig. 7). The dashed portions correspond to starting conditions, the solid ones to normal operating conditions. Such a situation usually prevails with separately excited and shunt motors. In a series-excited machine, instead, the flux varies with the armature current. Neglecting iron remanence and saturation effects, one may assume

$$\Phi = k_f I_a, \qquad (23)$$

so that according to Eq. (14),

$$I_a = V_a/(R_a + k k_f \omega_m),$$

and according to Eq. (10),

$$T_{em} = k k_f I_a^2 = k k_f \left(\frac{V_a}{R_a + k k_f \omega_m}\right)^2. \qquad (24)$$

As can be seen by comparing Fig. 8 for the series motor with Fig. 7 for the separately excited and shunt motors, the former exhibits a much stronger dependence of T_{em} on

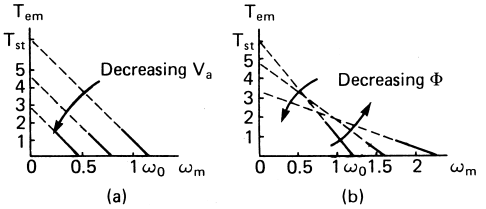


FIGURE 7 Torque–speed characteristics of separately excited and shunt-excited dc motors. (a) V_a as constant parameter; (b) Φ as constant parameter.

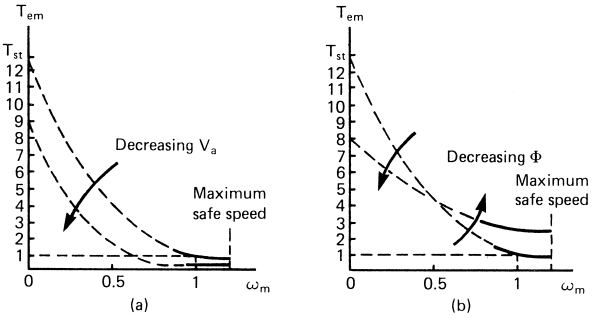


FIGURE 8 Torque–speed characteristics of series-connected dc motor. (a) V_a as constant parameter; (b) Φ as constant parameter.

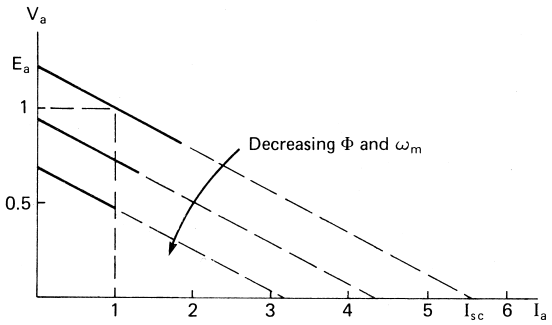


FIGURE 9 Voltage–current characteristics of separately excited dc generators.

ω_m. In particular, the strong excitation at starting ($\omega_m = 0$) provides a high torque, while dangerously high speeds can be reached when the load is removed ($T_a = 0$). These features are used to advantage in electric traction applications in which high accelerations are needed at starting.

Motors with positive compound excitation exhibit torque–speed characteristics that are intermediate between those of Figs. 7 and 8. With negative compound excitation, instead, a practically constant speed independent of load can be attained. An index of the dependence of the speed on torque is given by the speed regulation defined as the ratio of the difference between the no-load speed ω_{NL} and full-load speed ω_{FL} to the full-load speed, or

$$\text{Reg} = (\omega_{NL} - \omega_{FL})/\omega_{FL}. \quad (25)$$

As mentioned in the Introduction (Section I), the great advantage of dc motors is that their speed can be easily adjusted. As seen in Figs. 7 and 8 this can be achieved by controlling either V_a or Φ. Voltage control involves a larger power and therefore is more expensive. However, in constrast with field control, it affords control over the whole speed range.

The dc machine operates as a generator when the current changes direction. This occurs when the electromotive force E_a exceeds the armature terminal voltage V_a. The characteristics of interest in the case of generators are the V–I characteristics. As can be seen from Eq. (21), in the case of separately excited generators these are straight lines (Fig. 9).

The situation is more complicated in the case of shunt-excited generators, because here the field current is derived from the armature terminals and therefore the flux becomes a function of the terminal voltage. If the relation between flux and mmf were linear, such as the air-gap line of Fig. 4, then by introducing $V_f = V_a$ and $I_f = -I_a$ into Eqs. (13) and (14), one would obtain

$$(L_a + L_f)(dI_a/dt) - (kk_f\omega_m - R_a - R_f)I_a = 0 \quad (26)$$

as the equation for the unloaded generator. Accordingly, I_a would rise or decay exponentially depending on whether $kk_f\omega_m > R_a + R_f$. No steady state in the excitation of the machine could be achieved because of the feedback dependence of the field current on the armature voltage. As always happens with self-excitation, one can rely on the effect of nonlinearities, in this case iron saturation, to limit the output. However, in this case the V–I characteristics of the machine must be obtained graphically (Figs. 10a and 10b).

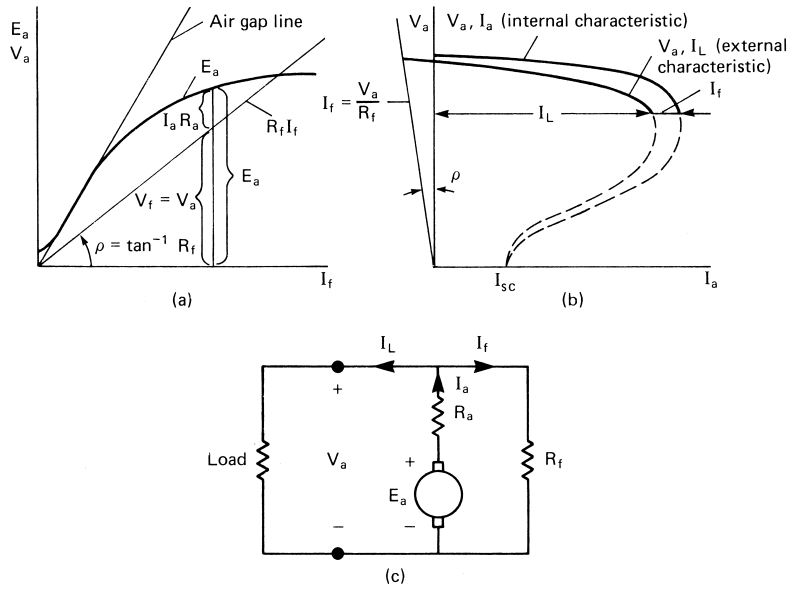


FIGURE 10 Voltage–current characteristics of shunt-excited generator. (a) Field circuit; (b) load circuit; (c) schematic diagram of generator.

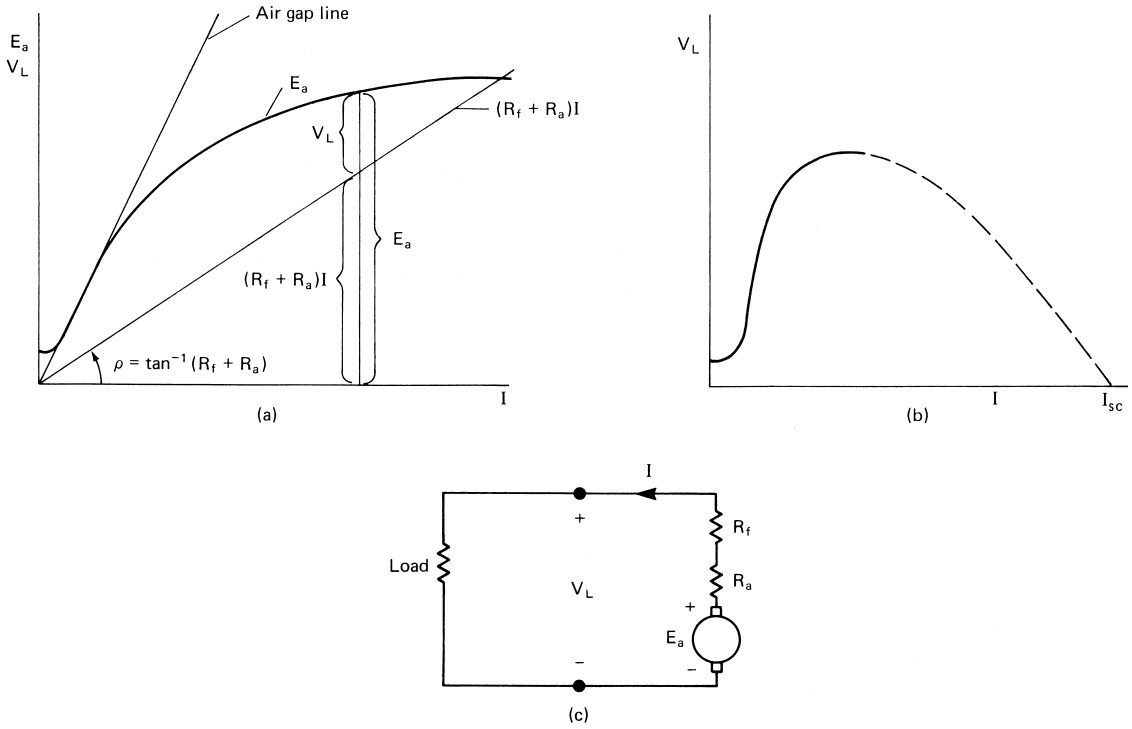


FIGURE 11 Voltage–current characteristics of series-excited generator. (a) Field excitation; (b) load circuit; (c) schematic diagram of generator.

First, the characteristics of field excitation E_a versus I_f and V_f versus I_f are drawn. Since $V_f = V_a$, their difference is $R_a I_a$ (Fig. 10a). Then, observing that this difference is proportional to I_a and that the distance measured along the field resistance line is proportional to V_a, one can draw the V_a–I_a (internal) characteristic of Fig. 10b. Finally, the external characteristic, which gives the armature voltage V_a as a function of the load current I_L, is obtained by subtracting I_f from I_a.

The V–I characteristics of the series-connected generator can be drawn in a similar way by letting $I_a = I_f = I_L = I$. The field excitation relations are shown in Fig. 11a. Then, observing that the voltage in the load V_L is obtained by the difference between the magnetization characteristic E_a and the resistive drop line and that the distances measured along this line are proportional to I, one can draw Fig. 11b.

The fundamental difference in performance between the shunt- and series-excited generators can be detected by comparing Figs. 10 and 11. Whereas the shunt generator can be operated in a region in which the load voltage is only slightly dependent on the load current, the series generator can maintain the load current practically constant over a wide range of load voltages. These voltage–source and current–source features can be accentuated by compounding the excitation with the appropriate polarity. A constant voltage supply is desirable in the common case in which the electrical loads are connected in parallel. When, instead, the loads are connected in series, a constant current is preferred. An important application is the supply of an electric arc of variable length, as in the case of welding.

Direct current generators are gradually being replaced by alternators connected to solid-state rectifiers, because the latter require less maintenance. Conversely, direct current motors are still widely used, especially in control systems. For these applications the transient performance is important.

F. Transient Performance

For the purpose of system analysis the dc machine can be viewed as the three-terminal-pair network shown in Fig. 12. The equations governing its performance [Eqs. (12)–(14)] form a nonlinear set. When one is

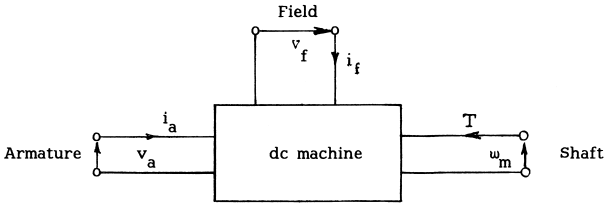


FIGURE 12 Block representation of dc machine.

studying such problems as small-system stability of feedback control systems, it is expedient to linearize the equations around the steady-state operating point. Denoting the incremental variables by lowercase letters and the differential operator d/dt by p, and letting $T_{F+W} = \alpha\omega_m$, one obtains the block diagrams of Fig. 13. It is seen from Fig. 13a that the field voltage v_f controls the speed ω_m through two cascaded time constants: the field time constant $\tau_f = L_f/R_f$ and the mechanical time constant $\tau_m = \Theta/\alpha$. The applied shaft torque T_a affects the speed ω_m through the single time constant τ_m.

In the case of armature control (Fig. 13b) the armature voltage v_a controls the speed through two cascaded time constants with negative feedback around them. The time constant of the armature $\tau_a = L_a/R_a$ is much smaller than τ_f and τ_m and is often neglected. The shaft torque affects the speed through one time constant, τ_m, with the other time constant in a negative feedback loop. In many problems the motor can be assumed to be driving a pure inertia load ($\alpha = 0$). In this case Fig. 13b yields the following transfer functions:

$$\left.\frac{i_a}{v_a}\right|_{T_a=0} = \frac{1/R_a}{1 + \tau_a p + 1/\tau_\Theta p} \tag{27}$$

and

$$\left.\frac{\omega_m}{v_a}\right|_{T_a=0} = \frac{1/kk_f I_f}{1 + \tau_\Theta p + \tau_\Theta \tau_a p^2}, \tag{28}$$

where

$$\tau_\Theta = \Theta R_a/(kk_f I_f)^2.$$

As should be expected Eq. (27) is consistent with the equivalent circuit of Fig. 6 and with Eq. (15).

In the case of field control of a separately excited generator (Fig. 13c) the field voltage controls the load current through two cascaded time constants.

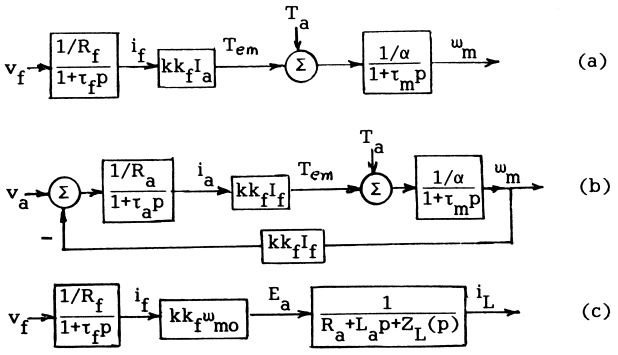


FIGURE 13 Block diagrams of dc machine. (a) Field control of motor speed; (b) armature control of motor speed; (c) field control of load current in separately excited generator.

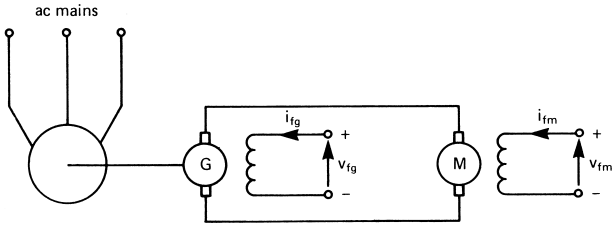


FIGURE 14 Ward-Leonard method of speed control.

Armature speed control of a motor (Fig. 13b) is often accomplished by means of field control of a dedicated generator (Fig. 13c) with the arrangement known as a Ward-Leonard system, shown in Fig. 14. The generator is usually driven by an induction motor. Despite the need for three machines instead of one, this method of speed control is very popular in industrial drives because of its smoothness and versatility.

G. Armature Reaction and Commutation

One may have wondered why in the preceding discussion the magnetic flux Θ was associated exclusively with the current I_f flowing in the field winding. Certainly the current I_a flowing through the armature winding must contribute to the excitation mmf and therefore should be taken into account in the evaluation of Φ. The reason the contribution of the armature, or, as it is called, armature reaction, could be ignored is that it was assumed implicitly that the current changes direction in the coils when the emf induced in their conductors is zero. According to Eq. (2), this means that $B = 0$, and if the field current is considered to be the only source of excitation, B vanishes in the plane equally spaced between adjacent north and south poles (Fig. 15a). In a two-dimensional representation this neutral position is identified as the quadrature axis, while the axis of the poles is called the direct axis.

One observes that all conductors above the brushes carry currents in the direction into the paper, and all conductors below the brushes carry current in the opposite direction. One can therefore imagine that the armature current flows in a fictitious coil, the axis of which is aligned with the quadrature axis. The corresponding mmf, also called armature reaction, drives a flux Φ_a in the direction opposite to that of the quadrature axis. However, the armature mmf faces the interpolar space, which interposes a large reluctance to the passage of the magnetic flux. As a consequence, for equal field and armature mmf's Φ_a is much smaller than Φ_f. Only a few lines of force close through the pole faces, which strengthens the field at the leading pole tip and weakens it at the trailing one. These

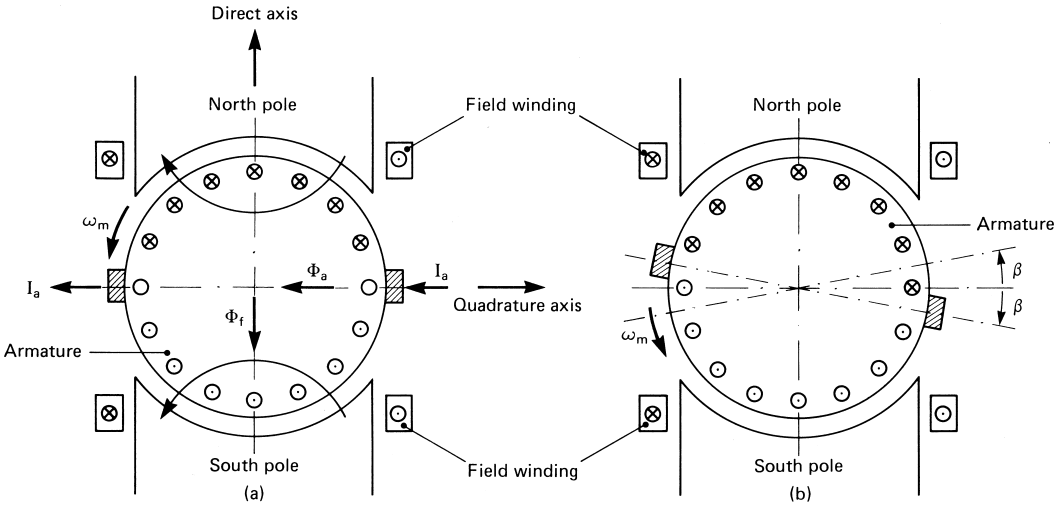


FIGURE 15 Field and armature reaction excitations in dc motor. (a) Brushes in the mechanical neutral plane; (b) brushes shifted to facilitate commutation (⊗, current into paper; ⊙, current out of paper).

local effects, however, nearly compensate for each other, and the flux Φ per pole and, therefore, the overall performance of the machine are unchanged, except when the pole tips are highly saturated.

Armature reaction, however, has an undesirable effect on the commutation of the current from one direction to the opposite. This process can be understood with the help of Fig. 16, which depicts a machine with single-turn coils. The conductors of the upper layer are drawn as solid lines, while those of the lower layer are dashed lines. Each section of the commutator joins one upper and one lower conductor. It should be noted that the strong asymmetry in the

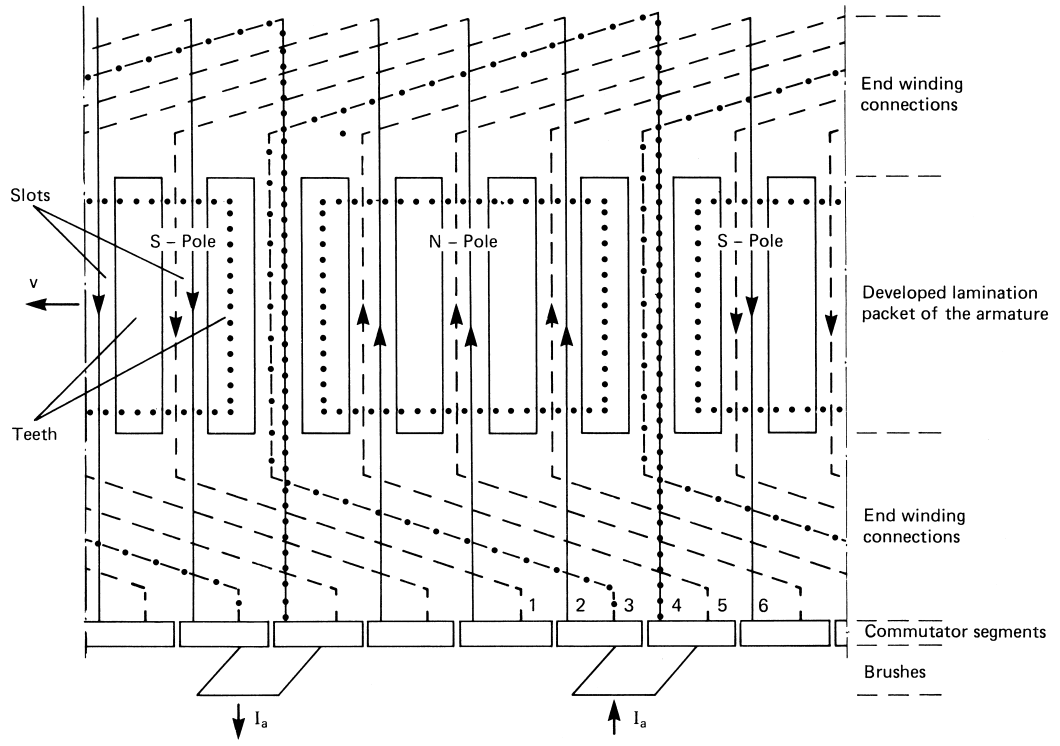


FIGURE 16 Distribution of the armature current with brushes in the neutral plane (—, conductors of the upper layer; – – –, conductors of the lower layer; •–•–, –·–, coils undergoing commutation).

end winding connection permits one to place the brushes in the neutral plane, as shown in Fig. 1a. This conceptually and structurally desirable position is adopted only in small machines wound with thin wires. In large machines the conductors consist of thick bars, which are not bent easily. This forces the adoption of symmetric end winding connections with the resultant placement of the brushes in the direct axis, as sketched in Fig. 3a. The direction of the currents corresponds to operation as a motor when the pole faces, shown as dotted rectangles in Fig. 16, are situated above the conductors. The brushes are seen as bridging two commutator segments each. Considering the brush to the right, the coil formed by conductors 3 and 4 is short-circuited and therefore no external current can enter this coil. Instead, the current I_a divides into two parts. One part enters conductor 2, returns through conductor 1, and advances to the left until it is collected by the brush to the left. The other part enters conductor 5, returns through conductor 6, and advances to the right.

Now consider the short-circuited coils, which have been singled out by the addition of black dots in Fig. 16. The motion of the armature with respect to the brushes transfers their conductors from a zone in which the current flows in one direction into a zone in which it flows in the opposite direction. For instance, conductor 4 of Fig. 16 has just left a south pole zone and is entering a north pole zone. Its current must change from a downward to an upward direction. However, in an inductive circuit such as the short-circuited coil, the current cannot die instantaneously and continues to circulate from conductor 4, through the brush, into conductor 3. When further armature motion opens the contact through the brush, the current must create a conducting path through the air, that is, an electric arc. Such arcs may pit the commutator segments and damage the insulation between adjacent segments. The situation is aggravated by the armature reaction. As can be seen with the help of Figs. 3b and 15a, the conductors leaving the south pole in the lower right corner are exposed to an armature B field, which induces an emf in the direction out of the paper. This emf promotes the flow of current in this direction and therefore opposes its change of direction.

So that the effect of armature reaction can be neutralized and reversal of the current can be accelerated, the conductors undergoing commutation must be exposed to a field that induces an emf in the direction opposed to the flow of the current. In motors this can be accomplished by shifting the brushes against the direction of motion, so that they approach the training end of the poles (Fig. 15b). In a generator the brushes must be shifted forward instead. There is, however, a price paid when commutation is improved by this method. With the shift in the brush axis the armature reaction acquires a component along the direct axis that opposes the field mmf. As indicated in Fig. 15b by the drawing of an axis shifted forward by the same angle β as the brush axis, the armature winding now can be viewed as consisting of two coils: one in quadrature with the field coil as before and one aligned with it but carrying current in the opposite direction. It appears that brush shifting introduces a differential compounding effect. Since brush shifting is not feasible in most situations, arc-free commutation is achieved by another means. In the spaces midway between the field poles (i.e., in the quadrature axis), additional narrow poles, appropriately called interpoles, are placed. These poles are excited by coils connected in series with the armature winding with such a polarity as to oppose the armature reaction. In large machines additional compensating windings are embedded in faces of the main poles.

H. Special Machines

There are a variety of commutator machines, each particularly suited for a specific application. Only two types will be mentioned here: the machine with permanent magnet (PM) excitation and the universal motor.

It is well known that the magnetic field excited by a dc current is indistinguishable from that excited by a permanent magnet. The main advantage of the latter is that it does not require expenditure of energy in, the ohmic losses. When field excitation by means of PMs is applied to dc machines, the main drawback is that it is not easily controlled. Other disadvantages such as weight and cost are gradually being eliminated with the development of better and less expensive materials.

The universal motor has widespread application in household appliances. It is a commutator machine with series-connected excitation. It is called universal because it can operate with ac as well as dc supplies. The reason is that, according to Eq. (24), its torque is proportional to the square of the current and therefore is not affected by the alternation in the sign of I_a. It should be noted, however, that when this motor is operated from ac supplies the impedance of the armature winding includes a significant reactive component and the expression on the right-hand side of Eq. (24) should be corrected accordingly. The practical result is that for the same current and therefore, effective torque, the motor requires a much higher voltage when it is supplied by an ac source than when it is supplied by a dc source.

III. SYNCHRONOUS MACHINE

Almost all electric power generation is accomplished by mechanically driving a synchronous machine (or alternator) and taking electric power from stationary windings

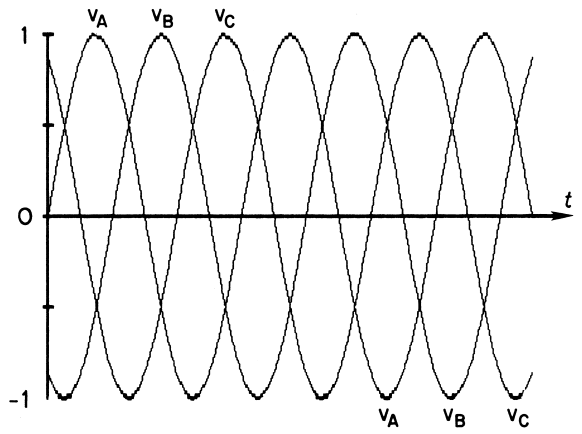


FIGURE 17 Voltages to ground appearing at the three terminals of a three-phase synchronous machine.

in the form of three-phase ac. Frequency is 60 Hz in the United States and Canada, and usually 50 Hz elsewhere. The voltages to ground that appear at the three terminals of the machine, v_A, v_B, v_C, are of equal amplitude, vary sinusoidally with time, and are displaced from one another by one-third of a cycle (Fig. 17).

Electricity generated in a large power plant must be transmitted elsewhere, often many miles away. For this to be accomplished economically, much higher voltages are used (e.g., 345 kV between lines) than are present at the generator terminals (~20 kV between terminals or ~11 kV between each terminal and ground). Large transformers accomplish the task of increasing the generated voltage to a level suitable for transmission.

A. Construction

Figure 18 shows a synchronous machine designed to generate power in the 200-MW (megawatt) range in a thermal power plant. The driving steam turbine is on the same shaft. High-pressure, high-temperature steam (~3000 psi, 1000°F in a coal- or oil-fired plant) is directed at the turbine blades, providing the motive power for the rotor, which induces three-phase ac voltages in the stator windings. Such a machine turns at 1800 or 3600 rpm if it generates 60 Hz and 1500 or 3000 rpm if it generates 50 Hz. In general, because of the relation between economical speed and pressure in the steam turbine, the higher generator speeds are used for the high-pressure turbine stages in fossil-fuel plants and the lower-generator speeds for the low-pressure turbine stages in fossil-fuel plants and in nuclear plants.

The rotor is basically a rotating magnet structure with poles of alternating polarity (Fig. 19), created by appropriate windings placed in slots on the silicon steel rotor structure and firmly held there against the action of strong centrifugal forces. Direct currents are fed into the rotor windings by dc voltages often created by dc generators ("exciters") mechanically driven by the turbine also, but through a reducing gear arrangement. Such a rotor is called a round rotor, as opposed to the salient-pole rotor used in hydroelectric plant generators, which are driven by water turbines, turn much more slowly, require many more poles, and have a cross section exhibiting one saliency per pole.

When a separate dc exciter is used, the dc current is supplied to the rotor through two "slip rings" mounted on and concentric with the rotor shaft, but insulated from it, through stationary brushes. Other arrangements for field excitation are also used, one being a separate alternator mounted on the rotor shaft with direct rectification of the ac voltage. This eliminates the need for slip rings.

Even though large generators have efficiencies well in excess of 90%, the energy dissipated (as heat) per second may amount to several megawatts. Means of adequate cooling must therefore be provided. The best coolant is water, which is made to flow inside the hollow copper conductors. In view of the difficulty of maintaining the necessary electrical insulation in the stator and leakproof seals in the rotor, such a cooling method is justified only in sizes above 200 MW. In special applications water or oil outside the conductor is used. Hydrogen is the second best choice as a coolant but, because of its explosive nature, is not used in sizes below ~20 MW. All other machines are cooled by air, and this includes most of the motors.

Synchronous motors are used mainly because of their capacity to improve the power factor of a system. At the lower end of the power range are the clock motors. Altogether, synchronous machines cover a power range that extends over 12 orders of magnitude.

B. Principles of Operation

In small machines, the roles of stator and rotor are inverted: The field winding is placed on the stator and the ac winding, also called the armature, on the rotor. This helps to explain the relationship with the dc machine discussed in Section II. A sketch of the connections of a single-turn coil is shown in Fig. 20. Comparison of this figure with Fig. 3 shows that the synchronous and dc machines appear to differ only in the way the armature coils are connected to the external circuits. In the dc machine the connections are switched periodically by the commutator, whereas in the synchronous machine the slip rings maintain constant connections.

It is convenient to explore the operation of the synchronous generator in terms of such an inverted machine. It is assumed that the magnetic flux, created by the stationary field winding, is sinusoidally distributed along the periphery of the gap, so that this density varies with the

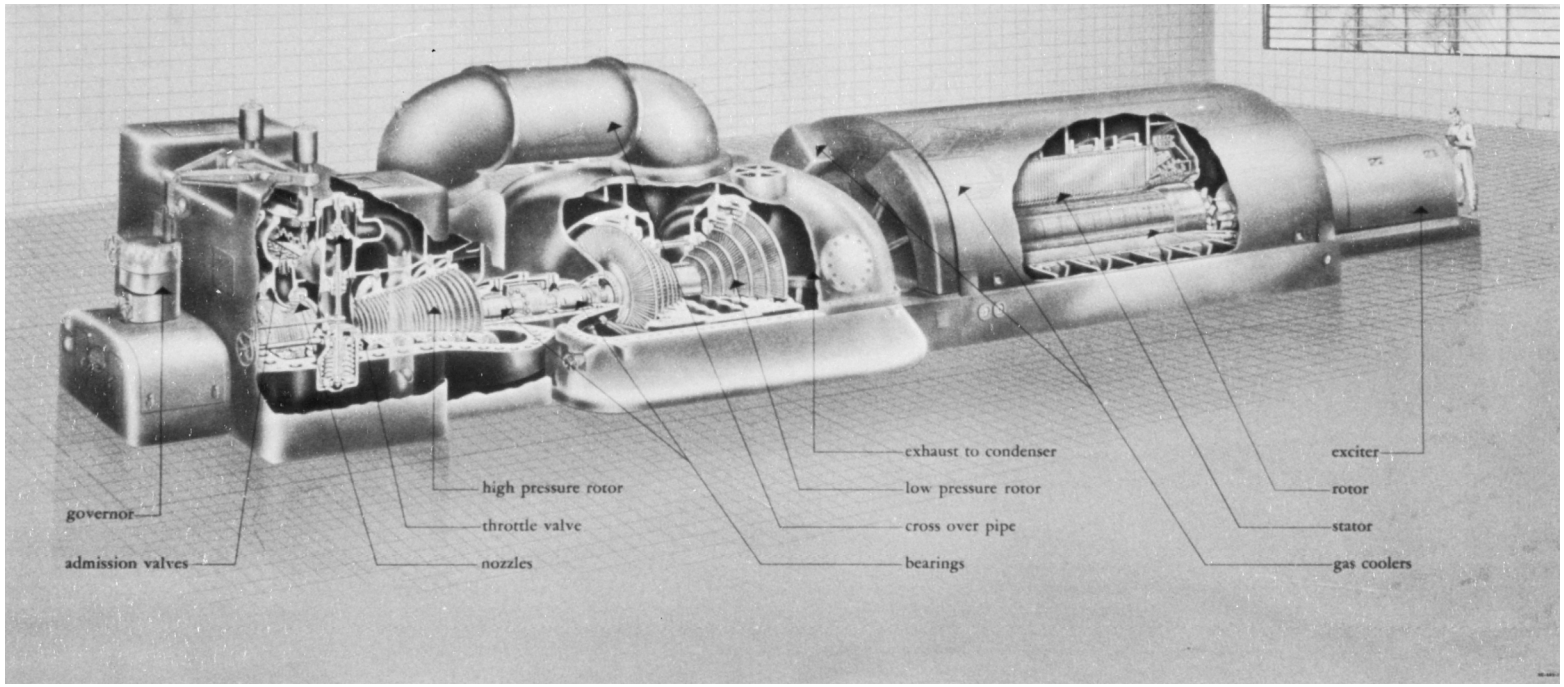


FIGURE 18 Sketch of 3600-rpm steam turbine generator unit.

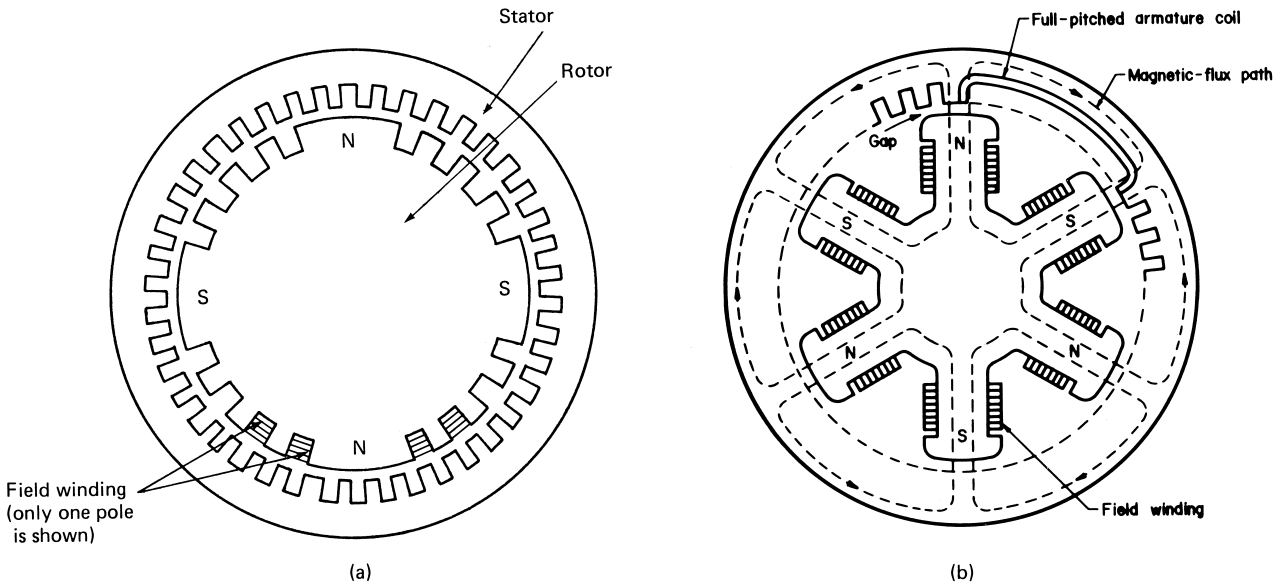


FIGURE 19 Types of rotors in a synchronous machine. (a) Round rotor; (b) salient-pole rotor.

angle ϑ (measured counterclockwise from the direct axis; i.e., the center line of a field pole) as

$$B(\vartheta) = B_{\mathrm{m}} \cos p\vartheta, \tag{29}$$

where p is the number of pole pairs. If the conductors rotate at constant angular velocity ω_{s}, it follows that their position varies with time as

$$\vartheta = \omega_{\mathrm{s}} t + \vartheta_0. \tag{30}$$

Making use of Eq. (2), one finds that the emf induced in each conductor becomes

$$E_{\mathrm{co}} = vLB_{\mathrm{m}} \cos p\vartheta = vLB_{\mathrm{m}} \cos(p\omega_{\mathrm{s}} t + p\vartheta_0). \tag{31}$$

It appears that the emf varies in time with an electrical radian frequency

$$\omega = p\omega_{\mathrm{s}} \tag{32}$$

and that the frequency is

$$f = \frac{\omega}{2\pi} = p\frac{\omega_{\mathrm{s}}}{2\pi} = p\frac{\mathrm{rpm}}{60}. \tag{33}$$

This rigid relation between electrical frequency and mechanical speed is called the synchronous tie and gives the synchronous machine its name. Equation (31) also establishes the relation between the location ϑ_0 of the conductors on the rotor and the phase of their emf. It follows that the three-phase voltages of Fig. 17 are supplied by conductors, which are displaced on the rotor by an angle $2\pi/3p$. Summing up the contributions of all the conductors in the N turns of one of the phase windings, one obtains

$$E_{\mathrm{mph}} = 2N_{\mathrm{eff}} vLB_{\mathrm{m}}, \tag{34}$$

where

$$N_{\mathrm{eff}} = k_{\mathrm{dp}} N \tag{35}$$

and k_{dp} is a winding factor that takes into account the fact that the series-connected conductors are distributed in different slots and, therefore, their emf's are not in phase.

Finally, introducing

$$v = \omega_{\mathrm{s}} D/2 = (\omega/p)(D/2), \tag{36}$$

the average value of B per pole

$$B_{\mathrm{av}} = (2/pi)B_{\mathrm{m}}, \tag{37}$$

and the flux per pole

$$\Phi_{\mathrm{m}} = \left(\frac{2}{\pi} B_{\mathrm{m}}\right)\left(\frac{\pi DL}{2p}\right) = \frac{DL}{p} B_{\mathrm{m}}, \tag{38}$$

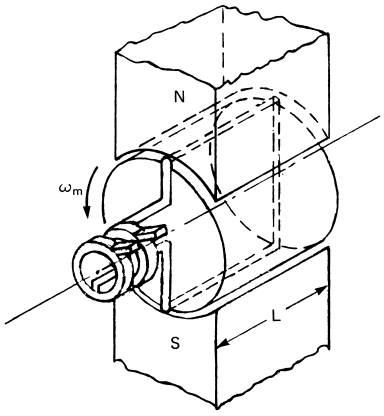


FIGURE 20 Isolated single-turn coil showing connection to slip rings.

one obtains for the rms value of the emf per phase

$$E_{\text{ph}} = \left(1/\sqrt{2}\right)\omega N_{\text{eff}}\Phi_{\text{m}} = 4.44 f N_{\text{eff}}\Phi_{\text{m}}. \tag{39}$$

Except for the substitution of N_{eff} for N, this result is identical with Eq. (9) of Transformers. It follows that the emf can also be viewed as resulting from the rate of change with time of the magnetic flux interlinked with the coils.

When the terminals of the phase windings are connected to a balanced load, the currents that flow in the coils create an armature reaction flux Φ_{a}, as was the case with the dc machine. This flux is sinusoidally distributed along the periphery of the gap. It represents a wave that rotates with phase velocity $-\omega_{\text{s}}$ with respect to the armature conductors, as can be inferred from Eq. (31), but that is at standstill with respect to the poles. Therefore, the flux that induces the emf E_{ph} when the machine is loaded is the resultant of the field and armature fluxes.

At this point it is expedient to introduce complex or phasor notation to represent sinusoidally distributed quantities and write

$$\bar{\Phi} = \bar{\Phi}_{\text{f}} + \bar{\Phi}_{\text{a}}. \tag{40}$$

As explained in Section II.B, the flux and the excitation mmf are related by the reluctance $\Re$. Assuming linearity and referring at first to a machine with uniform air gap, such as a turbogenerator, one can write

$$\bar{E}_{\text{ph}} = j\omega N_{\text{eff}}\bar{\Phi} = j\omega N_{\text{eff}}(\Re N_{\text{eff}}\bar{I}_{\mu}), \tag{41}$$

where $\bar{I}_{\mu}$ is the magnetizing current. Then, defining a mutual or magnetizing inductance

$$M = N_{\text{eff}}^2\Re, \tag{42}$$

one arrives at the relation

$$\bar{E}_{\text{ph}} = j\omega M\bar{I}_{\mu} = j\omega M(\bar{I}'_{\text{f}} + \bar{I}_{\text{a}}), \tag{43}$$

where the prime indicates that the quantity, in this case the field current, is referred to a similar quantity in the other winding on the basis of equality of mmf's and emf's.

With this in mind and introducing the impedance of the armature, which consists of a resistance R_{a} and a leakage inductance L_{a}, one can write the equation governing the performance of a synchronous machine with uniform gap:

$$\begin{aligned}\bar{V}_{\text{ph}} &= (R_{\text{a}} + j\omega L_{\text{a}})\bar{I}_{\text{a}} + \bar{E}_{\text{ph}} \\ &= R_{\text{a}}\bar{I}_{\text{a}} + j\omega(M + L_{\text{a}})\bar{I}_{\text{a}} + j\omega M\bar{I}'_{\text{f}}.\end{aligned} \tag{44}$$

In the case of a generator the armature current is usually defined with the opposite sign and the equation is written

$$\bar{V}_{\text{ph}} = \bar{V}_{\text{a}} = \bar{E}'_{\text{f}} - (R_{\text{a}} + j\omega L_{\text{s}})\bar{I}_{\text{a}}, \tag{45}$$

where

$$\bar{E}'_{\text{f}} = j\omega M\bar{I}'_{\text{f}}, \tag{46}$$

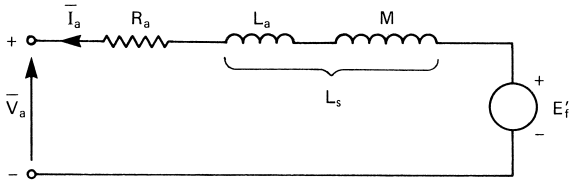


FIGURE 21 Equivalent circuit of generator with uniform air gap.

and

$$L_{\text{s}} = M + L_{\text{a}} \tag{47}$$

is the self-inductance or synchronous inductance of a phase winding. This relation can be represented by the equivalent circuit of Fig. 21.

Considered next is the case of a machine with salient poles. Since the reluctance in the direct axis now differs from that of the quadrature axis, one must define an inductance M_{d} for the direct axis and an inductance M_{q} for the quadrature axis and divide the armature reaction mmf into two components, so that

$$\bar{I}_{\text{a}} = \bar{I}_{\text{d}} + \bar{I}_{\text{q}}. \tag{48}$$

Equation (44) thus becomes

$$\begin{aligned}\bar{V}_{\text{a}} = {} & R_{\text{a}}\bar{I}_{\text{a}} + j\omega(M_{\text{d}} + L_{\text{a}})\bar{I}_{\text{d}} \\ & + j\omega(M_{\text{q}} + L_{\text{a}})\bar{I}_{\text{q}} + j\omega M_{\text{d}}\bar{I}'_{\text{f}}.\end{aligned} \tag{49}$$

If one denotes

$$x_{\text{d}} = \omega(M_{\text{d}} + L_{\text{a}}) \tag{50}$$

and

$$x_{\text{q}} = \omega(M_{\text{q}} + L_{\text{a}}) \tag{51}$$

as the armature reactances in the direct and quadrature axes, respectively, Eq. (49) leads to the phasor diagram of Fig. 22. The equation for the field winding is identical with Eq. (13).

C. Steady-State Performance

The power developed by a machine having m_{a} phases is

$$P = m_{\text{a}}V_{\text{a}}I_{\text{a}}\cos\vartheta, \tag{52}$$

where ϑ is the phase angle between V_{a} and I_{a} and $\cos\vartheta$ is the power factor.

Neglecting R_{a}, which, because of efficiency requirements, is always small, one can derive from the phasor diagram an expression for the power in terms of $\bar{V}_{\text{a}}$ and $\bar{E}'_{\text{f}}$ and the power angle δ between them. This is

$$P = -\frac{m_{\text{a}}}{x_{\text{d}}}V_{\text{a}}E'_{\text{f}}\sin\delta - \frac{m_{\text{a}}}{x_{\text{d}}x_{\text{q}}}(x_{\text{d}} - x_{\text{q}})V_{\text{a}}^2\frac{\sin 2\delta}{2}, \tag{53}$$

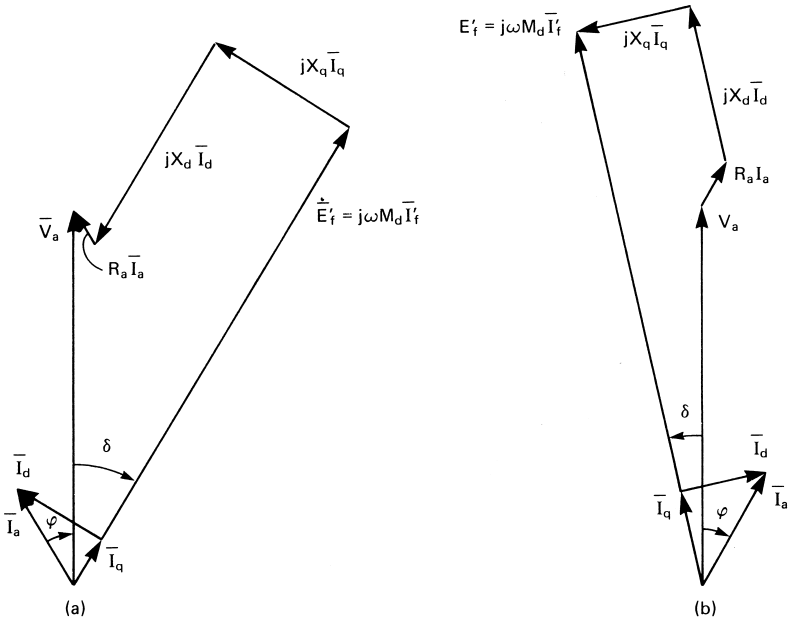


FIGURE 22 Phasor diagram of synchronous machine with salient poles. (a) Motor; (b) generator.

with δ negative in motors and positive in generators. Since the synchronous machine operates only at constant speed ω_s, its torque

$$T_{em} = P/\omega_s \tag{54}$$

is proportional to P. It appears that salient-pole machines, in which $M_d \neq M_q$, can develop a torque even when there is no field winding ($E'_f = 0$). This feature is utilized in the so-called reluctance motors, which find application in record players and electric clocks. In the presence of field excitation, however, this torque component is small and will be neglected.

The synchronous machine has two electrical ports and one mechanical port. In Eq. (53) they are characterized by the armature and field excitation voltages V_a and E'_f and by the power angle δ. Although the latter was defined in electrical terms as the phase difference between $\bar{V}_a$ and $\bar{E}'_f$, it has a mechanical significance, because $\bar{E}'_f$ is induced by a magnetic field distribution rigidly attached to the field structure and $\bar{V}_a$ by a wave rotating at speed $-\omega_s$ with respect to the armature structure. It follows that δ represents the cyclic position of the shaft and can be measured by using a stroboscopic light fed by V_a.

Usually, synchronous machines are connected in parallel in large systems that maintain V_a constant. The voltage E'_f supplied by the exciter can also be considered constant, because even though it is subject to control, it cannot change rapidly due to the large inductance of the field circuit. It is therefore of interest to study the performance of the machine at constant V_a and E'_f. The power and torque characteristics of a round rotor machine have, then, the sinusoidal trend shown in Fig. 23. The most striking feature is that the torques have maximum, or pull-out, values T_{po} in both the motoring and generating regimes. If the torque applied to the shaft exceeds these values, the machine starts to decelerate and hence loses synchronism between the mechanical speed and that of the electromagnetic wave or, as it is said, falls out of step. As δ changes with time, the electromagnetic torque alternates in sign with sometimes dire consequences. The situation is best studied by drawing the locus of the armature current with the help of Eq. (44) and by neglecting R_a. This is a circle of radius $(M/L_s)\bar{I}'_f$ centered at the tip of a modified magnetizing current

$$\bar{I}^{(m)}_{\mu} = \bar{V}_a/j\omega L_s, \tag{55}$$

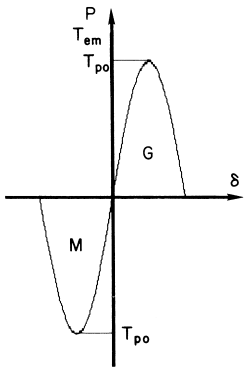


FIGURE 23 Power and torque as a function of the power angle δ (M, motoring regime; G, generating regime).

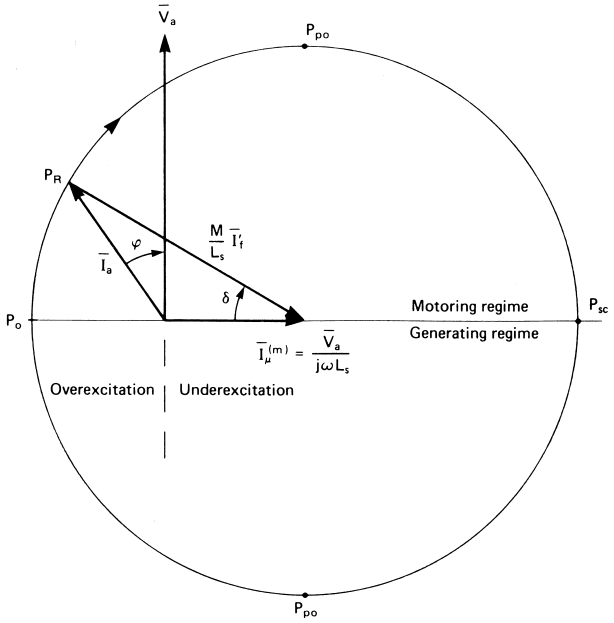


FIGURE 24 Circle diagram of synchronous machine.

as shown in Fig. 24. Assuming that initially the machine was operating as a motor at the rated point P_R, an increase in the mechanical torque applied to the shaft tends to slow the rotor, thus increasing the angle δ in the negative direction. This brings about an increase in the in-phase component of $\bar{I}_a$ and the desired increase in the electromagnetic torque T_{em}. A limit of *steady-state* stability is reached, however, at point P_{po}, beyond which the in-phase component of $\bar{I}_a$ and with it T_{em} decrease. This causes the rotor to decelerate at an increasing rate until, at point P_{sc}, T_{em} changes sign and the machine switches into the generating mode. Unless the armature terminals are rapidly disconnected from the supply, the machine will be damaged, thermally because of the large currents and mechanically because of the rapid alternations in torque. A similar situation prevails when the machine operates initially as a generator. It follows that the rated point of operation P_R is chosen to allow for a substantial overload capability (OC). From Fig. 24 it is seen that this is

$$\text{OC} = \frac{(M/L_s)I'_f}{I_a \cos\phi_R}. \tag{56}$$

The stability problem is compounded when the dynamic transient is taken into account. The combination of the inertia of the rotating masses with the inductance of the armature circuit produces a resonance effect, as discussed earlier in Section II.C. Under certain circumstances the oscillations of the rotor around its steady-state angle δ may grow as a result of a negative damping introduced by the armature resistance R_a.

Whereas stability is the most critical aspect in the operation of a synchronous machine, a very positive feature is that it not only can draw its own flux excitation current from the field side, but also, when overexcited, can operate with the armature current leading the voltage. As a result, whether it absorbs real power as a motor or delivers it as a generator, it can supply reactive volt amperes (VARs) to an inductive load. In particular, at point P_o of Fig. 24, I_a leads V_a by 90°; the real power is zero and the machine operates as a capacitor or, as it is called, a synchronous condenser. The role played by the field current I'_f can be understood by drawing the locus of $\bar{I}_a$ when the power and electromagnetic torque are maintained constant. With constant V_a this means that $I_a \cos\phi$, the in-phase component of armature current, is maintained constant. The locus of $\bar{I}_a$ is then a horizontal straight line, as shown in Fig. 25 for a motor. It appears that the field excitation current cannot be reduced below a value

$$I'_f = (L_s/M)I_a \cos\phi, \tag{57}$$

which corresponds to the pull-out point P_{po}. At this point the machine draws all its magnetizing current requirements from the ac side. The excitation characteristics of

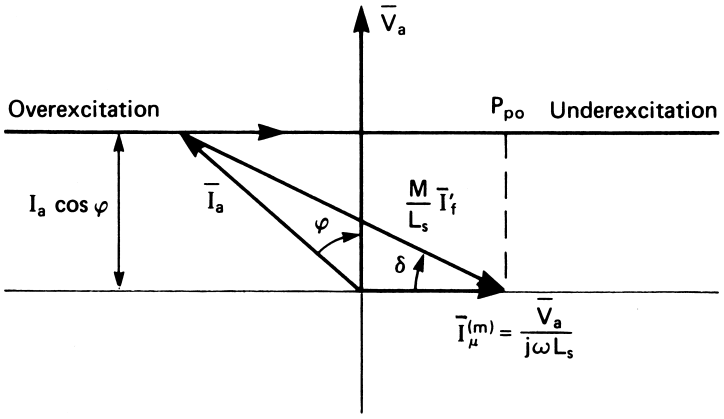


FIGURE 25 Locus of the armature current with constant power and torque.

the machines are often plotted in terms of I_a as a function of I'_f with the power P as a parameter. These curves can be derived from Fig. 25 and are known as V curves because of their shape.

The synchronous machine can also operate when singly excited on the field side with the armature terminals connected to a passive load, but only as a generator. In this case it is important that with constant field excitation the output voltage varies only slightly with load. The index of performance is then the voltage regulation between no-load (NL) and rated (R) conditions, or

$$\text{Reg} = \frac{|\bar{V}_{aNL}| - |\bar{V}_{aR}|}{|V_{aR}|} = \frac{|\bar{E}'_f|}{|V_{aR}|} - 1. \qquad (58)$$

If the parameters of the machine are known, the regulation can be calculated by making use of Eq. (45). If not, the parameters can be evaluated experimentally. First, the open-circuit characteristic E'_f versus the dc field current I_f is determined (Fig. 26). Then, the value of I_f required to drive the current I_a under short-circuit conditions at rated speed (segment OB, Fig. 26) is measured. Letting $\bar{V}_a = 0$ in Eq. (45), one finds

$$(\bar{I}'_f)_{sc} = \bar{I}_a + \frac{R_a + j\omega L_a}{j\omega M}\bar{I}_a \sim \bar{I}_a + \frac{L_a}{M}\bar{I}_a. \qquad (59)$$

It follows that the field mmf consists of a component that balances the armature mmf and a component that is required to drive I_a through the armature impedance. The impedance drop, given by the segment CA in Fig. 26, must now be found. It is noted that if a variable inductance replaces the short circuit and I_a is kept fixed, CA is also constant. Therefore, triangle ABC (named after Potier) traces the voltage characteristic BB'' because it moves parallel to itself as I_f increases, upward along the open-circuit characteristic. Once into saturation, a horizontal line is drawn through B'', and the distance $O''B'' = OB$ is marked off. Through O'', a line is drawn parallel to the air-gap line. The point of intersection C'' with the open-circuit characteristic determines the Potier triangle. The leakage reactance ωL_a is the ratio $C''A''/I_a$. The synchronous reactance x_s in per unit of the base impedance of the machine V_{aR}/I_{aR} is given by the inverse of the short-circuit ratio, that is, the ratio of the field current $(I_{fR})_{oc}$ required to generate the rated voltage V_{aR} at open circuit to the field current $(I_{fR})_{sc}$ required to drive the rated current I_{aR} under short-circuit conditions. This can be seen by letting $V_a = V_{aR}$ and $I_a = 0$ in Eq. (45) and $R_a = 0$ and $I_a = I_{aR}$ in Eq. (59). Graphically an *unsaturated* per-unit value of x_s is obtained from Fig. 26 as the ratio of $OB' = (I_{fR})_{oc}$ to OD, which represents $(I_{fR})_{oc}$ following the air-gap line. A smaller *saturated* value is obtained as OB'/OE, where OE corresponds to $(I_{fR})_{oc}$ in the case that the mmf in the iron is taken into account.

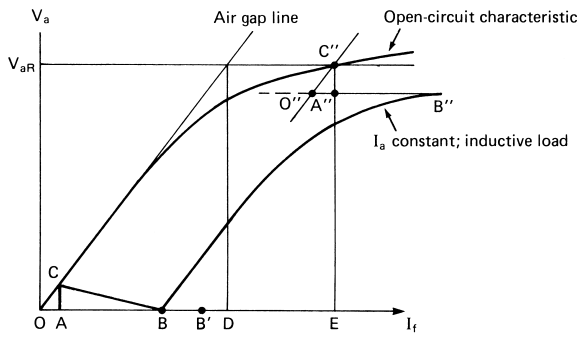


FIGURE 26 Open-circuit and inductive-load characteristics; Potier triangle.

IV. INDUCTION MACHINE

A. Construction and Principle of Operation

The induction (asynchronous) machine, similar in some respects to the synchronous machine and in others to the transformer, consists of two windings. One (the primary) is embedded in the stator lamination slots and the other (the secondary) in the rotor lamination slots. The stator winding is identical with the polyphase armature winding of a synchronous machine. When its terminals are connected to a balanced three-phase supply, the three stator currents create, at any instant of time, in the air gap, as in the synchronous machine, a magnetic field with alternating north and south poles. This field distribution rotates at the synchronous speed, which is given by Eq. (32). This speed depends only on the frequency of the ac supply and the number of poles, the latter being determined by the winding.

If attention is now focused on the rotating field rather than on the currents that create it, it is seen that, in the primary windings, an emf is excited. The emf may be viewed as being induced either by the relative motion of the conductors with respect to the magnetic field according to Eq. (31) or alternatively by the rate of change of magnetic flux interlinked with the coils according to Eq. (39). Since almost all the flux ϕ crosses the air gap and is interlinked also with the secondary winding, an emf is induced also in these conductors, as long as there exists relative motion between them and the rotating field. This is certainly the case when the rotor is at standstill. In this state, the machine is indistinguishable from a three-phase transformer, and if the secondary circuit is closed, currents will start to flow in the rotor conductors. These currents interact with the magnetic field, according to Eq. (1), to give rise to a torque on the rotor, which forces it to start turning in the direction of the field. The machine now operates

as a motor, and if the electromagnetic torque T_{em} it develops is larger than the torque T_a applied by the load on the shaft,the rotor gains speed according to Eq. (12). The motor eventually reaches a steady-state speed, when the developed torque equals the applied torque. The steady-state speed, of necessity, is always somewhat less than the synchronous speed; otherwise, there would be no relative motion between the rotor conductors and the rotating magnetic field, so that there would be no induced rotor emf and no rotor currents to give rise to a turning torque.

An important condition, however, has to be satisfied for the development of the electromagnetic torque: The magnetic field, set up by the induced polyphase rotor currents and rotating with respect to the rotor in its own direction of movement, must be stationary with respect to the synchronously rotating magnetic field set up by the stator currents. Otherwise, the phase angle between I and B would vary cyclically with time, which would cause the torque to vary cyclically with time and yield a zero average torque. This relative standstill condition is certainly satisfied before the rotor has started to turn, because by symmetry it is clear that the emf waves induced in the stator and rotor windings both rotate with synchronous speed ω_s with respect to the conductors. Yet to be proved is that the necessary condition for torque production prevails also when the rotor turns with a nonzero angular velocity ω_m. When it does, its conductors slip past the synchronous wave with relative angular velocity $\omega_s - \omega_m$ or with linear velocity $v_s - v_m$. First, it must be observed that, according to the synchronous tie of Eq. (32), the frequency of the voltages and currents induced in the rotor conductors is

$$\omega_2 = p(\omega_s - \omega_m), \tag{60}$$

or introducing a slip,

$$S = (\omega_s - \omega_m)/\omega_s = (n_s - n)/n_s = (v_s - v_m)/v_s, \tag{61}$$

one obtains

$$\omega_2 = Sp\omega_m = S\omega_1. \tag{62}$$

It will soon be apparent that slip is one of the most important parameters of an induction machine. Making use again of the synchronous tie relation, one finds that a system of currents such as the one flowing in the polyphase rotor winding and alternating with radian frequency ω_2 sets up an mmf wave that rotates with respect to the rotor with angular velocity

$$\omega_{2m} = \omega_2/p. \tag{63}$$

This is exactly the velocity that when added to the rotor velocity ω_m gives a velocity ω_s with respect to the stator conductors, as can be seen by assigning to ω_2 the value given by Eq. (62). It is now seen that the synchronous machine is a special case of the asynchronous one, corresponding to a situation in which the slip and, therefore, the secondary frequency are both zero.

If one would like the machine to operate at variable speed from a constant frequency supply, the secondary winding can be fed by a second source at variable frequency in accordance with the slip and Eq. (62). However, a much simpler method is used to operate the induction motor. The rotor currents that are required to produce the torque can be obtained simply by relying on the induced emf's and by closing the secondary circuit through passive resistances or through a short circuit. The first type of machine is built with a wound rotor and slip rings to connect with the external resistances. In the second the active conductors are bars short-circuited by rings at both ends (Fig. 27). The resulting construction resembles a squirrel cage and therefore goes by that name. The simplicity of this kind of secondary winding, its low cost, its ruggedness, and the absence of maintenance-requiring sliding contacts account for the popularity of the squirrel-cage motor.

The difference in velocity between the primary and secondary conductors with respect to the rotating electromagnetic wave has another consequence. Since the emf's induced in the primary and secondary conductors are proportional to the relative speed v, one can write, according to Eq. (2),

$$(E_{co})_2/(E_{co})_1 = (v_s - v_m)/v_s = S. \tag{64}$$

The vanishing of the secondary emf with zero slip establishes the connection with the synchronous machine,

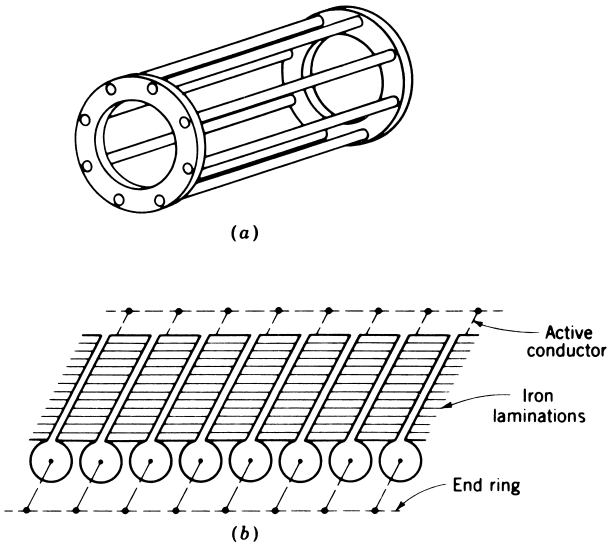


FIGURE 27 Sketch showing (a) conductors of a squirrel-cage rotor, with short-circuiting end rings; (b) a developed view of the laminated iron rotor within the slots of which the conductors of the squirrel-cage winding are normally housed. [Courtesy of Krieger Publishing Company, Melbourne, Florida.]

which requires an external source to drive current in the field winding.

B. Describing Equations

The similarities established above with both the synchronous machine and the transformer suggest the following form for the equations describing the performance of the induction machine. In steady state all variables alternate sinusoidally in time and therefore can be expressed in terms of phasors as

$$\bar{V}_1 = (R_1 + j\omega_1 L_1)\bar{I}_1 + \bar{E}_1 \tag{65}$$

for each phase of primary winding, and

$$\bar{V}_2 = (R_2 + j\omega_2 L_2)\bar{I}_2 + \bar{E}_2 \tag{66}$$

for each phase of the secondary winding.

A third equation relating the emf's to the currents on both sides of the air gap is obtained by expressing the currents on the basis of equal mmf's. This means referring all secondary quantities to a primary having the same number of effective turns and phases. The result is

$$\bar{E}_1 = j\omega_1 L_m(\bar{I}_1 + \bar{I}'_2) \tag{67}$$

or

$$\bar{E}'_2 = j\omega_2 L_m(\bar{I}_1 + \bar{I}'_2) = S\bar{E}_1, \tag{68}$$

where the primes denote quantities referred to the primary. It follows that Eq. (66) divided by S and referred to the primary becomes

$$\bar{V}'_2/S = (R'_2/S + j\omega_1 L'_2)\bar{I}'_2 + \bar{E}_1. \tag{69}$$

Equations (65), (67), and (69) can be represented by the phasor diagram of Fig. 28. In this diagram the magnetizing

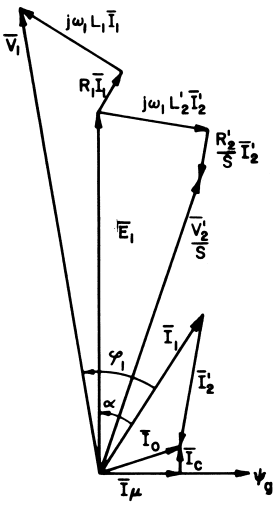


FIGURE 28 Induction motor phasor diagram showing relations among the primary and secondary voltages and currents.

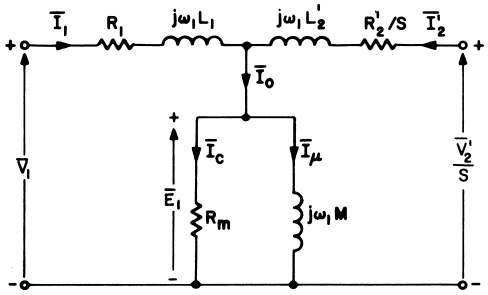


FIGURE 29 Equivalent circuit of induction motor.

current $\bar{I}_\mu$ has been augmented by a component $\bar{I}_c$ in phase with the voltage $\bar{E}_1$ to give an approximate representation to the losses in the iron core due to hysteresis and eddy currents. The result is

$$\bar{I}_0 = \bar{I}_\mu + \bar{I}_c = \bar{I}_1 + \bar{I}'_2. \tag{70}$$

The describing equations can also be reduced to the equivalent circuit of Fig. 29, where the resistance R_m has been added to represent the iron losses.

C. Torque–Slip Characteristic

The performance of the induction machine with short-circuited secondary is best understood by referring to a simplified equivalent circuit in which the primary resistance R_1 is neglected and the field excitation branches are moved forward toward the supply. The circuit (Fig. 30) also splits the resistance R'_2/S into the equivalent resistance of the secondary winding and an additional resistance $(1-S)R'_2/S$. This is done so that the power P_2 that is transferred from the primary into the secondary through the air gap can be divided into a component

$$P_{2\text{diss}} = m_1 R'_2 I'^2_2 = SP_2, \tag{71}$$

which is dissipated as heat in the conductors of the m_1 phases of the secondary winding, and a component

$$P_{\text{conv}} = m_1 \frac{1-S}{S} R'_2 I'^2_2 = (1-S)P_2, \tag{72}$$

which represents the power converted into mechanical form.

It now becomes possible to define a secondary efficiency:

$$\eta_2 = P_{\text{conv}}/P_2 = 1 - S. \tag{73}$$

This relation clearly shows that although the induction machine can operate at any value of speed and therefore S, in practice efficiency considerations force continuous operation at low values of S, that is, at speeds that differ from the synchronous speed ω_s by only a small percentage. Observing that the primary current I_1 differs only little

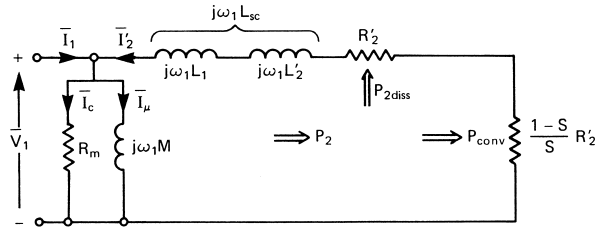


FIGURE 30 Simplified equivalent circuit of induction machine with short-circuited secondary.

from I_2', one can account also for the losses in the primary winding and write

$$P_{\text{diss}} \sim m_1(R_1 + R_2')I_2'^2 = \frac{R_1 + R_2'}{R_2'}SP_2. \tag{74}$$

Still to be accounted for, however, are the losses in the iron core,

$$P_{\text{iron}} = m_1 V_1^2 / R_{\text{m}}, \tag{75}$$

and the mechanical losses $P_{\text{F+W}}$ due to friction in the bearings and windage. These must be subtracted from P_{conv} to obtain the net power available at the shaft.

The electromagnetic torque is given by

$$\begin{aligned} T_{\text{em}} &= P_{\text{conv}}/\omega_{\text{m}} \\ &= P_{\text{conv}}/(1-S)\omega_{\text{s}} = P_2/\omega_{\text{s}}, \end{aligned} \tag{76}$$

and if one introduces the appropriate values of P_2 and ω_{s},

$$\begin{aligned} T_{\text{em}} &= \frac{1}{\omega_{\text{s}}} m_1 \frac{R_2'}{S} I_2'^2 \\ &= m_1 p \frac{V_1^2}{\omega_1} \frac{R_2'/S}{(\omega_1 L_{\text{sc}})^2 + (R_2'/S)^2}, \end{aligned} \tag{77}$$

since, from Fig. 30,

$$\bar{I}_2' = \frac{\bar{V}_1}{R_2'/S + j\omega_1 L_{\text{sc}}}, \tag{78}$$

where $L_{\text{sc}} = L_1 + L_2'$.

This equation allows one to draw the torque–slip characteristics and, by making use of Eq. (61), the torque–speed curves. For this purpose one observes that the torque peaks when the slip satisfies the relation

$$\omega_1 S_{\max} = \omega_{2\max} = R_2'/L_{\text{sc}} \tag{79}$$

and attains a value

$$T_{\max} = m_1 p \left(\frac{V_1}{\omega_1}\right)^2 \frac{1}{2L_{\text{sc}}}. \tag{80}$$

One can see that the maximum torque is a function only of m_1, p, and L_{sc}, which are parameters characterizing the machine, and depends on the supply only through the ratio V_1/ω_1, which, according to Eq. (41) and if one lets $E_{\text{ph}} \sim V_1$, represents the flux interlinked with one phase winding. Most induction motors are operated from constant-voltage, constant-frequency supplies. In the few applications in which efficient operation at reduced speed is desired the synchronous speed ω_{s} must be reduced by reducing the supply frequency ω_1. If V_1 were maintained constant, the flux would tend to increase and the iron would be driven into saturation. For this reason V_1 is also reduced in proportion to ω_1, so that V_1/ω_1 can be considered constant.

If follows that $T_{\max}$ is practically constant and can serve as a convenient normalization parameter. Thus, dividing Eq. (77) into Eq. (80), one obtains

$$\frac{T_{\text{em}}}{T_{\max}} = \frac{2}{S/S_{\max} + S_{\max}/S}. \tag{81}$$

The torque–slip characteristic sketched in Fig. 31 highlights three regimes of operation. The machine is a motor in the interval between synchronous speed ($S = 0$) and stalling ($S = 1$). It is a generator when driven above synchronous speed, so that the slip and therefore the equivalent resistance $(1 - S)R_2'/S$ are negative. The machine operates as a brake when the rotor rotates in the direction ω_{m} opposite to ω_{s} and $S > 1$.

D. Starting Performance

Although Eq. (81) is useful in outlining a general trend, it cannot be relied on to yield accurate numerical results. The reason is not so much that it is based on a simplified equivalent circuit but in larger measure that the parameters of the machine vary with the conditions of operation. The inductances vary because of saturation in the iron, and the resistances, because of heating. In particular the starting performance of squirrel-cage motors is strongly affected by changes in the resistance and leakage inductance of the secondary circuit that accompany changes in ω_2.

This phenomenon can be understood with the help of Fig. 32. A rotor bar can be viewed as consisting of layers

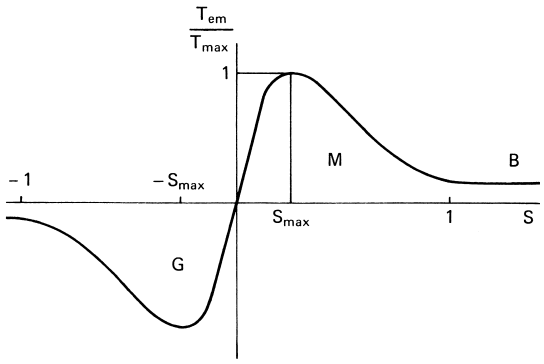


FIGURE 31 Torque–slip characteristic of induction machine (G, generating regime; M, motoring regime; B, braking regime).

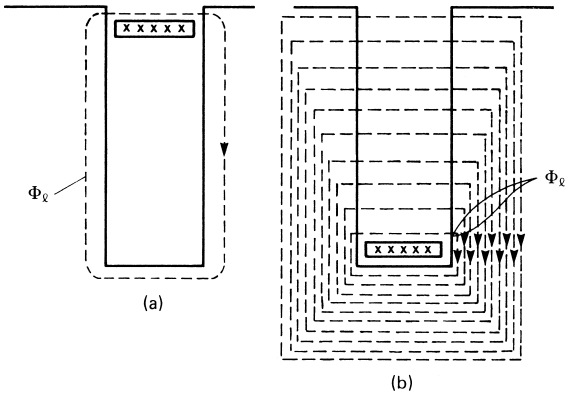


FIGURE 32 Leakage flux interlinked with current in upper layer (a) and lower layer (b).

lying in the slot one on top of the other and connected in parallel. The current is not necessarily distributed evenly among the layers but tends to concentrate in those that have the least impedance. At stalling, the secondary frequency ω_2 is high and equal to ω_1, because $S = 1$. Therefore, the dominant part of the impedance is the leakage reactance. As illustrated in Fig. 32 the current-carrying upper layer is interlinked with much smaller leakage flux than the bottom one. As a result of their low reactance the current tends to concentrate in the upper layers. This so-called skin effect not only reduces the overall leakage inductance, but also reduces the cross-sectional area available for the passage of current, which thereby increases the effective resistance. According to Eq. (79) an increase in R_2' and the decrease in L_{sc} that accompanies a decrease in L_2' lead to an increase in S_{max}. As can be seen from Fig. 31 a shift of S_{max} to the right results in an increase in the normalized torque at stalling ($S = 1$). This effect is compounded by the simultaneous increase in T_{max}, according to Eq. (80). To take advantage of this skin effect and further enhance the starting torque, one can make the rotor bars deep and narrow. Otherwise, one can use a double-squirrel-cage rotor with the upper cage made of high-resistance material.

As the motor picks up speed, ω_2 and with it the importance of the secondary reactance are reduced. The secondary current tends to distribute itself evenly and R_2' is reduced.

The increase in starting torque as a result of the skin effect is obtained at a relatively low increase in manufacturing cost. However, a price is exacted in terms of overall performance. First, part of the increase in torque comes at the expense of a higher inrush current I_{st}. This can be seen with the help of the equivalent circuit of Fig. 30. At starting with $S = 1$ the current is limited essentially by L_{sc}, and any decrease in L_{sc} leads to an increase in I_{st}. Next, as can be inferred from Fig. 32, the distribution of the current over a deep bar under running conditions leads to an increase in the leakage flux and with it x_2'. This lowers the rated power factor of the motor.

These drawbacks become especially significant in large motors. When large starting torques are needed with low inrush currents, the desired increase in the slip S_{max} at which maximum torque occurs is achieved by inserting external starting rheostats in the secondary circuit. This implies that the rotor must be wound and its terminals made accessible usually by means of slip rings. This construction and the additional apparatus required to short-circuit the external rheostats, once the motor has picked up speed, have a significant impact on the cost of the motor.

E. Single-Phase Motors

The power requirement for a single-family home or for a single housing unit is only a few kilowatts, usually supplied by a single phase. Therefore, the induction motor has been adapted to such supply conditions. When only one phase is available, the magnetic field distribution in the air gap no longer rotates, but instead is a stationary pulsating one. This can be viewed as consisting of the superposition of two waves rotating in opposite directions according to the trigonometric identity

$$E_m \cos \omega_1 t \cos p\theta = (E_m/2)[\cos(\omega_1 t + p\theta) + \cos(\omega_1 t - p\theta)]. \quad (82)$$

Each of these waves induces a system of currents in the squirrel cage of the rotor and therefore produces a torque. If one expresses each of these torques as a function of the speed $\omega_m = (1 - S)\omega_s$, instead of S, Fig. 31 leads to Fig. 33. Defining positive values of speed and torque as forward and negative as backward, it appears that at stalling positive and negative torques cancel out exactly. Therefore, the motor cannot start. If, however, it is given an initial acceleration in either direction, it will produce a

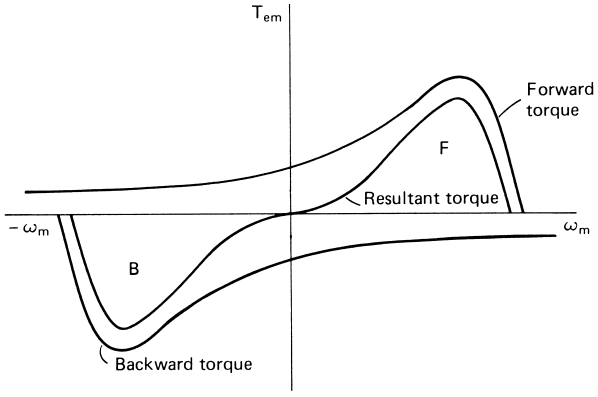


FIGURE 33 Torque–speed characteristics in single-phase motor (F, forward motion; B, backward motion).

net torque, accelerate, and, when it reaches steady speed, operate nearly like a polyphase machine.

The problem is, then, to enhance the rotating field in the desired direction at starting. This is done by providing an auxiliary phase winding. Various solutions have been adopted depending on how strong the starting torque may be.

In all cases the auxiliary winding is displaced in space from the main winding. The required phase difference in the currents flowing in the two windings when they are fed from the same voltage is obtained by the different nature of the circuit impedance. In order of decreasing magnitude of developed torque and cost, one may list the following:

1. Capacitor-run motors, such as used to drive the compressors in air-conditioning units. The auxiliary winding is connected permanently to the supply through a capacitor.
2. Capacitor-start motors, such as used in refrigerators. The auxiliary winding is connected through a capacitor only during starting.
3. Resistance split-phase motors. The auxiliary winding is wound with fewer turns and much smaller size of wire than that of the main winding, so that it has a considerably greater resistance-to-reactance ratio. The auxiliary winding is connected only during starting.
4. Shaded-pole motors, such as used in fans. The auxiliary winding is a so-called shading coil consisting of an uninsulated short-circuited copper strip embedded into one of the main poles and spanning about one-third of its pitch.

The direction of motion in motors of types 1, 2, and 3 can be changed by reversing the connections of either the main or the auxiliary windings.

V. STEPPER AND SWITCHED-RELUCTANCE MOTORS

Reluctance motors were mentioned in Section III.C with reference to synchronous motors without field excitation. These machines have a smooth armature usually placed on the stator and a rotor having a number of salient poles equal to the number of poles in the armature winding. Recently, with the advent of digital computers, a new machine has gained popularity. Its main advantage is that it can provide very accurate positioning when it is energized by an appropriate sequence of digital signals generated by a microprocessor and amplified by a solid-state power conditioner. Therefore, it has found many applications, especially in computer drives, robots, and machine tools. Another advantage is its simple and sturdy construction. Its simplest form is sketched in Fig. 34.

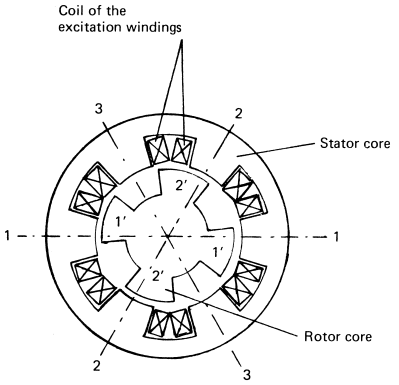


FIGURE 34 Sketch of a reluctance machine having a different number of poles on stator and rotor.

The motor consists of stacks of laminations for both stator and rotor. In contrast with conventional machines the motor is doubly salient; that is, it has salient poles on both sides of the air gap. Also, the numbers of poles on the stator and rotor are not equal. Diametrically opposite stator poles carry coils connected in series and constitute one phase of the winding. Each phase is fed a voltage pulse in sequence from a dc source. The solid-state switching arrangement is called a chopper. So that saturation of the iron is avoided, the flux must be "reset" at the end of each pulse by impressing a voltage of opposite polarity on the winding. This is accomplished by the circuits of Fig. 35. In circuit (a) a separate dc source is needed; in circuit (b) the phase winding consists of a bifilar arrangement of

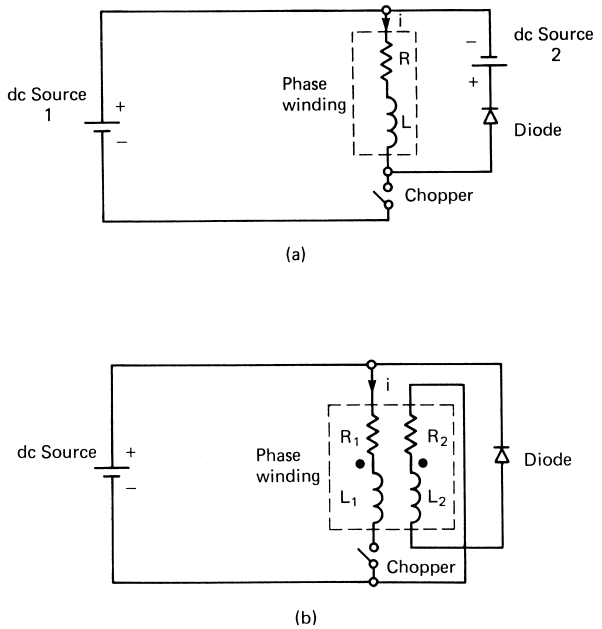


FIGURE 35 Supply schemes for reluctance motors. (a) Circuit with two dc sources; (b) circuit with bifilar-wound coils.

two tightly coupled coils. The diode allows the current to flow in the auxiliary loop only after the switch is opened and allows recovery of the energy stored in the magnetic field and its transfer back to the dc source.

Operation of the motor can be understood with the help of Fig. 34. Torque is developed by the tendency of the magnetic circuit to adopt a configuration of minimum reluctance, that is, for the rotor poles to come into alignment with the stator poles, so that the inductance of the energized coils is maximized. Therefore, if phase winding 1 is energized first, the rotor poles $1'$ will align with poles 1 of the stator. Since in this figure the number of rotor poles is smaller than that of the stator poles, the rotor pole pitch is larger. It follows that when poles $1'$ and 1 are aligned poles $2'$ overlap the stator pole faces only partially and in that portion farthest away from the axis of phase 1. If phase 2 is energized next, the rotor poles $2'$ will align with poles 2 of the stator. The motor will thus move one step clockwise. If, instead of phase 2, phase 3 is energized after phase 1, the rotor will move anticlockwise. Therefore, when the pole pitch of the rotor is larger than that of the stator, the rotor moves in the opposite direction than the sequence of phase energization. When the pole pitch of the rotor is smaller, the rotor follows the sequence of phase energization.

It can be seen that in the case of Fig. 34 the angular displacement of the rotor in one step is

$$\alpha = 90^\circ - 60^\circ = 30^\circ, \tag{83}$$

or if N_s and N_r denote the number of poles in the stator and rotor, respectively,

$$\alpha = 2\pi/N_r - 2\pi/N_s. \tag{84}$$

The relation between f_{ph}, the pulse frequency in one phase winding, and rpm can be found by observing that, in a motor having q phases, the rotor advances q steps or an angle equal to $q\alpha$ during one period. Therefore, the number of pulses (in one phase winding) per revolution of the rotor is $2\pi/q\alpha$, and the frequency is

$$f_{ph} = \frac{2\pi}{q\alpha}\frac{\text{rpm}}{60} = \frac{N_r N_s}{q(N_s - N_r)}\frac{\text{rpm}}{60}. \tag{85}$$

The frequency of the pulsed source that sequentially pulses the q stator phases is then

$$f_s = \frac{N_r N_s}{N_s - N_r}\frac{\text{rpm}}{60}. \tag{86}$$

An approximate relation between the torque developed by the motor and the current i flowing in the phase winding can be obtained by assuming that i remains constant during a pulse and by neglecting saturation effects in the iron. In this case the energy and coenergy stored in the magnetic field are equal to

$$W_{mag} = \tfrac{1}{2}Li^2, \tag{87}$$

where L is the inductance of the phase winding. The principle of conservation of energy then leads to the following functional dependence of the torque T on the rotor angular displacement θ and the excitation current i:

$$T(\theta, i) = \frac{\partial W_{mag}(\theta, i)}{\partial \theta} = \frac{1}{2}i^2\frac{dL}{d\theta}. \tag{88}$$

If one considers two limiting positions—(1) when stator and rotor poles are aligned and (2) when the stator pole faces the interpolar space of the rotor—the angular displacement is π/N_r and the average torque is

$$\bar{T} = \frac{1}{2}\frac{N_r}{\pi}i^2(L_1 - L_2). \tag{89}$$

The torque actually vanishes at these two limiting positions and peaks somewhere between. As a first approximation, the torque with only one phase energized displays a sinusoidal variation with the angular displacement. A similar curve displaced by the step angle α, with the appropriate sign according to Eq. (84), is exhibited when one phase is de-energized and the adjacent one is energized. Stepping operations under the assumption that electrical and mechanical transients are neglected are plotted in Fig. 36. Full clockwise (path a–d–e) and counterclockwise (path a–b–c) steps are shown with zero load torque T_L in (a) and with finite load torque in (b). Since the difference between the torque developed by the motor and the one applied to the shaft is the accelerating torque, the rotor will reach its new resting position faster in the direction favored by the applied torque.

The machine takes different forms, depending on its application. When it is used as a positioning device, it is desirable that the step α be small. According to Eq. (84), this implies a large number of poles. Because it would be impractical to have each stator pole energized separately, one uses a vernier arrangement by which each phase energizes a large core comprising a number of poles on its face. A typical cross section of such a machine is shown in Fig. 37. The step can be reduced to one-half the original one in the "half-stepping" mode of operation by exciting sequentially one or two adjacent phase windings at any one time. Operation resulting in a clockwise half step is shown in Fig. 38. The torque developed by a motor of given size can be increased by providing the rotor with permanent magnet excitation. In a common construction the rotor is split into two stacks with poles staggered by one-half pitch. The permanent magnet is oriented axially and is placed between the two rotor stacks under their salient poles.

A torque versus displacement angle plot is shown in Fig. 39. A small (less than 10%) *detente* torque T_d is developed even in the absence of stator excitation, because of

the dependence of the reluctance on the rotor position. Permanent magnet (PM) machines provide rotation in both directions with two phases, instead of the minimum of three needed with the variable reluctance (VR) machines. The consequent simplification in the power conditioner must be weighed, however, against the need in PM machines for either a bidirectional current supply or a full-sized bifilar winding.

Machines used for accurate positioning are called *stepper* or *stepping* motors. When speed control rather than positioning is desired, a large number of poles is not needed, and, according to Eq. (86), it would lead to a large source frequency. These machines, having a small number of poles, are called switched-reluctance motors and have been built in sizes up to a few tens of kilowatts.

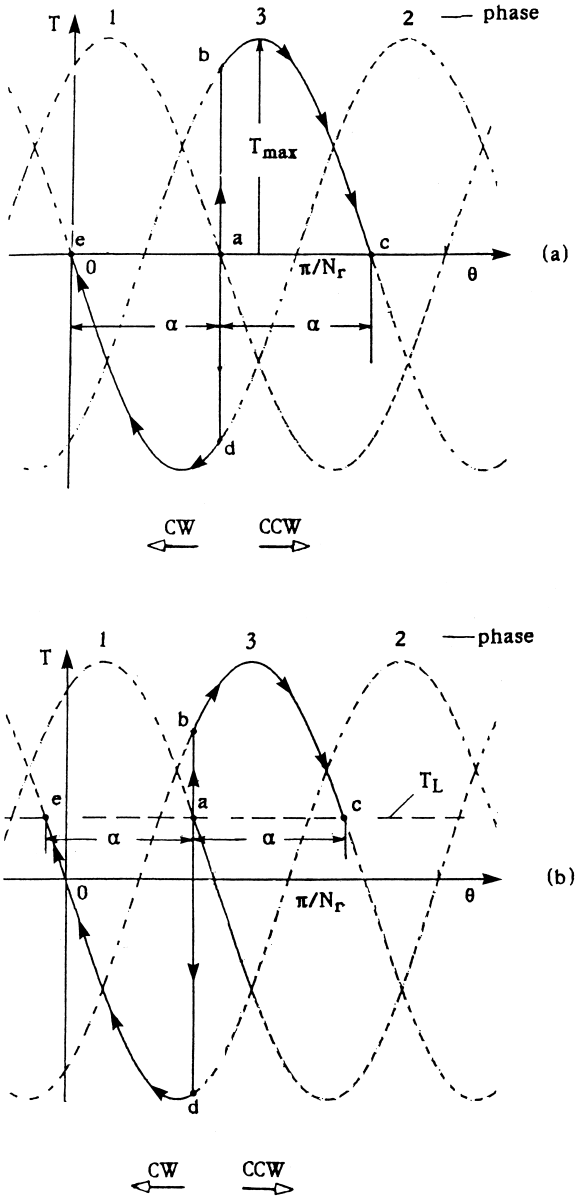


FIGURE 36 Stepping operation of a three-phase stepper motor with $N_s = 6$ and $N_r = 2$. (a) $T_L = 0$; (b) $T_L \neq 0$.

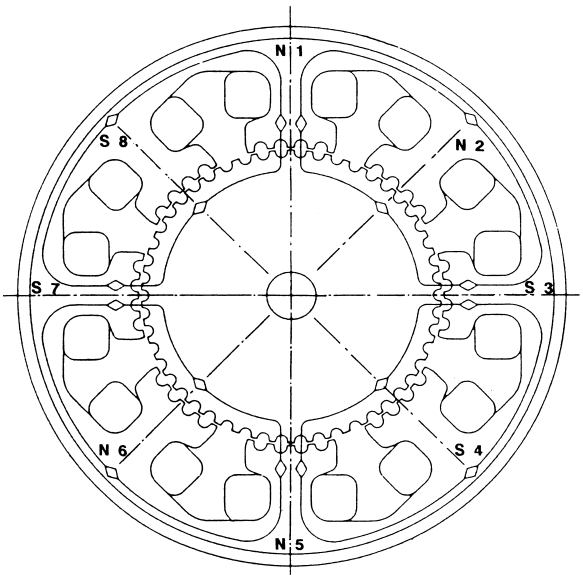


FIGURE 37 Cross section of stepper motor with vernier arrangement.

VI. WRITTEN-POLE[1] MOTOR AND GENERATOR

A. Introduction

The Written-Pole machine is a hybrid technology: a squirrel-cage/synchronous machine. It was developed by the Precise Power Corporation, with EPRI (Electric Power Research Institute) support, to overcome some fundamental drawbacks of conventional motors, such as high starting current. Therefore, it is particularly valuable for remote and rural applications, where single-phase power supply is the norm. Typically, the Written-Pole motor (WPM) requires a starting current that is less than twice its running current. An additional advantage is the motor's very high efficiency—92% in comparison with 85% for conventional motors.

The Written-Pole machine is also being developed in motor–generator sets that can be used as an uninterruptible power supply to substitute for maintenance-intensive battery plants.

[1]Written-Pole is a 1996 trademark of Precise Power Corporation, Bradenton, Fla. This company was granted, at least, eight patents, some of which have already expired. The oldest patent is U.S. Patent No. 4,168,459, "Non-Interruptible Power Supply Systems." Issued September 18, 1979, it expired October 25, 1997. It was also covered in Australia, Canada, Great Britain, and Japan.

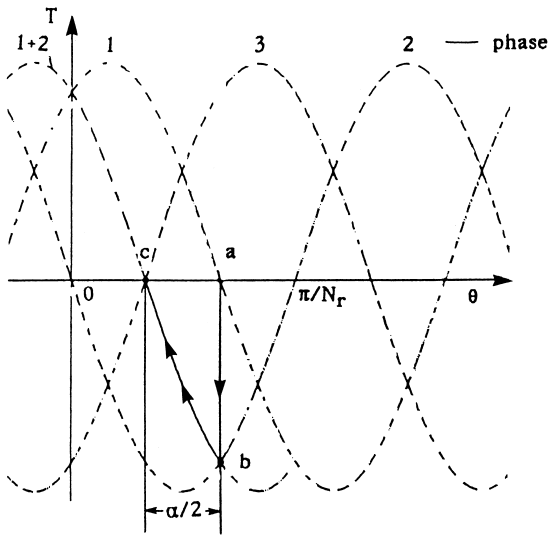


FIGURE 38 Half-step operation of a three-phase variable reluctance (VR) stepper motor.

In 1997, the WPM won the New Product Award of the National Society of Professional Engineers. In 1994, the technology also won an R&D 100 Award from *R&D Magazine*, which recognized the WPM as one of that year's 100 most technologically significant products.

B. Principles of Operation

The magnetic poles of conventional PM motors are fixed in one position and the rotational speed is tied in with the frequency. In a WPM the magnetic poles are more flexible; the machine incorporates a special coil that "imprints" poles onto a ring of magnetic material attached to the spinning rotor. This coil operates in much the same way as a tape recorder, but, unlike a tape recorder, which is designed to erase and record sound only occasionally, the Written-Pole coil writes and erases the magnetic poles with the line frequency during startup. The alternating magnetic poles serve to pull the rotor forward, which increases the rotor's speed during each rotation until the motor reaches full design speed. When the full speed is achieved, the pole writing procedure ceases and the poles remain as last written.

The Written-Pole technology does not work on its own to start the motor and it is not activated until the motor is brought up to 80% of its design speed by a squirrel-cage rotor. Whereas in ordinary motors the squirrel cage is made of low-resistance material in order to reduce the losses during running, the WPM uses high-resistance material in

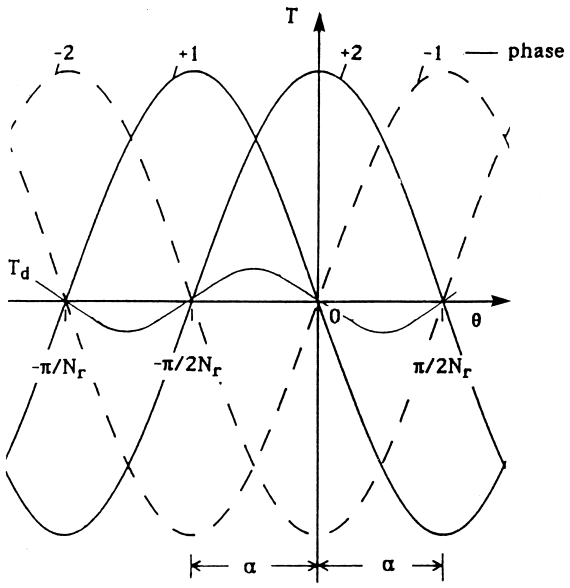


FIGURE 39 Torque as a function of the displacement angle for a permanent magnet (PM) stepper motor.

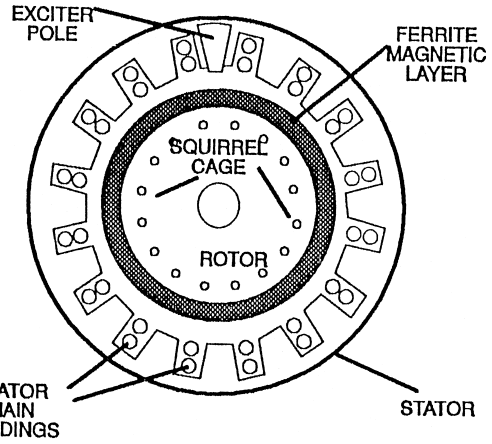


FIGURE 40 Written-Pole motor, internal rotor construction.

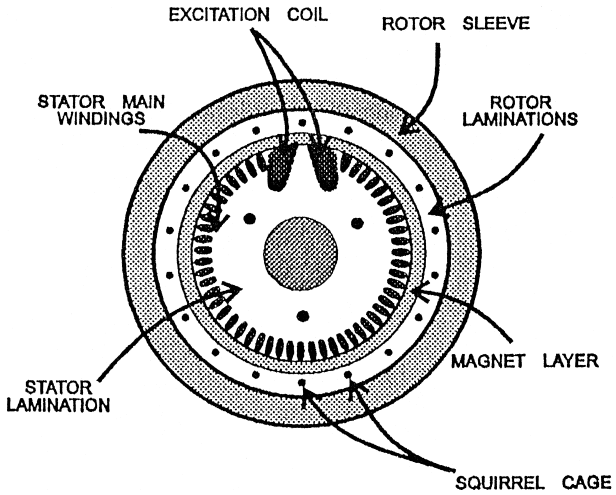


FIGURE 41 Written-Pole motor, external rotor construction.

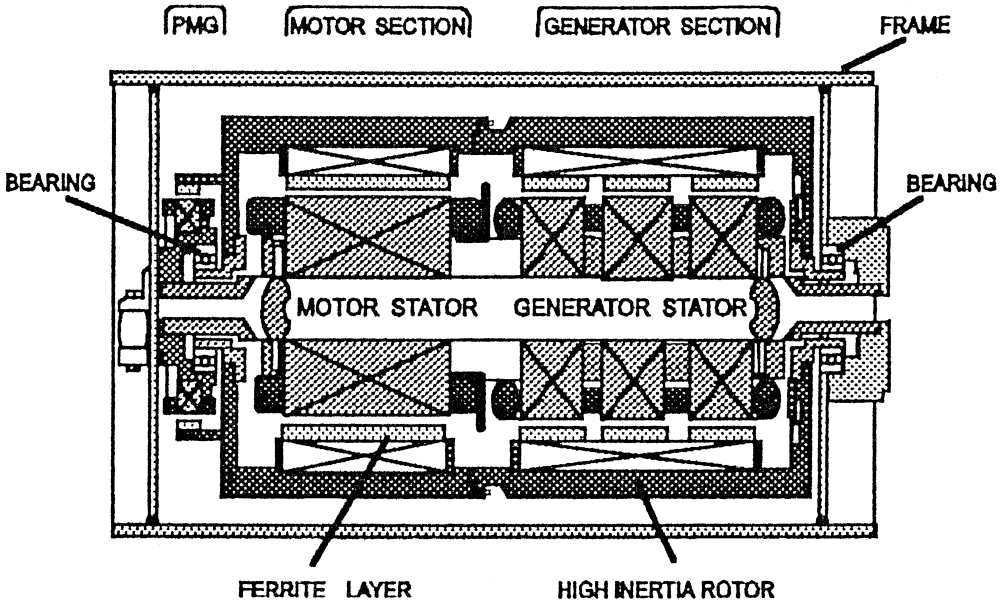


FIGURE 42 Three-phase Written-Pole motor–generator assembly, axial construction.

order to reduce the starting current. The lower starting current and, hence, the lower generated heat allow the WPM to start loads more than 10 times as heavy as those that can be started with conventional motors of the same size.

C. Construction

The Written-Pole machine, as seen in Fig. 40, has a continuous layer of "permanent" magnetic material on the surface of the rotor. Typically, this material is a highly anisotropic S_rO–Fe_2O_3 composition ferrite ranging in thickness in the oriented direction from 15 to 30 mm, depending on the machine size. The main windings of the stationary parts of the machine are similar to those of conventional machines, with an additional concentrated excitation coil around an exciter pole. The magnetic layer can be magnetized (or "written") into any desirable magnetic pole pattern by the exciter pole *while the machine is operating*.

So that the machine can ride through momentary power interruptions of 10–20 sec or longer, the WPM is also built with a high-inertia rotor external to the stator, as shown in Fig. 41, Motor–generator (M-G) sets utilizing Written-Pole technology provide additional ride-through for all voltage sags and even complete blackouts. The synchronous motors in these new machines operate at unity power factor. The M-Gs prevent harmonic feedback from the load to the utility. The combination of motor and generator on a common shaft is shown in Fig. 42.

SEE ALSO THE FOLLOWING ARTICLES

ELECTRIC MACHINES, DESIGN • ELECTRIC MACHINES, DRIVES • ELECTROMAGNETICS • MACHINE DESIGN • TRANSFORMERS, ELECTRICAL • TURBINE GENERATORS

BIBLIOGRAPHY

Del Toro, V. (1990). "Basic Electric Machines," Prentice-Hall, Englewood Cliffs, N.J.
Fitzgerald, A. E., Kingsley, C., and Umans, S. D. (1990). "Electric Machinery," 5th ed., McGraw-Hill, New York.
French, P. S.-J. R. (1984). *IEE Proc.* **131,** Pt. B, No. 5, 209–219.
Frieren, D., and St. George, D. (1996). "Independent Laboratory and Field Test of a WPM," *Proceedings of WESCON'96*, Anaheim, Calif., pp. 210–215.
Krause, P. C., and Wasynczuk, O. (1989). "Electromechanical Motion Devices," McGraw-Hill, New York.
Leehouts, A. C. (1987). "The Art and Practice of Step Motor Control," Intertec Communications Inc., Ventura, Calif.
Levi, E., and Panzer, M. (1982). "Electromechanical Power Conversion," Krieger, Melbourne, Fla.
Matsch, L. W. (1977). "Electromagnetic and Electromechanical Machines," 2nd ed., Harper & Row, New York.
Morash, P. J. (1996). "An Overview of Written-Pole Technology," presented at the Winter Meeting, P.E.S., Tampa, Fla.
Nasar, S. A., and Unnewehr, L. E. (1983). "Electromechanics and Electric Machines," 2nd ed., Willey, New York.
Ray, W. F. *et al.* (1984). *IEE Proc.* **131,** Pt. B, No. 5, 220–225.
Slemon, G. R., and Straughen, A. (1980). "Electric Machines," Addition-Wesley, Reading, Mass.
Wildi, T. (1981). "Electrical Power Technology," Wiley, New York.
"The Written-Pole revolution," *EPRI J.* (May/June 1997), 27–35.



Fire Dynamics

D. D. Drysdale
University of Edinburgh

I. Introduction
II. Modes of Burning
III. Flammability Limits
IV. Ignition
V. Steady Burning of Combustible Liquids and Solids
VI. The Fire Plume
VII. Fires in Compartments
VIII. Smoke Production and Movement

GLOSSARY

Backdraft Sudden increase in the intensity of a fire caused by the admission of air after a prolonged period of burning under poorly ventilated conditions.

Combustion Chemical reaction between a fuel and oxygen.

Compartment Enclosure within a building. The compartment boundaries can be regarded as confining the fire, at least until flashover.

Explosion Process producing a shock wave, most frequently created by the violent bursting of a container or containing structure. The energy may come from combustion (e.g., burning of a confined flammable gas/air mixture).

Extinction Process by which a self-sustaining combustion reaction is terminated.

Fire In the present context, uncontrolled combustion.

Firepoint Limiting condition for establishing a flame at the surface of a combustible liquid or solid. It is expressed as a bulk temperature for liquids.

Flashover Stage in a compartment fire during which a localized fire spreads very rapidly to become one in which all combustible surfaces are involved.

Flashpoint Minimum liquid fuel temperature at which the vapor–air mixture at the surface is flammable.

Fire resistance Normally applied to elements of structure, fire resistance is determined as the time to failure in a large-scale standard fire test.

Fire severity Ill-defined term referring to the ability of a fire to cause damage.

Flammability limit Limiting concentration of a flammable gas in air marking the boundary between mixtures that are flammable and those that are nonflammable.

Flammable Literally, "easily set on fire." It is commonly applied to gases that form flammable mixtures with air, as such mixtures can be ignited by minute amounts of

energy (e.g., a spark from static electricity). The term has a specific meaning when applied to combustible liquids but has no strict definition when applied to solids.

Flame Self-sustaining reaction in the gas phase between fuel vapor and oxygen (from the air).

Ignition Process by which a self-sustaining combustion process is initiated.

Smoldering Self-sustaining combustion process involving the formation and oxidation of char. Only materials that produce a char on heating can exhibit this type of combustion.

Wood crib Sticks of wood arranged regularly as a cross pile. Cribs are commonly used as the fuel bed for standard fires, as they are well understood and reasonably reproducible.

FIRE DYNAMICS deals with those topics that are relevant to our understanding of the behavior of unwanted fire or "wildfire." It draws together aspects of a number of fundamental disciplines, including chemistry, physics, and the engineering sciences, particularly heat and mass transfer and fluid mechanics. The topics of ignition, spread of flame, and steady burning are discussed in association with a review of the concepts of flammability and their bearing on the identification of those parameters that determine the "fire properties" of combustible materials. This is followed by a more detailed discussion of fire growth in compartments, leading to an examination of the fully developed fire. The relevance of this to life safety in buildings and to the concepts of fire severity and fire resistance is examined. Fire dynamics also includes smoke movement in buildings.

I. INTRODUCTION

The scientific study of unwanted fire was largely neglected until about 30–35 years ago, when it became apparent that the lack of understanding of the mechanisms of fire development and spread was a serious hinderance to progress in the area of fire safety. Although fire research groups—sponsored mainly by government and insurance organizations—had been set up some 20 years earlier, their fundamental work was not widely disseminated. A large amount of valuable information on fire behavior began accumulating from about 1950 onward, but little of this was published in the open literature. At the same time, the need for such information began to increase as new materials became available and began to be used widely. It was then no longer possible to rely on experience of fires involving the "traditional" materials (wood, wood products, wool, hair, etc.) to predict (or in some cases understand) the behavior of fires involving synthetic polymers.

Tests that had been developed to assess the "fire properties" of traditional materials were found to be grossly inadequate when applied to the new generation of materials. In many well-documented cases, materials that were deemed to be "safe" on the basis of their performance in one of these "standard tests" were found to be hazardous in the context of the full-scale fire scenario. This situation arose because there existed no framework of knowledge that could be used to apply the results of these tests to predict likely fire behavior. There was a lack of appreciation of several factors that are now known to be highly significant in determining how materials burn in real fires. These include the role of radiative heat transfer in a developing fire, the effect of scale (i.e., the physical size of the fire), and the importance of the physical form and orientation of a material or "fuel bed."

These difficulties were created because of a lack of motivation to bring together the constituent disciplines required to describe fire processes. While fire is an extremely complex phenomenon, it is basically a manifestation of combustion in which a fuel reacts with oxygen from the air to produce combustion products with the release of energy (heat). Combustion has received considerable attention by research scientists and engineers, as it is the major source of energy for industrial and domestic purposes. Understandably, research was directed toward improving combustion systems, such as furnaces and internal combustion engines. While this work has become highly sophisticated, involving the application of several branches of science and engineering to solve complex problems, it cannot be applied readily to the fire process. Combustion systems that have been studied have tended to involve closed systems, which are amenable to mathematical modeling because the boundaries of the problems are well defined. While the same chemical and physical laws apply, detailed understanding of "unwanted fire" requires that a much wider range of boundary conditions be considered, many of which cannot easily be specified in advance. The systematic study of fire has thus been neglected by the research worker simply because it has been regarded as too complex. However, with the advent of the computer, it has now become possible to contemplate mathematical models of fire: this has given a new impetus to fire research at a time when an improved understanding of fire processes has become necessary to underpin the application of fire safety engineering to performance based fire regulations."

II. MODES OF BURNING

In the context of fire, "fuels" may take many forms, encompassing gases, liquids, and solids. However, in the flaming mode, the combustion process involves a reaction

in the gas phase between fuel vapor and oxygen from the air. Clearly, except when the fuel is initially gaseous (e.g., hydrogen, methane), conversion to the gaseous state must first take place. For a liquid, this is achieved by evaporative processes corresponding to boiling of the surface layer. However, in the case of combustible solids, chemical decomposition must take place to convert large, involatile fuel molecules into relatively small molecular fragments that can escape (vaporize) from the surface at the temperatures at which the decomposition has taken place. In the steady burning process, the energy necessary to maintain this flow of flammable vapors is supplied by heat transfer from the flame, where energy is liberated by combustion. This feedback mechanism is illustrated in Fig. 1 and is discussed further in Section V. However, it is appropriate to emphasize here the importance of heat transfer at all stages of a fire—initiation, growth, steady burning, and extinction. The three heat-transfer mechanisms (conduction, convection, and radiation) all contribute, although their relative importance changes as fire develops. Thus, convection tends to be important when the fire is small, but once the area of burning fuel grows to beyond the point at which its diameter is 0.2–0.3 m, thermal radiation from the flame becomes increasingly important (Section V). Flame emissivity is strongly dependent on whether or not the flame contains particulate matter (i.e., minute particles of "soot"), some of which may burn up while others escape the flame to form smoke. The presence of these species gives most diffusion flames their characteristic yellow luminosity. Nonluminous flames, such as those burning above methanol or poly(oxymethylene), have very low emissivities, as the only emitting species are the gases carbon dioxide and water vapor. This has important implications regarding rates of burning and fire detection (Section V).

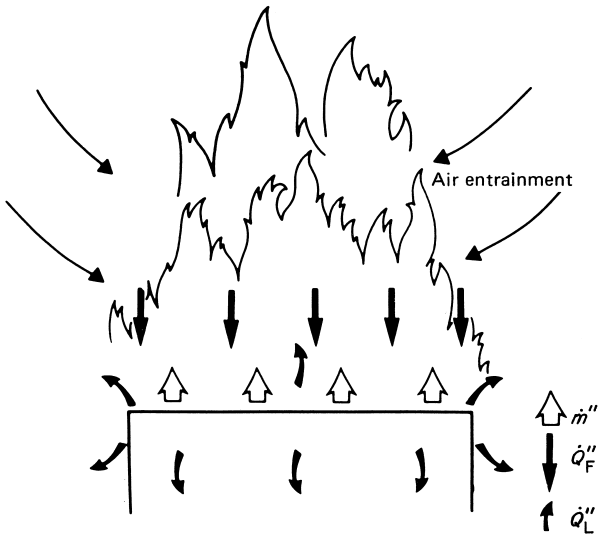


FIGURE 1 Representation of a burning surface, showing the heat- and mass-transfer processes: $\dot{m}''$, mass flux from the surface; $\dot{Q}''_F$, heat flux from the flame to the surface; $\dot{Q}''_L$, heat losses, expressed as a flux through the surface. [From Drysdale, D. (1998). "Introduction to Fire Dynamics," John Wiley and Sons, Chichester.]

In Fig. 1, flame exists in those regions where fuel vapor and air mix. In the idealized model first proposed by Burke and Schumann, the flame corresponds to a thin reaction zone at the interface where the fuel–air mixture is stoichiometric. This is known as a diffusion flame, as diffusive mixing is an integral part of the burning mechanism. This type of flame is obtained when a jet of flammable gas burns in air and is well known as the flame that is obtained from a Bunsen burner with the air inlet closed. Its structure and appearance are influenced by the flow velocity of the gas issuing from the mouth of the burner. At sufficiently high velocities, the flame becomes turbulent and is dominated by the momentum of the flow of fuel. During the burning of a condensed fuel (i.e., a combustible liquid or solid), the momentum of the fuel vapor is relatively unimportant and the behavior is determined by the upward, buoyant movement of the burning gases.

An alternative mode of burning is possible if the flammable gas (or vapor) and air are intimately mixed before ignition: then, provided that the fuel concentration is within the limits of flammability (Section III), combustion can be initiated by means of a "pilot" ignition source, such as a spark or small flame. A stable premixed flame is obtained with a Bunsen burner if the air inlet port is open. The resulting rate of burning of the mixture is very high and proceeds as the reaction zone (i.e., the flame) propagetes into the "unburnt mixture," converting the fuel into products of combustion at high temperatures, typically $>2000^{\circ}$C. The resulting expansion can be as high as sevenfold. If initially the mixture is enclosed, high pressures will be produced, capable of causing substantial damage to any confining structure. Such is the mechanism of the gas explosion. A similar mode of burning is observed with condensed fuels if they are in a finely divided state of dispersion, such as mists of combustible liquids and suspensions of combustible dusts in air.

Some solids can undergo a mode of combustion known as smoldering, for which the prerequisite is that the solid must be capable of yielding a rigid carbonaceous char when heated. Surface oxidation of the char releases heat, sufficient of which is transferred into the unaffected fuel ahead of the oxidation zone to cause decomposition of the fuel. Oxidation of the fresh char will begin almost as soon as it is formed. The rate of burning is extremely slow in comparison to flaming. The gaseous decomposition products escape more or less unburnt to produce an aerosol smoke, which will tend to form an oily deposit on cool

surfaces. Smoldering will not be discussed further here, although it is recognized that many fires begin with a smoldering phase, initiated, for example, by a carelessly discarded cigarette. Some of the materials which have been commonly used in furnishings can be set smoldering in this way and can burn slowly for a considerable time before undergoing a transition to flaming combustion.

III. FLAMMABILITY LIMITS

As indicated already, mixtures of flammable gases and vapors in air will burn rapidly if ignited. However, not all such mixtures will propagate flame: only those mixtures whose compositions lie between well-defined limits are capable of burning. For example, methane–air mixtures will not burn (at ordinary temperatures and pressures) if the methane concentration is less than 5% (fuel lean) or more than 15% (fuel rich). All fuel gases and vapors exhibit such limits, which are known as the lower and upper flammability (or explosion) limits, respectively. Examples are given in Table I. This knowledge is particularly important in the petrochemical and processing industries, where flammable materials are handled. Accidental formation of a flammable mixture within an enclosure is particularly hazardous, as ignition would give rise to an explosion.

The severity of such an explosion will depend on the nature of the flammable gas and its concentration in air (within the limits). Generally speaking, stoichiometric mixtures—in which there is an exact equivalency between fuel and air—will give the most violent explosions, as the rate at which flame propagates through the unburnt mixture (the burning velocity) is a maximum for any gas at that particular concentration. The variation of burning velocity with mixture composition for propane is shown in Fig. 2: the maximum burning velocities for a number of gases is shown in Table I. Burning velocities are increased if the oxygen concentration in the air is increased: the rate of burning is also enhanced if the initial mixture is in a turbulent state. Indeed, self-generated turbulence, which can occur if a premixed flame is propagating along a pipe or duct, can in certain circumstances give rise to sufficient flame acceleration to produce a detonation, in which the combustion is associated with a shock wave. A detonation is an extremely destructive event, traveling at velocities of up to 2.5 km/sec and generating high pressures.

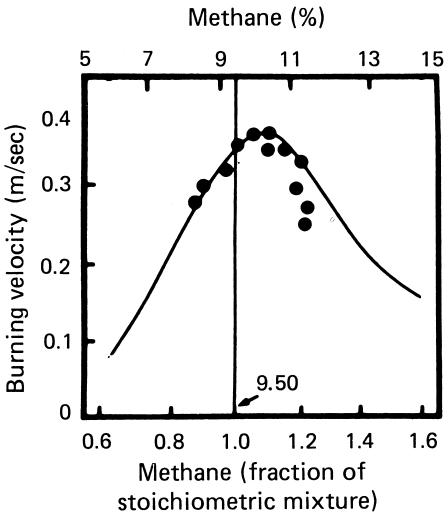


FIGURE 2 Variation of burning velocity with composition: propane/air mixtures at ambient temperatures. [From Lewis, B., and von Elbe, G. (1961). "Combustion, Flames and Explosions in Gases," Academic Press, New York and London.]

Several mechanisms have been proposed to explain the existence of flammability limits, ranging from aerodynamic instability at low flame speeds to critical heat losses

TABLE I Flammability Data for Gases and Vapors[a]

	Lower flammability limit (% vol.)	Upper flammability limit (% vol.)	Stoichiometric concentration (% vol.)	Burning velocity (m/sec)
Hydrogen	4.0	75	30.8	3.2
Carbon monoxide	12.5	74	29.8	0.43
Methane	5.0	15	9.4	0.37
Propane	2.1	9.5	4.0	0.42
Butane	1.8	8.4	3.1	0.42
n-Octane	0.95	—	1.6	—
Ethylene	2.7	36	6.6	>0.69
Acetylene	2.5	(100)[b]	7.7	1.7
Methanol	6.7	36	12.2	0.52

[a] From Zabetakis, M. G. (1965). U.S. Bureau of Mines Bulletin 627.
[b] Pure acetylene will undergo explosive decomposition.

TABLE II Thermal Capacities of Common Gases at 1000 K

Gas	$C_p^{1000\,K}$ (J/mol K)	Concentration required to inert stoichiometric CH_4–air mixture (% vol.)
Carbon dioxide (CO_2)	54.3	23
Water (vapor) (H_2O)	41.2	27
Nitrogen (N_2)	32.7	37
Helium (He)	20.8	48
Halon 1301 (CF_3Br)	69.5	8[a]

[a] Refers to *n*-hexane–air mixture.

at the limit. The latter theory relies on the proposition that flame can no longer propagate if the rate of heat loss from the reaction zone equals or exceeds the rate of heat release by combustion. This is consistent with the observation that the temperature of a flame propagating through a near-limit mixture is still relatively high (approximately 1300°C for most hydrocarbons and hydrocarbon derivatives).

The temperature of a premixed flame is determined by the heat of combustion of the fuel, but limited by the thermal capacity and thermal dissociation of the products (including nitrogen and any excess oxygen from the air). The maximum flame temperature is achieved with a stoichiometric mixture, and progressive dilution by excess fuel (fuel-rich), excess air (fuel-lean), or another chemically inert gas, such as carbon dioxide, will reduce the flame temperature toward the limit. Extinction at the limit occurs discontinuously, because the rate of heat release decreases with temperature very much more rapidly than the rate of heat loss: the latter becomes dominant as the limit is crossed.

Any flammable mixture can be rendered nonflammable by dilution by "inert" gases such as carbon dioxide and nitrogen. Their effectiveness is in proportion to their respective thermal capacities (Table II). However, there is a group of extinguishing agents which are effective at very much lower concentrations. These act as chemical inhibitors, interfering with the free radical chain mechanisms in the flame. An example is Halon 1301 (CF_3Br) (Table II), but unfortunately this and similar compounds are know to have adverse effects on the ozone layer and have been banned under the Montreal protocol.

The thermal theory of extinction may be used semiquantitatively to explain the operation of flame arresters. These may be used to prevent the accidental propagation of flame along a pipe containing a flammable gas/air mixture. They consist of arrays of narrow channels through which the flame must pass in order to continue traveling along the pipe and operate by removing sufficient heat from the flame to cause "quenching." It may tacitly be assumed that the flame is cooled to 1300°C or less, but the mechanism is likely to be more complex, as the surfaces of the arrester will also inhibit the reaction by heterogeneous destruction of free radicals from the flame.

The concept of quenching a flame by cooling can also be applied to the limit of stability of a diffusion flame at the surface of a condensed fuel. This is discussed in more detail in Section IV.

Flammability limits are of importance in fire dynamics for a number of reasons. Not only do they define the range of concentrations within which an explosion hazard exists in an enclosure following a leak of flammable gas, but also they determine the conditions under which a solid or liquid fuel may be ignited by means of a small pilot ignition source. The vapor above the surface of a combustible liquid or solid can only be ignited if the vapor/air mixture is within the appropriate flammability limits at the location where the pilot is introduced. Whether or not sustained flaming will develop from ignition in this mixture will depend on the nature of the fuel and the temperature of the surface. These concepts are examined in the next section.

IV. IGNITION

Ignition can be defined as that process by which a rapid exothermic reaction is initiated, which propagates, or is self-sustaining, and causes the material involved to undergo change, reaching temperatures greatly in excess of ambient. For premixed gas–air mixtures within the flammability limits, it is possible to identify a minimum ignition energy below which ignition will not occur. This varies with concentration, exhibiting a minimum close to the stoichiometric ratio of fuel in air (Fig. 3). A perceptible

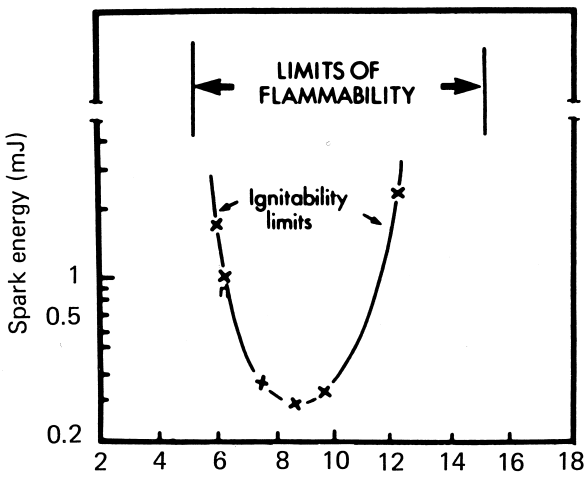


FIGURE 3 Ignitability curve and limits of flammability for methane/air mixtures at atmospheric pressure and 26°C. [From Zabetakis, M. G. (1965). U.S. Bureau of Mines Bulletin 627.]

TABLE III Flashpoints and Firepoints of Liquid Fuels

	Closed-cup flashpoint[a] (°C)	Open-cup flashpoint[b] (°C)	Firepoint$_b$ (°C)
Gasoline (100 octane)	−38	—	—
n-Hexane	−22	—	—
Cyclohexane	−20	—	—
n-Octane	13	—	—
iso-Octane	−12	—	—
n-Decane	46	52	61.5
n-Dodecane	74	—	103
Methanol	11	1(13.5)[c]	1(13.5)[c]
Ethanol	13	6(18.0)[c]	6(18.0)[c]
Propanol	26	16.5(26.0)[c]	16.5(26.0)[c]
n-Butanol	35	36(40.0)[c]	36(40.0)[c]
i-Pentanol	41	—	57
p-Xylene	27	31	44

[a] "National Fire Protection Association Handbook."
[b] Glassman, I. and Dryer, F. (1980/81). *Fire Safety J.* **3,** 123–128.
[c] Figures in brackets refer to ignition by a spark. All others refer to ignition by flame.

static spark (of the order of 1.0 mJ) is quite sufficient to ignite most stoichiometric mixtures; this observation serves to emphasize why in the interests of safety it is more prudent to avoid the formation of flammable mixtures rather than rely on the elimination of ignition sources.

One of the necessary requirements for the ignition of combustible liquids and solids is that the temperature of the fuel is high enough for there to be a flammable mixture above the surface. Liquid fuels are classified according to their closed-cup flashpoints, that is, the bulk temperature of the liquid below which the vapors cannot be ignited in a small experimental enclosure (e.g., the Pensky–Martens closed cup apparatus). The vapor pressure of the liquids at this temperature can be shown to correspond to the lower flammability limit of the vapor in air. The flashpoints of a number of common liquids are given in Table III. They are used to classify liquid fuels into several groups according to their ignition hazard: thus, in the U.K., the dividing line between "highly flammable liquids" and "flammable liquids" is 32°C, while a combustible liquid is defined as one with a flashpoint above 60°C. A comparison with the classification system used in the U.S. is shown in Fig. 4.

While the flashpoint is appropriate in characterizing the fire hazard of a liquid, it is found that sustained burning at a liquid surface will not follow ignition of the vapors unless the temperature is higher than the "firepoint," which is characteristically several degrees higher than the closed-cup flashpoint. In fact, stabilization of flame at the firepoint marks the lower limit of stability of a diffusion flame at a surface: it has been shown that there is a critical rate of supply of fuel vapors necessary for the establishment of a flame. At lower flowrates, a diffusion flame would lose too much heat to the surface for combustion to be sustained.

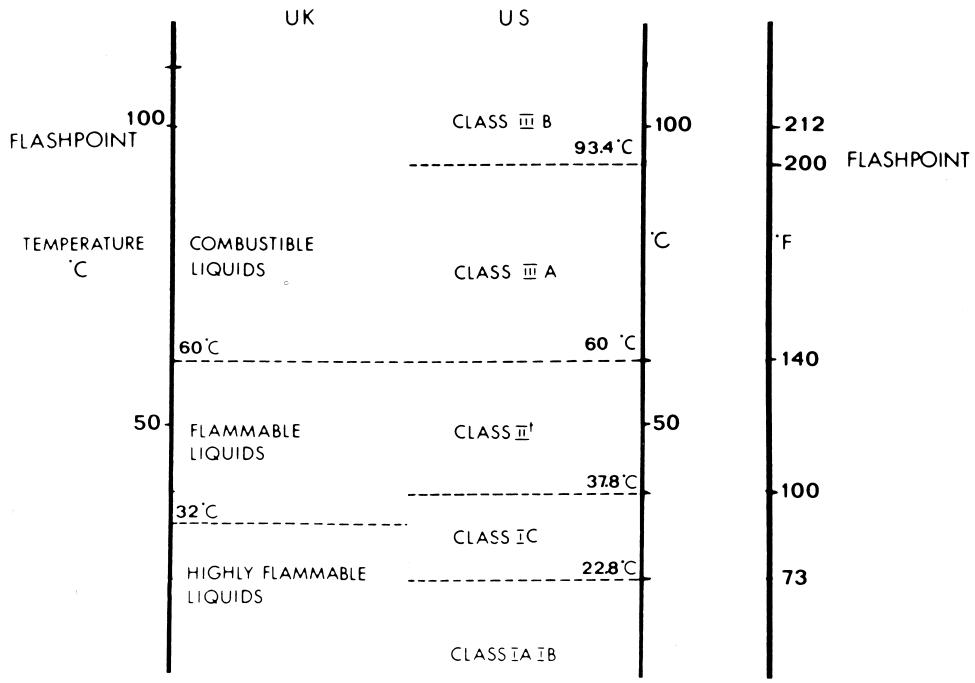


FIGURE 4 A comparison of the U.K. and U.S. classifications of flammable and combustible liquids. [From Drysdale, D. (1998). "Introduction to Fire Dynamics," John Wiley and Sons, Chichester.]

However, unlike liquids, uniform bulk heating of a solid to the required temperature is not possible experimentally, and the firepoint (or flashpoint) must be determined as a surface temperature achieved as a result of the imposed external heat flux.

Thus, piloted ignition of a condensed fuel can be regarded as involving several identifiable steps. In the first, the fuel, or fuel surface, is heated to a temperature at which sufficient flammable vapors are being evolved to produce a flammable vapor–air mixture close to the surface. In the second, a "pilot" initiates a premixed flame, which will propagate through the volume in which a flammable mixture exists and will extinguish when this mixture has been consumed. The stage that follows is the establishment of the diffusion flame: if the fuel temperature is at or only slightly above the firepoint, the flame will be very weak but will strengthen and grow in size as heat is transferred from the flame to the surface, whose temperature will then begin to rise, increasing the flow of fuel vapors to the flame. At higher surface temperatures, ignition of the vapor–air mixture near the surface will immediately give rise to a strong diffusion flame, whose vigor will increase with temperature. These features are summarized in Fig. 5, in which it is tacitly assumed that the source of energy responsible for the initial heating of the fuel is independent of the source of "pilot ignition," although in fact this need not be the case. For example, a flame in direct contact with a combustible material will act both as the primary source of energy and the "pilot."

A pool of combustible liquid (flashpoint $>60°C$) is difficult to ignite by direct application of flame. This is because surface-tension-driven flows dissipate the heat from the point of flame contact, and a substantial volume of liquid has to be heated before flame can establish itself at the surface. On the other hand, if the same liquid is absorbed as a thin film on a porous material, then surface-tension-driven flows are inhibited, and direct application of a flame (e.g., by a match) gives very rapid ignition and flame stabilization. This is the "wick" effect, best known in association with candles. A pool of a high-flashpoint liquid may be ignited if flame is first established on a wick—such as a rag dipping or floating in the pool—but a considerable amount of time may elapse before flame begins to spread across the surface.

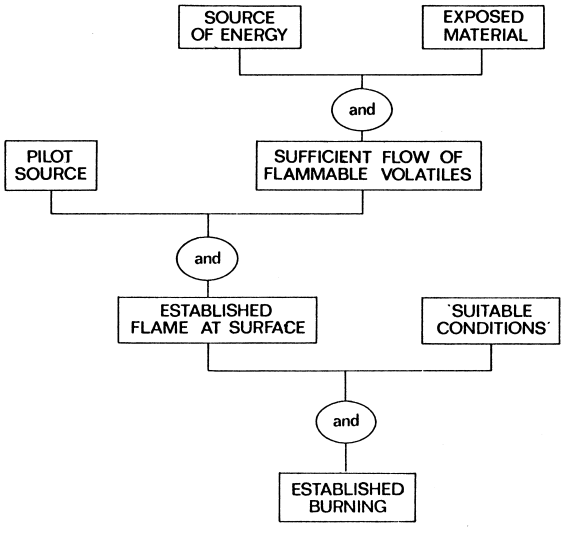


FIGURE 5 The ignition scenario. [From Drysdale, D. (1985). *Cell. Polym.* **4,** 405–419.]

If the temperature of the fuel is high enough, then spontaneous ignition of the vapor–air mixture above the surface can occur. Under these circumstances, the resulting fire will achieve a high rate of burning almost instantaneously, as the rate of production of fuel vapors may already be approaching that for steady burning (Section V). However, while it is possible to define firepoint temperatures for combustible liquids and solids, the fuel temperature at the moment of spontaneous ignition depends not only on the nature of the fuel, but also on the number of factors which relate to the mixing of the hot fuel vapor with the ambient air (including its temperature and flow characteristics).

As with combustible liquids, it is appropriate to attempt to categorize the ignition hazard of combustible solids in terms of their flashpoints or firepoints (i.e., piloted ignition), rather than their potential to ignite spontaneously. As yet, there is no recognized method by which this can be done. The firepoint can be identified by a critical surface temperature (Table IV) or a critical rate of production of

TABLE IV Criteria for Ignition

Material	Fire point[a] (°C)	Critical radiant heat flux (kW/m^2)	
		Pilot	Spontaneous
"Wood"	—	12[b]	28[b]
Chipboard	—	28[c], 18[d]	—
Hardboard	—	27[c]	—
Polymethylmethacrylate	309 ± 6	21[c]	—
Flexible polyurethane foam	~325	16[d]	—
Polyoxymethylene	281 ± 5	17[d]	—
Polymethylene	363 ± 3	12[d]	—
Polyethylene/42% Cl	—	22[d]	—

[a] Surface temperature for piloted ignition (horizontal samples). [H. E. Thomson and D. D. Drysdale, *Fire Mater.* **11,** 163–172.]

[b] General value for wood, vertical samples [Lawson, D. I., and Simms, D. L. (1952). *Br. J. Appl. Phys.* **3,** 288–292.].

[c] Quintiere, J. G. (1981). *Fire Mater.* **5,** 52–60. These values seem high.

[d] From Tewarson, A. (1982). "Flame Retardant Polymeric Materials" (M. Lewin, S. M. Atlas, and E. M. Pearce, ed.), Vol. 3, pp. 97–153. Plenum Press, New York and London.

fuel vapors, although these are difficult to measure and are sensitive to the boundary conditions at the surface. Attempts have been made to define "ease of ignition" in terms of the minimum radiant heat flux necessary for piloted ignition to occur. Figures are quoted in the literature (Table IV), but these are related not only to the material tested, but also to parameters that determine heat losses from the surface, including orientation and thickness of the fuel and the thermal boundary conditions of the experimental configuration. The result can only be used for other configurations if all these factors are taken into account.

It should be noted that heating the surface of a solid produces thermal gradients within the solid that can affect subsequent behavior. If the imposed heat flux is suddenly removed after stabilization of a flame at the surface, extinction can occur if the rate of conductive heat transfer from the surface into the body of the fuel exceeds the rate of heat transfer to the surface from the flame. This will occur particularly under conditions of high heat flux when the surface temperature will rise rapidly, creating large subsurface temperature gradients.

Ease of ignition of solid fuels cannot be defined or quantified easily. It is possible to speak of "thick" and "thin" fuels, but these depend not only on the physical thickness of the fuel bed (τ), but also on the heat transfer conditions at the surface and the thermal properties of the solid. If the Biot number [$\mathrm{Bi} = h\tau/k$, where h is the convective heat transfer coefficient (W/m^2 K) and k is thermal conductivity (W/m^2 K)] is less than 0.1, then there will be no effective temperature gradient between the front and rear faces of the specimen during the transient heating process. Using simple heat transfer theory and assuming that the fuel is an inert solid, it can be shown that the time to achieve the firepoint temperature (t_i) under an imposed heat flux is directly proportional to the thickness of the fuel. For convective heating on both faces

$$t_i = \frac{\tau \rho c}{2h} \ln\left(\frac{T_\infty - T_0}{T_\infty - T_i}\right) \tag{1}$$

where T_∞ is the ambient temperature, which must be greater than the firepoint temperature (T_i), and ρ is the density (kg/m^3). For thick solids, the subsurface temperature will be nonuniform while the solid is being heated, but the thickness of the heated layer can be approximated by writing $\tau = \sqrt{\alpha t}$, where t is the duration of heating and α is the thermal diffusivity of the solid (m^2/sec) [$k/\rho c$, where c is heat capacity (J/kg K)]. By substituting this into Eq. (1), it can be seen that the time to achieve the firepoint is proportional to $\rho c\sqrt{\alpha t_i}$, or

$$t_i \propto k\rho c \tag{2}$$

where the product $k\rho c$ is commonly known as the thermal inertia.

Thermal inertia has a very significant effect on the ease with which combustible solids may be ignited. It is common experience that unless suitably treated, low-density materials such as a standard polyurethane foam can be ignited extremely easily, for example, by a small flame. This is a direct consequence of the low thermal inertia, which permits a rapid surface temperature rise. This is illustrated directly in Fig. 6, which is derived from the equation for the time dependence of the surface temperature of a semi-infinite solid exposed to a convective heat flux.

It is appropriate to emphasize at this point that low-density materials, once ignited, will achieve their maximum rate of burning very rapidly because the surface is effectively insulated from the cooler layers beneath. This is a direct consequence of the low thermal conductivity (Table V). Equally, the ignition behavior of dense materials can be markedly influenced by the rear-face boundary conditions. An extreme example is a "thin" fuel that is firmly attached to a dense substrate, such as wallpaper on plaster. The plaster acts as an effective heat sink and makes the paper difficult to ignite in this situation.

Whenever the ignition of a combustible solid is being considered, it is necessary to examine the heat-loss conditions obtaining at the time of exposure.

A. Spread of Flame

Spread of flame over the surface of a combustible liquid or solid may be regarded as an advancing ignition front in which the leading edge of the flame acts both as the source of heat necessary to raise the temperature of the fuel to the firepoint and as the pilot to ignite the vapors (cf. Fig. 5). With liquid fuels, the mechanism by which the surface ahead of the flame is heated involves surface-tension-driven flows that cause hot liquid to be drawn forward from under the flame, displacing the surface layer of cooler liquid. With solids, the mechanism of heat transfer depends on several factors, including the orientation of the fuel. Figure 7 illustrates how the interaction of the established flame changes with orientation, from vertically downward (a) to vertically upward spread (e). For vertically downward spread over a thin fuel, a steady rate of spread is quickly achieved. Heat transfer is by conduction through the gas phase, and the rate is inversely proportional to the fuel thickness. However, the rate of spread down the surface of a "thick" fuel has been found to depend on heat transfer through the solid and is proportional to the inverse of the thermal inertia. These points are consistent with the observations on ease of ignition incorporated in Eqs. (1) and (2).

With vertically upward spread (Fig. 7e), the flame interacts directly with the fuel above the leading edge of

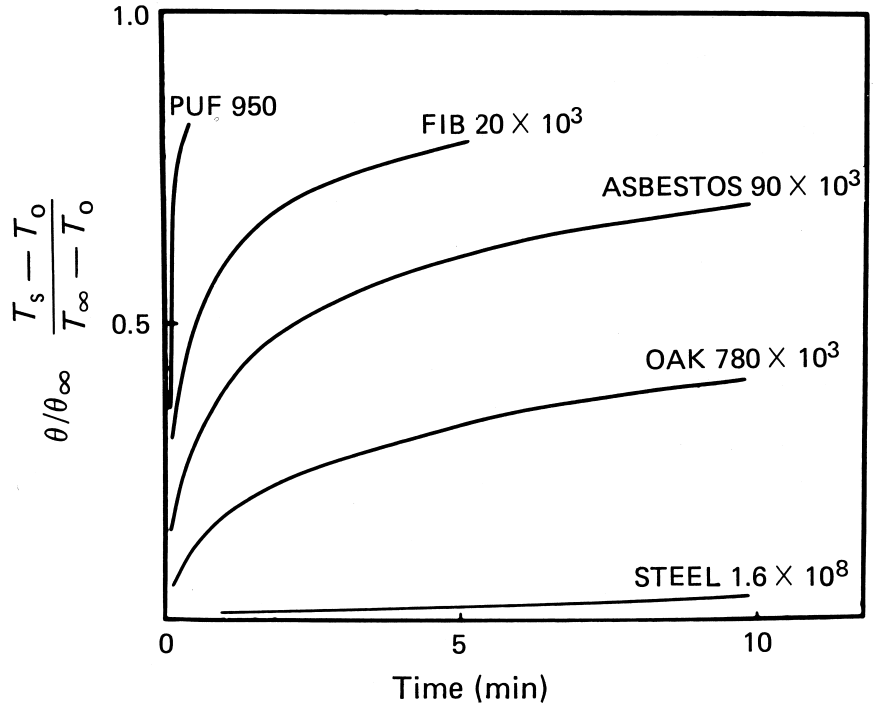


FIGURE 6 Effect of thermal inertia ($k\rho c$) on the rate of temperature rise at the surface of a semi-infinite solid exposed to a uniform convective heat flux. FIB, fiber insulation board; PUF, a typical polyurethane foam; T_s, surface temperature. Figures are values of $k\rho c$ in W^2 sec/m^4 K^2. [From Drysdale, D. (1998). "Introduction to Fire Dynamics," John Wiley and Sons, Chichester.]

the burning zone, filling the boundary layer and providing high rates of heat transfer. The rate of spread is greatly enhanced, in proportion to the height of the flame. For thin fuels, with flame spread on both sides and a relatively short burning zone, high rates of spread are achieved very quickly. Upward flame spread on thick fuels, on the other hand, develops much more slowly as there is a significant delay between ignition and achieving the maximum rate of burning at any one point (see Section V). In this case, however, the rate of spread will continue to increase as long as the length of the burning zone (which determines the flame height) continues to grow.

Upward spread is clearly dominated by the buoyant flow of hot, burning gases ahead of the flame front. This effect is important at angles of inclination significantly less than 90°; indeed, it first becomes noticeable for thick fuel beds

TABLE V Thermal Properties of Solids

Material	k (W/m K)	c_p (J/kg K)	ρ (kg/m³)	α (m²/sec)	$k\rho c$ (W^2 sec/m^4 K^2)
Brick (common)	0.69	840	1600	5.2×10^{-7}	9.3×10^5
Polymethylmethacrylate	0.19	1420	1190	1.1×10^{-7}	3.2×10^5
Polyethylene					
(low density)	0.35	1900	940	2.0×10^{-7}	6.2×10^5
(high density)	0.44	2300	970	2.0×10^{-7}	9.8×10^5
Polystyrene	0.11	1200	1100	8.3×10^{-8}	1.5×10^5
Oak	0.17	2380	800	8.9×10^{-8}	3.2×10^5
Yellow pine	0.14	2850	640	8.3×10^{-8}	2.5×10^5
Fiber insulating board	0.041	2090	229	8.6×10^{-8}	2.0×10^4
Polyurethane foam[a]	0.034	1400	20	1.2×10^{-6}	9.5×10^2
Polystyrene foam[a]	0.034	1520	20	1.1×10^{-6}	1.0×10^3

[a] Typical values.

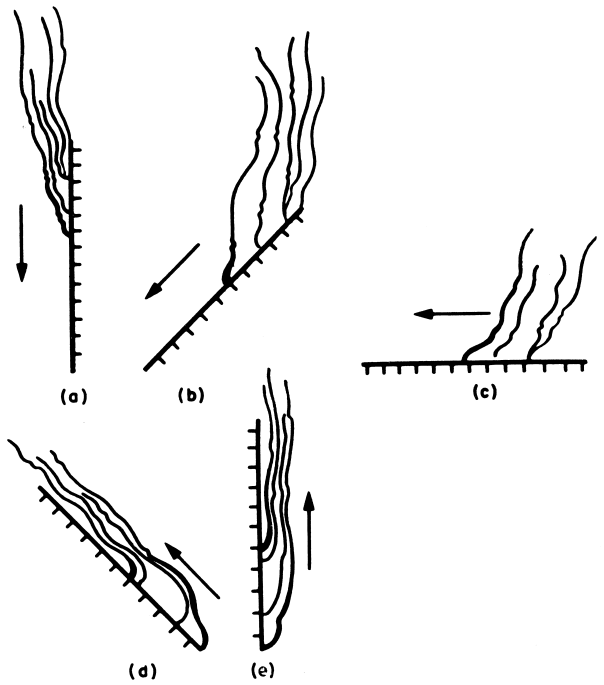


FIGURE 7 Interaction between the flame and the fuel for flame spread at different angles of inclination. (a) −90°; (b) −45°; (c) 0°; (d) +45°; (e) +90°. [From Drysdale, D. (1985). "Introduction to Fire Dynamics," John Wiley and Sons, Chichester.]

as an increase in the rate of flame spread at an inclination of around 20° from the horizontal and is greatly enhanced if air is prevented entering from the sides by the presence of sidewalls. This is shown in Fig. 8 where the rate of (upward) flame spread is plotted as a function of the angle of inclination of the surface: the influence of sidewalls is very marked. (This is the process which is believed to have been responsible for the very rapid spread of fire on a wooden escalator at the King's Cross Underground Station in London on November 18, 1987.) Note that with thin fuels, provided that there is no restriction to flow on the underside of the material, the rate of flame spread increases monotonically as the inclination is increased from the horizontal (0°).

The rate of spread of flame is also increased if there is an external source of heat, such as radiation from a nearby fire. Indeed, during the early stages of a compartment fire, radiation levels gradually increase, enhancing not only the rate of burning of items already involved (see Section V), but also the rate of flame spread over combustible surfaces. Unless there is direct flame contact, spread of flame to an adjacent item normally requires that radiation levels have increased sufficiently for spontaneous ignition to occur. The importance of radiation in flame spread through other discontinuous fuel beds such as wood cribs, bush, and forests is well known.

V. STEADY BURNING OF COMBUSTIBLE LIQUIDS AND SOLIDS

The rate of burning of a condensed fuel is controlled by the net rate of heat transfer to the fuel surface. Expressing the rate of burning as a mass flux $\dot{m}''$ (g/m^2 sec),

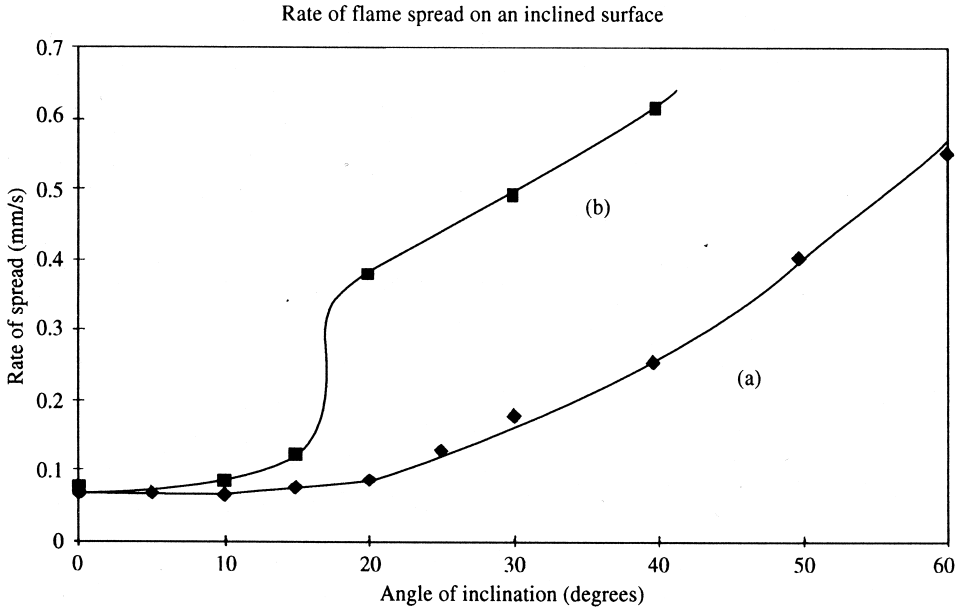


FIGURE 8 Rate of spread of flame on an inclined surface. 60-mm wide samples of poly(methyl methacrylate) with (■) and without (♦) "sidewalls." [From Drysdale, D. (1998). "Introduction to Fire Dynamics," John wiley and Sons, Chichester.]

TABLE VI Limiting Regression Rates for Liquid Pool Fires[a]

Liquid fuel	Limiting regression rate		$\dot{m}''$ (g/m^2 sec)
	R (mm/min)	R (m/s)	
LNG[b]	6.6	1.1×10^{-4}	51.3
n-Butane	7.9	1.3×10^{-4}	79.1
n-Hexane	7.3	1.2×10^{-4}	80.3
Xylene	5.8	0.97×10^{-4}	85.1
Methanol	1.7	0.28×10^{-4}	22.4

[a] From Burgess, D. S., Strasser, A., and Grumer, J. (1961). *Fire Res. Abstr. Rev.* **3,** 177–192.
[b] Liquefied natural gas.

this can be written

$$\dot{m}'' = \left(\dot{Q}''_{\mathrm{F}} + \dot{Q}''_{\mathrm{E}} - \dot{Q}''_{\mathrm{L}}\right)/L_{\mathrm{v}} \tag{3}$$

where $\dot{Q}''_{\mathrm{F}}$ is the heat flux (kW/m^2) from the flame to the fuel, $\dot{Q}''_{\mathrm{E}}$ is the heat flux originating from an external heat source, $\dot{Q}''_{\mathrm{L}}$ is the rate of heat loss expressed as a flux through the surface, and L_{v} is the heat required (kJ/g) to produce the volatiles (heat of gasification).

Early studies of fires involving pools of liquid fuels revealed that the mode of heat transfer from the flame to the fuel surface (i.e., $\dot{Q}''_{\mathrm{F}}$) changes from mainly convective to predominantly radiative as the pool diameter is increased from approximately 0.1 m to 1.0 m. Indeed, the evidence suggests that there is a limiting, radiation-controlled burning rate for pools of diameter greater than about 1 m. This occurs because the effective flame thickness (as viewed from the surface) increases with size of the fuel bed, and consequently the emissivity (ε) will increase toward a maximum value of unity, at least for luminous flames:

$$\varepsilon = 1 - \exp(-Kl) \tag{4}$$

where K is an effective emission coefficient (m^{-1}) and l is the thickness of the flame (m) as viewed from the surface: this will be related to the fuel-bed diameter [see Eq. (7) and following]. Limiting rates for some typical liquid fuels are shown in Table VI, expressed as regression rates (m/sec, equivalent to m^3/m^2 sec) and as mass fluxes (g/m^2 sec). The rates of burning of the hydrocarbon fuels are similar, while that for methanol is substantially less. There are two contributory factors here: (1) L_{v} for methanol is much greater than that for the hydrocarbons, and (2) the effective emission coefficient of the methanol flame is small (see Table IX).

The heat-loss term ($\dot{Q}''_{\mathrm{L}}$) includes radiative losses from the surface and conductive losses into the bulk of the fuel. To achieve steady-state burning, a quasi-equilibrium must exist. This is rarely achieved: as the fuel burns away, the boundary conditions will change and a stage will be reached when heat losses from the liquid to the "base" of the container become significant and influence the rate of burning. With high flashpoint liquids, burning will not be sustained if these heat losses become so great that the fuel surface temperature falls below T_i, the firepoint temperature. The decrease in $\dot{m}''$ is accompanied by a reduction in $\dot{Q}''_{\mathrm{F}}$ as the flame becomes weaker and its size diminishes: a point will eventually be reached when $\dot{m}'' < \dot{m}''_{\mathrm{cr}}$, the critical flowrate of the volatiles at the firepoint. As an example, consider the burning of a layer of oil floating on water. If the firepoint of the oil is significantly greater than the temperature of the water, heat losses to the water will reduce the rate of burning and will cause the flame to self-extinguish when the layer is less than a few millimeters thick.

The same principles apply to the burning of solids, both in the pool configuration and other orientations. Tewarson has made a careful study of the burning of combustible solids, basing the analysis of his data on Eq. (3). This enabled him to deduce values of L_{v} for a range of materials under steady-burning conditions in his apparatus by varying an externally applied heat flux ($\dot{Q}''_{\mathrm{E}}$) while maintaining the other experimental parameters constant. The results gave, *inter alia*, values of $\dot{Q}''_{\mathrm{F}}/L_{\mathrm{v}}$, which Tewarson called the "ideal burning rate," and proposed it for use as a measure of potential burning intensity. Typical values are given in Table VII, although they are strictly applicable only to Tewarson's experimental situation.

These results may be used to illustrate why some materials can burn in isolation while others require an external heat flux: Tewarson was also able to estimate $\dot{Q}''_{\mathrm{L}}$ and found that for some, $\dot{Q}''_{\mathrm{L}} \geq \dot{Q}''_{\mathrm{F}}$, indicating that without an external heat source, $\dot{m}''$ in Eq. (3) would be negative or zero.

TABLE VII Flammability Parameters[a]

Combustibles	L_{v} (kJ/g)	$\dot{Q}''_{\mathrm{F}}$ (kW/m^2)	$\dot{Q}''_{\mathrm{L}}$ (kW/m^2)	$\dot{m}''_{\mathrm{ideal}}$ (g/m^2 sec)
FR phenolic foam (rigid)	3.74	25.1	98.7	11
Polyoxymethylene (solid)	2.43	38.5	13.8	16
Polyethylene (solid)	2.32	32.6	26.3	14
Polypropylene (solid)	2.03	28.0	18.8	14
Wood (Douglas fir)	1.82	23.8	23.8	13
Polystyrene (solid)	1.76	61.5	50.2	35
Polymethylmethacrylate (solid)	1.62	38.5	21.3	24
Polyurethane foam (rigid)	1.52	68.1	57.7	45
Polyurethane foam (flexible)	1.22	51.2	24.3	32
Methyl alcohol (liquid)	1.20	38.1	22.2	32
Styrene (liquid)	0.64	72.8	43.5	114
Methylmethacrylate (liquid)	0.52	20.9	25.5	76
Heptane	0.48	44.3	30.5	93

[a] From Tewarson, A., and Pion, R. F. (1976). *Combust. Flame* **26,** 85–103.

There are several factors that can give rise to this situation. Heavily fire-retarded materials may release volatiles of low flammability (low heat of combustion), while others decompose to yield a layer of char that will provide protection for the bulk of the fuel below. A flow of volatiles sufficient to support flame will only be maintained at a high surface temperature, which will be accompanied by significantly greater radiative heat loss. Wood in its natural state is a char-forming material. The combination of high surface temperature, of the accompanying large temperature gradients within the material, and of the fact that the volatiles are of relatively low flammability is responsible for its unusual fire properties. Thus a "massive" piece of wood, such as a log, cannot be made to sustain flaming combustion in isolation. However, when built up with other logs to make a fire, cross radiation between burning surfaces supplies the "external" heating ($\dot{Q}''_{\mathrm{E}}$) necessary for vigorous burning. A thin piece or splint of wood, on the other hand, can be ignited and will continue to burn in isolation because internal temperature gradients will quickly be reduced to a level at which the heat losses to the interior of the material are no longer significant.

Equation 3 describes the rate of release of fuel vapors ($\mathrm{g/m^2 s}$) which in turn determines the rate at which heat is released ($\dot{Q}_c$ (kW)). As will become clear, the rate of heat release of a fire is a very important parameter and a number of significant features of the fire process may be correlated with it. It may be expressed as

$$\dot{Q}_{\mathrm{c}} = \dot{m}'' \chi \Delta H_{\mathrm{c}} A_{\mathrm{F}} \tag{5}$$

where ΔH_{c} is the heat of combustion (kJ/g) of the volatiles, χ is an efficiency factor to take into account the fact that combustion is rarely complete, and A_{F} is the area ($\mathrm{m^2}$) of the fuel bed. (Tewarson has reported that χ lies between 0.7 and 0.4, the lower values being associated with materials containing a high proportion of halogen in the form of flame retardant.) Substituting for $\dot{m}''$ [from Eq. (3)], but writing

$$\dot{Q}''_{\mathrm{net}} = \dot{Q}''_{\mathrm{F}} + \dot{Q}''_{\mathrm{E}} - \dot{Q}''_{\mathrm{L}}$$

we have

$$\frac{\dot{Q}_{\mathrm{c}}}{A_{\mathrm{F}}} = \dot{Q}''_{\mathrm{net}} \chi \left(\frac{\Delta H_{\mathrm{c}}}{L_{\mathrm{v}}} \right) \tag{6}$$

The ratio $\Delta H_{\mathrm{c}}/L_{\mathrm{v}}$ can be calculated from available experimental data using the heat of combustion of the solid. This ranges in value from 3 (for red oak) to 90 (for liquid heptane) (Table VIII). While materials may be ranked according to this ratio, its value as a means of selecting materials is somewhat limited as it refers to the steady-burning condition. Thus, methanol has a relatively low value, while a certain flexible polyurethane foam (PRC no. 25) is found to have a value almost 20% less than solid polyethylene. Yet it is known that both the methanol and the cellular polymer represent significant fire hazards on account of their ease of ignition and the rate at which steady-burning conditions can be achieved (Section IV). This point may be used to emphasize the importance of the growth period in a compartment fire when life safety is the prime consideration (Section VII).

This ranking also neglects χ and gives no indication as to how the rate of heat release may depend on an imposed heat flux. Fortunately, direct measurement of rate of heat release is possible making use of the fact that for every gram of oxygen consumed in combustion, a known amount of heat is released (3 kJ/g oxygen), provided that all the oxygen is converted to carbon dioxide and water vapor. (Corrections can be made for incomplete combustion.) Thus, by measuring the rate of oxygen consumption in a fire—which can be done using a large-scale test facility such as the Nordtest apparatus—the rate of heat release can be calculated. The method is also used in the which is a laboratory scale apparatus used to determine (among other things) the rate of heat release of a combustible material by burning small samples as they are exposed to a range of radiant heat fluxes.

TABLE VIII $\Delta H_{\mathrm{c}}/L_{\mathrm{v}}$ Values for Fuels[a]

Fuel[b]	$\Delta H_{\mathrm{c}}/L_{\mathrm{v}}^{c}$
Red oak (solid)	2.96
Polyoxymethylene (granular)	6.37
PVC (granular)	6.66
Polyethylene 48% Cl (granular)	6.72
Flexible PU foam (27)	12.26
Flexible PU foam (21)	13.34
PMMA (granular)	15.46
Methanol (liquid)	16.50
Flexible PU foam (25)	20.03
Rigid polystyrene foam (47)	20.51
Polypropylene (granular)	21.37
Polystyrene (granular)	23.04
Polyethylene (granular)	24.84
Styrene (liquid)	63.30
Heptane (liquid)	92.83

[a] From Tewarson, A. (1980). *Fire Mater.* **4,** 185–191.

[b] Numbers in parentheses are sample numbers of Products Research Committee, 1980 (PRC).

[c] ΔH_{c} measured in an oxygen bomb calorimeter and corrected for water as a vapor for fuels for which data are not available; L_{v} is obtained by measuring the mass loss rate of the fuel in pyrolysis in N_2 environment as a function of external heat flux for fuels for which data are not available. Note: If ΔH_{c} is replaced by the heat of combustion of the volaties ($\Delta H_{\mathrm{c}} + L_{\mathrm{v}}$), then the combustibility ratios are increased by 1.00 and the ranking order is unchanged.

The fire behavior of combustible solids can be modified by the incorporation of fire retardants. These can alter ΔH_c and/or L_v, but may also influence χ, $\dot{Q}''_F$, and $\dot{Q}''_L$. Surprisingly, application of fire retardants based on phosphorus and boron to wood decreases the effective L_v, but this is accompanied by a significant decrease in the heat of combustion of the vapors released.

Most of the above remarks refer to experimental studies carried out on horizontal samples, but burning at vertical surfaces must be considered. The principal difference between the two lies in the relationship between the flame and the surface. For a horizontal fuel bed, buoyancy assures that the flame will assume a position directly above the burning surface, while with a vertical surface, the flame will hug the fuel, filling the boundary layer as illustrated in Fig. 9. Experimental studies have shown the burning rate to vary significantly with height above the bottom edge. However, as above, the hazard is perhaps not so much the steady-burning rate, but the rapidity with which it can be achieved following ignition near the base. High flame spread rates will ensure rapid engulfment of the entire surface, which, in a real situation (e.g., flammable wall linings in conjunction with a low ceiling), can lead to very rapid fire development in a compartment (Section VII).

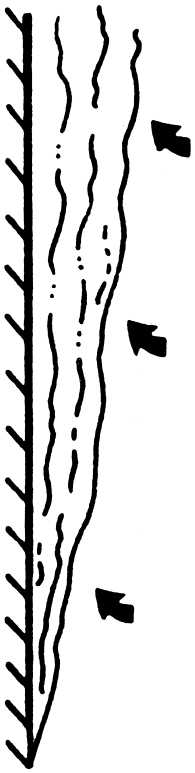


FIGURE 9 Illustration of burning at a vertical surface. The arrows represent air entrainment. [From Drysdale, D. (1998). "Introduction to Fire Dynamics." John Wiley and Sons, Chichester.]

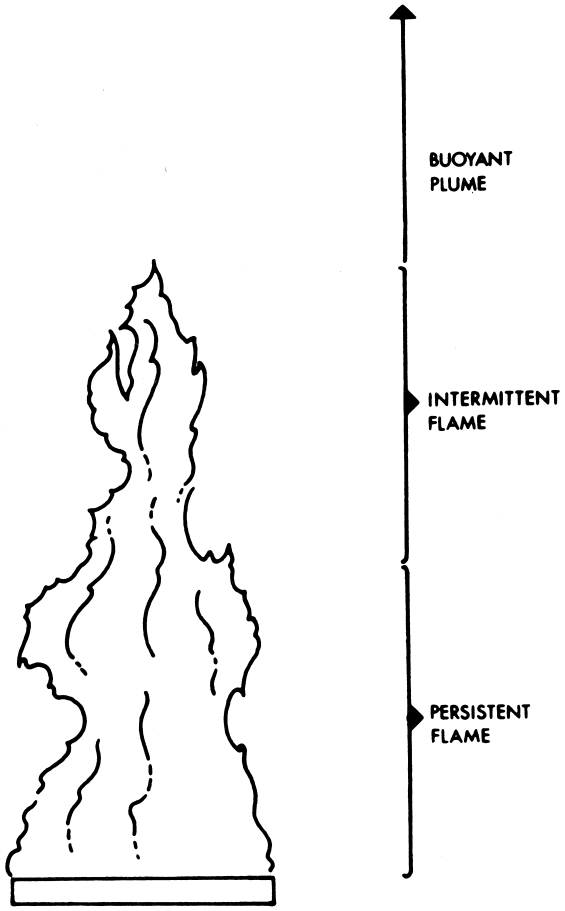


FIGURE 10 Diagram of the fire plume, showing the three regimes. [From Drysdale, D. (1998). "Introduction to Fire Dynamics," John Wiley and Sons, Chichester.]

VI. THE FIRE PLUME

In a natural fire, the behavior of the flame and the movement of the hot combustion products is determined by buoyancy forces due to differences in density between the fire plume and the surrounding air. The fire plume consists of three regions, the buoyant diffusion flame occupying the lower two and the "buoyant plume" of hot combustion products and entrained air occupying the third (Fig. 10). The upper part of the flame is intermittent, flickering or pulsating with a frequency that decreases as the diameter of the fuel bed increases. Most of the combustion occurs in the lower part of the flame, particularly in those locations where there is efficient mixing of fuel vapor and entrained air. However, the combined process of combustion and air entrainment introduces instabilities that create vortexlike structures that are shed, still burning, from the upper portion of the persistent flame: these create the appearance of intermittency referred to above.

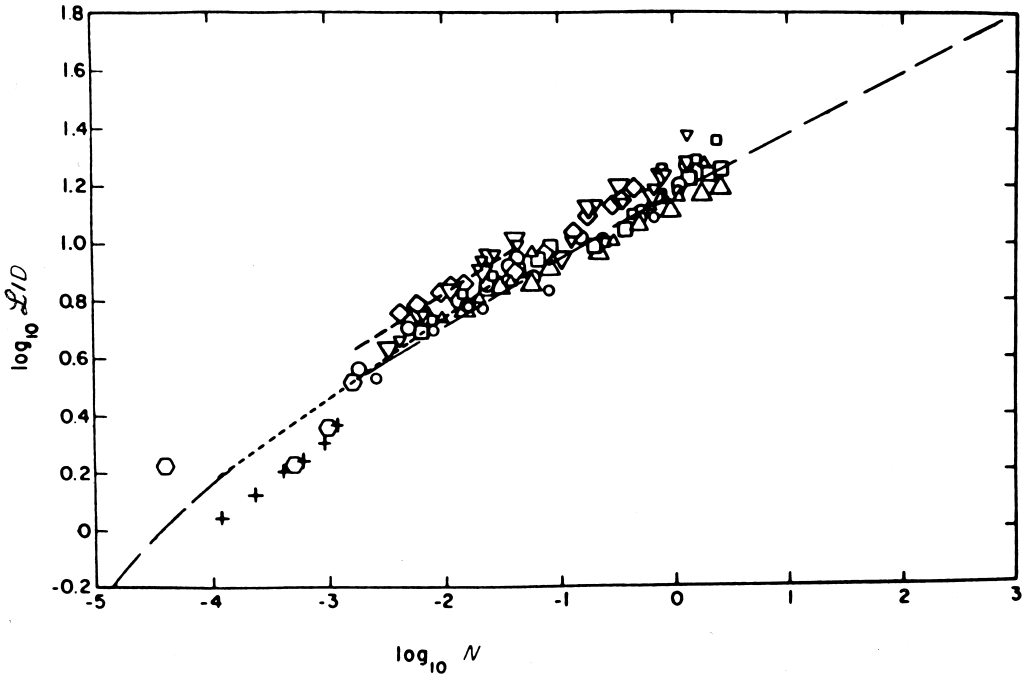


FIGURE 11 Correlation of flame height data from various sources. [From Heskestadt, G. (1983). *Fire Safety J.* **5,** 103–108.]

McCaffrey has shown that in the region of persistent flaming, the gases on the centerline of a diffusion flame above a 0.3-m^2 porous burner are accelerating, while in the intermittent region, the upward velocity is approximately constant. In the buoyant plume there is a slow deceleration, which is consistent with early theoretical studies of the buoyant plume rising from a point source.

The height of the buoyant diffusion flame is not easy to define or measure. Fundamentally, it is determined by the rate of production of fuel vapor and the rate at which air is entrained into the rising column of burning gas, however, it has been found that the amount of air entrained up to the flame tip greatly exceeds that required for complete combustion of the fuel vapors. The rate of entrainment is determined by the upward velocity of the flame gases, which in turn is dependent on the buoyancy of the flame gases and the shear forces generated at the flame boundary. These generate the instability that causes the "flickering" of the intermittent flame. Visual observation of the flame height averages out these fluctuations and the tip of the flame as judged by eye corresponds approximately to the height below which flame is present for at least 50% of the time. The average temperature at this height is between 500 and 600°C, falling to 300–400°C at the maximum reach of the intermittent flames.

Analyses have shown that the ratio of the flame height (l) to fuel-bed diameter (D) is a function of the ratio $(\dot{Q}_c)^2/D^5$, where $\dot{Q}_c$ is the rate of heat release. Heskestad correlated data from a large number of sources using the equation

$$l/D = 15.6N^{1/5} - 1.02 \tag{7}$$

in which N is a dimensionless number derived from a modified Froude number and is given by

$$N = \left[\frac{c_p T_\infty}{g\rho_\infty^2(\Delta H_c/r)^3}\right]\frac{\dot{Q}_c^2}{D^5} \tag{8}$$

The correlation is shown in Fig. 10. Note that as the terms within the brackets in Eq. (8) are known, Eq. (7) can be simplified to

$$l = 0.23\dot{Q}_c^{2/5} - 1.02D \tag{9}$$

Data may also be correlated using another form of the Froude number, namely,

$$\dot{Q}^* = \frac{\dot{Q}_c}{\rho_\infty c_p T_\infty \sqrt{gD}D^2} \tag{10}$$

Heskestad's correlation then reduces to

$$l/D = 3.7\dot{Q}^{*2/5} - 1.02 \tag{11}$$

and applies to the range $0.12 < \dot{Q}^* < 1.2 \times 10^4$. The modified Froude number Q^* is most commonly used in the fire research community to express rate of heat release in a nondimensional form.

TABLE IX Radiation Properties of Flames above 0.3-m-Diameter Pool Fires[a]

Fuel	Flame temperature (°C)	Emissivity (ε)
Alcohol	1218	0.066
Petrol	1026	0.36
Kerosene	990	0.37
Benzene	921	
After 2 min		0.59
After 5 min		0.70
After 8 min		0.72

[a] From Rasbash, D. J., Rogowski, Z. W., and Stark, G. W. V. (1956). *Fuel* **31,** 94–107.

The temperature of large diffusion flames are quoted as lying in the region of 800–1200°C, depending on the nature of the fuel and the conditions of burning (Table IX). These are averages and do not refer to localized reaction zones within the flame envelope, where much higher temperatures will exist. The apparent temperature depends on the amount of excess air entrained into the flame and the rate of heat loss by radiation, which can be as high as 30–40% for highly emissive (i.e., sooty) flames. Nonluminous flames, such as those from methanol and polyoxymethylene, lose a relatively small proportion of energy by radiation, and consequently their average flame temperatures are significantly higher.

McCaffrey's data on the temperature and velocity of the flame gases on the centerline of a buoyant methane diffusion flame are summarized in Table X. The upward velocity is a significant factor in determining the effectiveness of sprinklers in controlling large fires. If the fire has grown too large by the time the sprinkler begins to discharge, the downward momentum of the spray or the terminal velocity of the large drops may be insufficient to penetrate the flame and reach the burning surfaces.

In the open, a flame may be subjected to the effects of a wind that will cause flame deflection, to an extent depending on the size of the fire (i.e., the upward velocity of the flame gases) and the velocity of the wind. Some studies of this effect have been carried out, but a rough rule of thumb that is used in the petrochemical industry is that a 2 m/sec wind will deflect a flame of a large pool fire by 45°. The possible consequences of this within the confines of a chemical plant, for example, could be extremely serious, as items downwind could become exposed either to direct flame impingement or to increased radiant heat flux levels.

TABLE X Summary of Centerline Data for a Buoyant Methane Diffusion Flame on 0.3-m-Square Porous Burner[a] (Fig. 10)

Centerline velocity: $$\frac{u_0}{Q^{1/5}} = k\left(\frac{z}{Q^{2/5}}\right)^{\eta}$$

Centerline temperature: $$\frac{2g\Delta T_0}{T_0} = \left(\frac{k}{C}\right)^2\left(\frac{z}{Q^{2/5}}\right)^{2\eta-1}$$

Region	k	η	$z/Q^{2/5}$ (m/kW$^{2/5}$)	C
Flame	6.8 m$^{1/2}$/sec	$\frac{1}{2}$	<0.08	0.9
Intermittent	1.9 m/kW$^{1/5}$ sec	0	0.08 – 0.02	0.9
Plume	1.1 m$^{4/3}$/kW$^{1/3}$ sec	$-\frac{1}{3}$	>0.2	0.9

[a] From McCaffrey, B. J. (1979). National Bureau of Standards, NBSIR 79-1910.

The above remarks refer to an unconfined fire plume in which air has access from all sides and there is no restriction to the upward buoyant flow. While the earliest stages of a fire in a compartment may approximate to this, provided that the burning item is not against a wall or is significantly deflected by the air inflow through the ventilation opening, a stage will be reached when the fire plume impinges on the ceiling and the flow of hot gases will be deflected horizontally as a ceiling jet, as illustrated in Fig. 12. This is the mechanism by which smoke and hot gases are transported from the fire to ceiling-mounted fire detectors. A sprinkler is actuated only after a sprinkler head has been exposed to a stream of hot gas. Alpert correlated the data from a large number of experimental steady-state fires burning under an extensive flat ceiling (of height H) and showed that the maximum, gas temperature (T_{max}) at a radial distance r from the centerline of the plume could be described by the equations

$$\text{for} \quad r > 0.18H \quad T_{max} - T_0 = \frac{5.38(\dot{Q}_c/r)^{2/3}}{H} \tag{12}$$

$$\text{for} \quad r > 0.18H \quad T_{max} - T_\infty = \frac{16.9\left(\dot{Q}_c^{2/3}\right)}{H^{5/3}} \tag{13}$$

where $\dot{Q}_c$ (kW) is the rate of heat release, or fire intensity. (The latter equation refers to the area where the plume impinges on the ceiling.) Equivalent correlations were derived for the flow velocities under the ceiling. These equations from the basis for Appendix C of NFPA 72E, *Standard on Automatic Fire Detectors* (National Fire Protection Association, 1984).

If the fire is next to a wall, or in a corner, then the subceiling temperatures will be enhanced as a result of the restrictions on air entrainment into the fire plume. Approximate corrections for these changes may be incorporated into Eqs. (12) and (13) by multiplying $\dot{Q}_c$ by 2 and 4, respectively.

Direct interaction between the flame and the compartment boundaries causes significant changes in flame behavior. If the burning item is against a wall, then air can only be entrained through 180° rather than 360°, and the flame will tend to be deflected by the momentum of the inflowing air, as illustrated in Fig. 13.

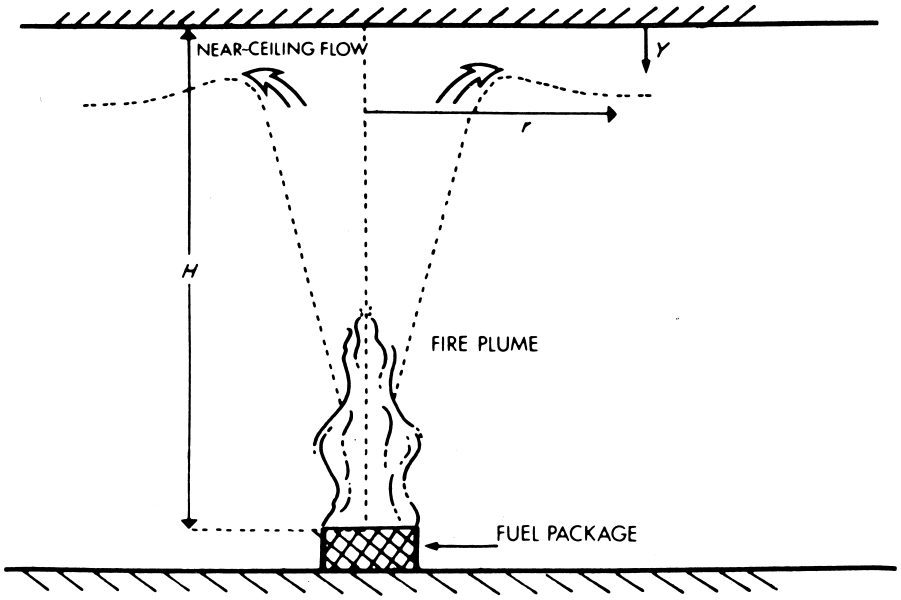


FIGURE 12 The fire plume and its interaction with a ceiling. [From Drysdale, D. (1998). "Introduction to Fire Dynamics," John Wiley and Sons, Chichester.]

For a flame against a plain noncombustible wall, the height is virtually unchanged, but because less air has been entrained due to the presence of the wall, the temperatures in the buoyant flow are greater. The effect is more pronounced if the burning item is in a corner.

The effect on the flame occurs regardless of the combustibility of the wall lining material. However, if the lining is combustible the type of fire illustrated by Fig. 13 represents a very effective means by which the wall lining can be ignited. Rapid fire growth with greatly enhanced flame extension can be anticipated, particularly in a confined or semiconfined situation (e.g., within a vertical duct or in a corner, respectively) when rates of heat transfer to the burning surface can become very high.

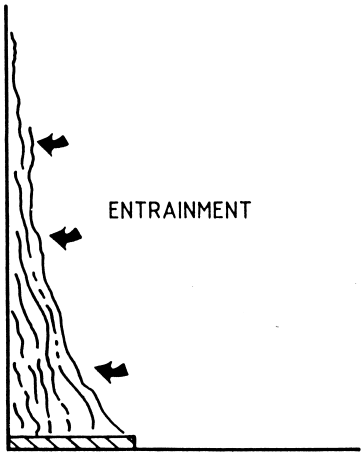


FIGURE 13 Flame deflection near a vertical surface. [From Drysdale, D. (1998). "Introduction to Fire Dynamics," John Wiley and Sons, Chichester.]

If the ceiling is sufficiently low, or the fire sufficiently large, then the flame will impinge on the ceiling and be deflected horizontally, undergoing extension in the process as air entrainment in this configuration (a layer of hot gas flowing over relatively cold gas) is inefficient. There has been only limited study of this effect. Babrauskas has reviewed the available data and suggested that the horizontal extension may be related to the "cut-off height" of the flame, that is, $(l - H)$, where l is the flame height in the unconfined situation [e.g., calculated from Eq. (7)] and H is the ceiling height, but the correlation is not simple, depending on the ceiling configuration: for example, the extension of the cut-off height may be up to 12-fold if the fire is at the closed end of a corridor. Hinkley and co-workers studied the extension of town gas flames under model corridor ceilings and were able to distinguish between "air rich" and "fuel rich" horizontal flames. The latter exhibited considerable extension, representing a fuel-rich gas flowing down the corridor with flaming at the interface with the cooler air below. This is observed with noncombustible ceiling linings and the extension of flaming will be even greater if the lining is combustible. This greatly extends and enhances the thermal radiation flux at floor level as well as the heat transfer to the ceiling and the upper walls.

VII. FIRES IN COMPARTMENTS

The term "compartment fire" is commonly used to describe a fire in a room or "compartment" within a building. If allowed to burn itself out, then assuming sufficient

fuel and ventilation, such a fire will progress through three main stages:

1. The growth, or preflashover, stage, when the fire is localized in the vicinity of its origin and the average temperature in the compartment is still relatively low.
2. The fully developed, or postflashover, fire, during which all the combustible items are involved and flames appear to fill the entire volume and emerge from the ventilation openings.
3. The decay period, when the fire begins to die down as the fuel is consumed. This is formally identified as the period after the average temperature has fallen to 80% of the maximum achieved during stage 2.

These stages are indicated in Fig. 14 in which the size of the fire (defined in terms of the rate of heat release (kW)) is shown as a function of time. The *average* temperature in the lower part of the compartment will remain low during the growth period, although a layer of hot smoke will accumulate under the ceiling and will increase in depth and temperature as the fire grows in size. The development of this layer is the key factor in the way in which a fire progresses to the fully developed stage. However, it is possible for a fire to self-extinguish if the item first ignited burns out before other items can become involved, or to be held in check if it becomes starved of air (oxygen) (Fig. 14). This can occur in a space where there is inadequate ventilation as will occur if the doors and windows are closed and well fitting.

This scenario is potentially hazardous for fire fighters and others. Under conditions of oxygen starvation, the fuel can continue to burn, slowly and inefficiently, producing flammable vapors which will accumulate within the compartment. If a door is opened (or a window breaks), there will be an inrush of fresh air and outflow of hot fire products. The fire will increase in intensity and the flammable vapors trapped in the compartment will burn vigorously as they mix with the air, expelling burning gases through the opening producing a fireball capable of causing serious injury to anyone in its path. This phenomenon is known as "backdraft."

If the compartment has "adequate" ventilation and a sufficient and well-distributed supply of combustible material, a fire once ignited is capable of developing to the fully developed stage.

The transition from the growth period to the fully developed fire is referred to as flashover. Anyone who has not escaped from the compartment before flashover is unlikely to survive, as high temperatures and intense fluxes of radiant heat will develop very quickly, and high levels of toxic and noxious gases will be present. Individuals occupying other parts of the same building are likely to be threatened directly after flashover has occurred in the room

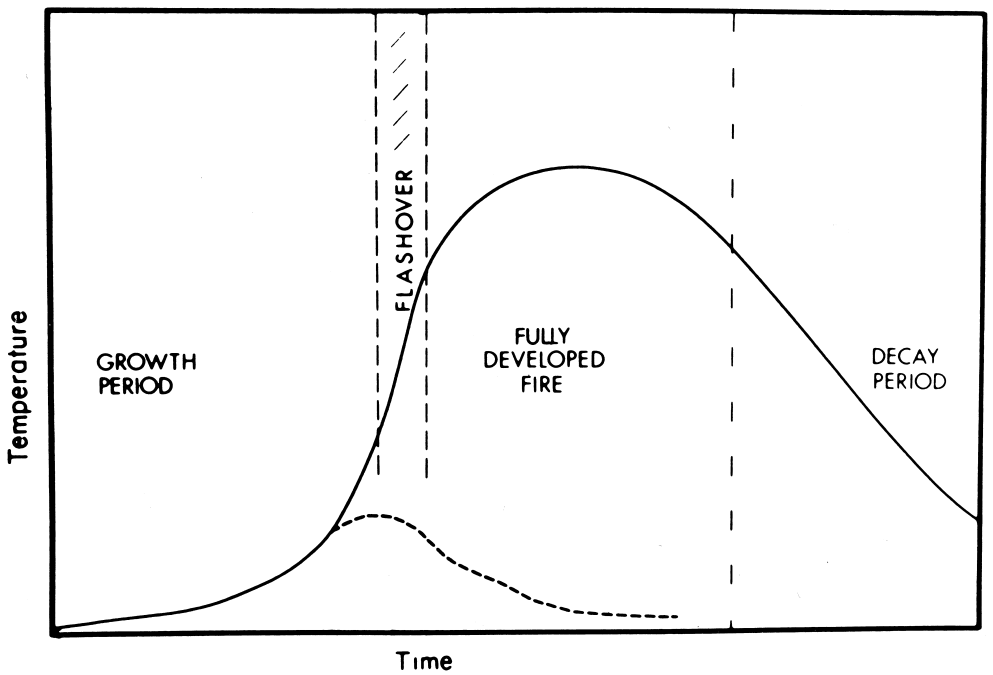


FIGURE 14 The course of a compartment fire, expressed as an average gas temperature as a function of time. The broken line can represent depletion of fuel before flashover has commenced, or depletion of oxygen due to very poor ventilation. [From Drysdale, D. (1998). "Introduction to Fire Dynamics," John Wiley and Sons, Chichester.]

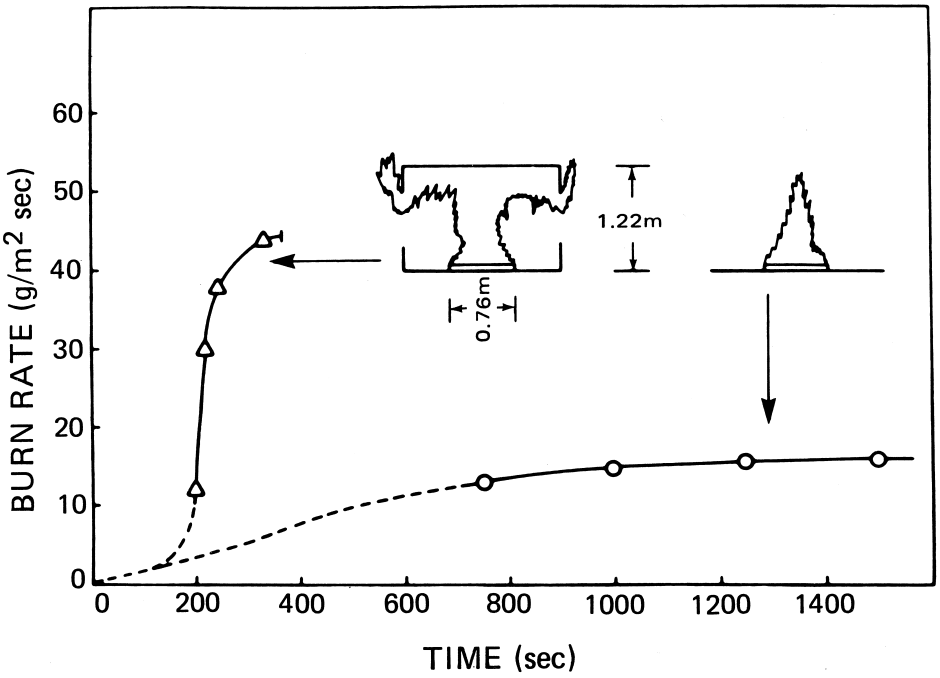


FIGURE 15 The effect of enclosure on the rate of burning of a slab of poly(methyl methacrylate) (0.76 × 0.76 m). [From Friedman (1975). International Symposium on Fire Safety of Combustible Materials, University of Edinburgh.]

of origin, as the rate of burning of the fully developed fire is sufficient to generate untenable levels of heat, smoke, and toxic gases at points remote from the fire if there is a route by which they can enter corridors, escape routes, and stairways. Clearly, it is important that the occupants of the building be evacuated as quickly as possible following ignition. Early detection will increase the chances of successful escape, but this may be jeopardized if the growth period of the fire is short. Consequently, it is important to be able to identify scenarios that could lead to rapid fire development and to design spaces and select their contents so that time to flashover would be maximized should fire occur.

The sequence leading from ignition to flashover can be described in terms of a fire that gradually increases in power (rate of heat release). The layer of hot smoky gases accumulating under the ceiling causes the temperature of the upper surfaces of the compartment to become hot. These surfaces *and* the smoke layer radiate downward, heating the combustible materials at low level at a rate which will increase as a rate which will increase as the fire grows in size. As discussed earlier (Sections IV and V), this will lead to enhanced rates of burning [increasing $\dot{Q}''_E$ in (3)] and high rates of spread of flame over combustible surfaces. Flashover is the transition from relatively localized burning when flame spreads rapidly to engulf all the combustible materials in the compartment.

This effect is shown clearly in Fig. 15 which compares the mass burning rates of two slabs of poly(methyl methacrylate), one burning in an unconfined atmosphere, the other under a hood intended to simulate the confinement offered by the ceiling of a compartment. The presence of the hood causes deflection of the fire plume and eventually creates a reservoir of hot burning gas, which acts as an intense source of radiant heat ($\dot{Q}''_E$) for the horizontal fuel slab. Not only is the final rate of burning several times greater than that for the free-burning slab, but the rate of development is also greatly enhanced. This behavior is understood to be typical of the processes that occur in the development of a real compartment fire.

In view of the importance of "flashover" in the compartment of origin to the safety of the occupants of the building, there has been much interest in developing an understanding of the process at a fundamental level. The key question is whether or not a fire in a compartment will progress to flashover and beyond. Intervention by trained firefighters can prevent flashover, but this is a hazardous operation if the fire is large. In recent years, the understanding of the mechanism has been combined with advanced technology in the use of water sprays and mists to provide a technique for preventing flashover. The technique, originally explored and developed in Sweden, demands high degrees of courage and skill.

Early work by Waterman drew attention to the fact that flashover will not occur unless the fire can grow beyond a certain critical size. Various "markers" have been used in experimental fires to define the onset of flashover, or flashover itself. Waterman concluded from a study of small-scale compartment fires that a satisfactory criterion for the onset of flashover is a radiant heat flux at floor level of 20 kW/m^2. This is substantially more than the minimum levels necessary for piloted ignition of most combustible solids, while somewhat less than the values associated with spontaneous ignition (Table IV). However, there is no doubt that flame spread rate will be greatly enhanced by a heat flux of this magnitude. Other criteria based on experimental observations have been suggested: thus, an upper gas temperature of 600°C has been observed to coincide with flashover in compartments approximately 2.5 m high, and flames are normally seen to emerge from the upper parts of ventilation openings. This is associated with the commencement of burning of the smoke layer under the ceiling and is followed by sustained external flaming as the fully developed fire begins. It is a clear indication that flashover has occurred. For fire safety design calculations, however, a radiant heat flux of 20 kW/m^2 provides a conservative criterion for the end of the growth period.

Modeling of this stage of the compartment fire is extremely complex because of the many and varied interactions that must be taken into account. There is a simple zone model in which the flashover condition is regarded as a thermal instability analogous to that invoked to account for the phenomenon of spontaneous ignition. Flashover is treated as the stage in the fire when the rate of heat production within the compartment first exceeds the rate at which it can be lost: under these conditions, the extent and intensity of the fire rapidly increase until the rate of burning reaches a new state of quasi-equilibrium, normally limited by the rate of air supply (ventilation).

Considerable effort has gone into developing more sophisticated models to describe the pre-flashover fire. A major project at Harvard University, under the direction of Emmons, led to the Harvard Fire Code, a "zone" model which has undergone successive modifications as our understanding of the component parts of fire dynamics has improved. It has been developed further as "FIRST" at the National Bureau of Standards (now the National Institute for Science and Technology) and a modification incorporated into HAZARD 1. The latest version of the Harvard zone model represents the most advanced of its kind, incorporating among other things the fluid dynamics of the flows of gas in and out of the compartment, the movement of the flame, radiation from the layer of hot smoky gases beneath the ceiling, and the response of combustible materials to an increasing radiant heat flux. It is an extremely valuable tool for developing an understanding of the mechanisms involved during the growth stage of a fire, but its value for design purposes has yet to be proven.

From the experimental point of view, there is now good evidence to suggest that before the onset of flashover in a compartment can be attained, the fire must grow to a certain minimum size. This is consistent with Waterman's earlier conclusions, but it was not until 1981 that sufficient data had been obtained to attempt some form of correlation. Thus, examination of a large amount of data on the upper room-gas temperatures in enclosure fires that had not actually achieved flashover (upper temperature <600°C) revealed that these data could be correlated using a simple zone model of the preflashover fire. Then, on the assumption that an upper gas temperature of 500°C would be a conservative indicator of the onset of flashover, the following equation was deduced:

$$\dot{Q}_{\mathrm{FO}} = 610(h_{\mathrm{k}} A_{\mathrm{T}} A_{\mathrm{w}} H^{1/2})^{1/2} \qquad (14)$$

where $\dot{Q}_{\mathrm{FO}}$ is the rate of heat release required to achieve flashover. A_{T} is the internal surface area, A_{W} and H are the area and height of the ventilation opening, respectively, and h_k is an effective surface heat-transfer coefficient (W/m^{2}K). for thermally thick walls, h_{k} is proportional to $(k\rho c)^{1/2}$, indicating that compartments lined with materials of low thermal inertia are likely to give flashover at lower values of $\dot{Q}_{\mathrm{FO}}$. This has been deduced from other studies, particularly in measurement of the time to flashover in compartments of different thermal mass. The effect is associated with the exposed surface, and not cavity insulation. Consequently, the effect is likely to be small with the limited range of wall lining materials currently in use, particularly in the case of fast-growing fires.

Factors affecting the time to flashover were studied in an international program of small-scale tests carried out under the auspices of the CIB (Conseil International du Bâtiment). The results were analyzed statistically, and it was found that the time to flashover was influenced by a number of factors, particularly those that led to rapid initial fire development and impingement of flame on the ceiling. These included size and location of the ignition source (or in terms of a real fire, those of the item first ignited), the height of the fuel bed, the fire properties of the wall lining materials, and the bulk density of the fuel, which influences the rate at which fire can spread from one item to its neighbor. This work, with subsequent studies at the National Bureau of Standards (now NIST) has shown the importance of the spacing of items of furniture in the early growth of fire. If an item is more than 1 m from an adjacent item, then it is unlikely that fire will spread directly from one to the other.

In the postflashover fire, all combustible surfaces are burning and the average temperatures within the

compartment will rapidly increase to a high level determined by the rate of heat release within the compartment, the rate of heat loss, and the duration of burning. A flow of hot combustion products is established through the upper parts of the ventilation openings, above a neutral pressure plane. This is driven by the positive overpressures that are maintained in the upper part of the enclosure by the buoyancy of the hot gases within the compartment. The corresponding pressure differential below the neutral plane causes air to be drawn into the fire (this is clear in Fig. 16a). The high levels of radiative and convective heating within the compartment maintain a high rate of production of fuel volatiles [$\dot{Q}''_{\mathrm{E}}$ in Eq. (3)] and apply a thermal load on the compartment boundaries, which may eventually lead to their failure (see below).

The rate of heat release within the compartment depends not only on the rate of production of fuel volatiles but also on the rate at which air is being drawn into the compartment through the ventilation opening as a result of buoyancy-induced flows. Early studies of experimental fires involving wood cribs led to the conclusion that the rate of burning could be related to a ventilation factor $A_{\mathrm{w}}H^{1/2}$, such that

$$\dot{m} = 5.5A_{\mathrm{w}}H^{1/2} \quad \text{kg/min} \tag{15}$$

where A_{w} and H are defined as in Eq. (12). The correlation of these data is shown in Fig. 16.

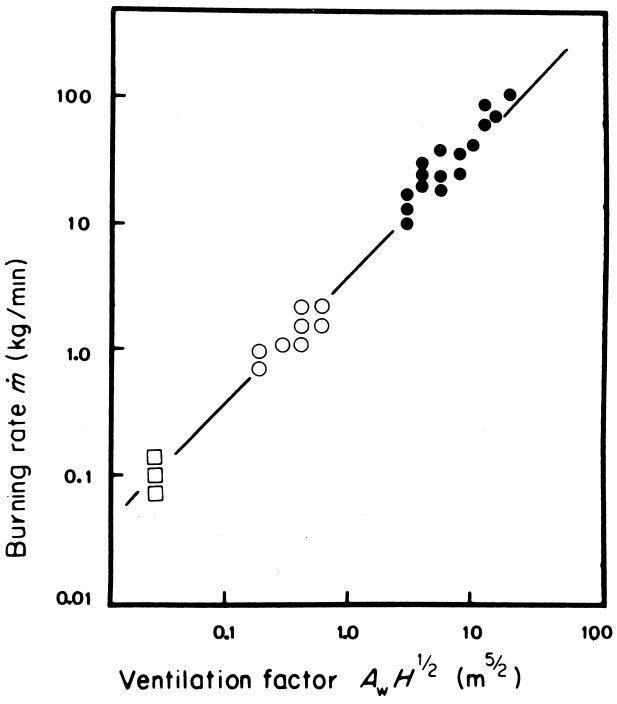


FIGURE 16 Mass burning rates of wood cribs in enclosures as a function of the ventilation parameter $A_{\mathrm{w}}H^{1/2}$ for ventilation-controlled fires. ●, Full-scale enclosures; O, intermediate-scale models; □, small-scale models. [From Lie, T. (1972). "Fire and Buildings," Elsevier Applied Science, Barking, England.]

Theoretical justification for such a relationship can be found by analyzing the buoyancy driven gas flows in and out of the compartment and assuming that stoichiometric burning takes place—that is, all the oxygen entering the compartment reacts quantitatively with the equivalent amount of fuel vapor. The implication of this is that the rate of burning and the rate of inflow of fresh air are directly coupled, although there appears to be no logical reason why this should be so. In fact, if the geometry or nature of the fuel is changed, then the relationship can break down dramatically. For example, if instead of wood cribs, the fuel was present as wall-lining material, then very much higher "rates of burning" are achieved, greatly in excess of the stoichiometric requirement for the inflow of air. The consequence is that some of the fuel vapors will not burn in the compartment, but will flow out of the ventilation opening to burn vigorously as air is entrained into the plume thus producing external flames. This is illustrated in Fig. 17a.

The size of such flames cannot be calculated with any degree of precision, although several important contributory factors have been identified. Thus, the greater the quantity of unburnt fuel escaping through the upper part of the ventilation opening, the larger will be the flame, although this will be moderated by the width of the opening. Some correlations have been developed for the purpose of estimating the fire protection necessary for external steelwork.

For a given distribution of fuel within a compartment, the rate of burning will be independent of the ventilation factor if the vent area is made large enough. The rate of burning of this type of fire is determined by the area of the fuel bed, and the appearance is quite different from the fully developed ventilation-controlled fire as may be seen by comparing Fig. 17a and b. Generally speaking, the maximum temperatures achieved in fuel-controlled fires are less than those reached in ventilation-controlled fires, unless the ventilation is particularly poor. This can be understood qualitatively, as the maximum temperatures will be achieved when the rate of heat release in the compartment is at a maximum. If excess air is entering the compartment, this will have a cooling effect, thus reducing the rate of production of the volatiles ($\dot{m}$), while if excess fuel vapors are being produced, there will be an increased flow of hot gases from the compartment, which at the same time will have the effect of throttling the inflow of air.

Steady-state analyses of various scenarios, in which the rates of burning have been decoupled from the rates of air inflow, have been carried out and show the importance of the heat of gasification (L_{v}) in determining the nature of the fire. Fuels which have a low heat of gasification (or vaporization) tend to have high rates of burning under steady-state conditions, [Eq. (3)]. This is demonstrated most clearly with liquid fuels such as gasoline or hexane



(a)

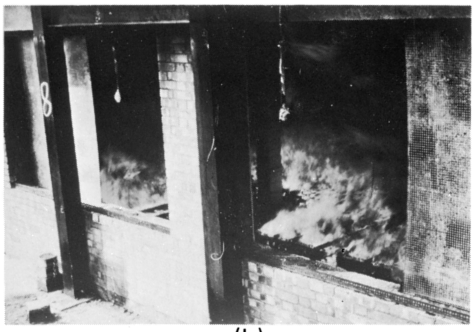
Walls and ceiling lined

with wood fiber insulating board

7.5 kg fuel/m^2 of floor

A_S = 65 m^2

Flames out of opening for 5½ min



(b)

Cribs

15 kg wood/m^2 of floor

A_S = 55 m^2

Flames out of opening for 0 min

FIGURE 17 The effect of a large exposed fuel surface area on fire behavior. (a) Ventilation-controlled fire. (b) Fuel-controlled fire. [Reproduced by permission of the Controller, HMSO. © Crown Copyright.]

(Table VII). With a solid fuel, the rate at which the maximum rate of burning will be achieved will depend on the thermal inertia, [Fig. (6)]. Thus, the transition to the fully developed fire (flashover) is not instantaneous, being prolonged by the natural delay that occurs before individual items reach their maximum rates of burning.

It is during the fully developed compartment fire that elements of structure are subjected to high thermal stress. They must be designed such that they continue to perform their function during a fire: that is, they must have a degree of fire resistance. Traditionally, fire resistance is measured in a large-scale standard test and taken as 80% of the time to failure in a "standard fire" which is defined precisely in terms of a standard temperature-time program. This is an unsatisfactory approach, as exposure within any of the existing test furnaces bears little resemblance to real fire exposure. This is particularly true if some of the fuel involved is noncellulosic (i.e., of synthetic origin, such as polypropylene). In the extreme situation of a fire on a petrochemical plant, elements of structure may be exposed to hydrocarbon pool fires which produce very high temperatures much more rapidly than the typical compartment fire. It has been necessary to develop more severe fire resistance tests for this scenario.

A design method for fire protection of structural steelwork in buildings was developed in Sweden in the 1970s; the potential fire exposure is first calculated from the rate of heat release, given by

$$\dot{q}_c = \Delta H_c \dot{m} \tag{16}$$

where $\dot{m}$ is given in Eq. (15), and the duration of burning which depends on the total amount of fuel present, that is, the fire load. By carrying out numerical integration of the equations describing the rate of heat generation ($\dot{q}_c$) within and the rate of heat loss from the compartment ($\dot{q}_w$, $\dot{q}_L$, and $\dot{q}_R$ in Fig. 18), the average gas temperature within the compartment can be calculated as a function of time from flashover.

Typical curves for four different ventilation parameters ($A_w H^{1/2}/A_t$, where A_t is the total internal surface area) and a range of fire load densities (expressed in terms of the available combustion energy, in MJ/m^2) are shown in Fig. 19. The appropriate one may then be selected and used to calculate from first principles whether or not a given element of structural steelwork will survive that particular fire on the basis that an increasing steel temperature is accompanied by a decrease in the load-bearing capacity. This type of calculation is accepted in Sweden as an alternative to the more conventional, but inflexible, building codes that embody the fire-resistance tests. However,

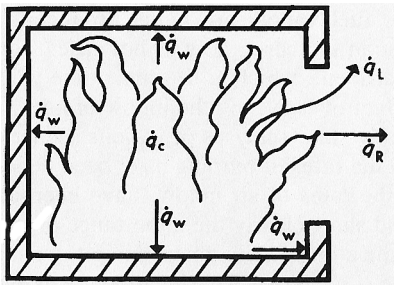


FIGURE 18 Heat losses during a fully developed compartment fire [From Pettersson, O., Magnusson, S. E., and Thor, J. (1976). Swedish Institute of Steel Construction, Publication 50.]

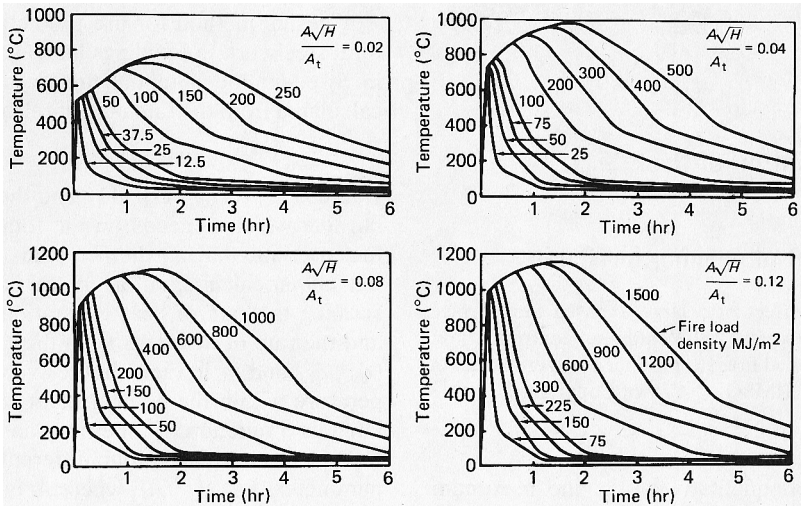


FIGURE 19 Theoretical temperature–time curves for compartment fires with different fire load densities (MJ/m^2) and opening factors, $A_w H^{1/2}/A_t(m^{1/2})$. [From Pettersson, O., Magnusson, S. E., and Thor, J. (1976). Swedish Institute of Steel Construction, Publication 50.]

both approaches are essentially flawed in that they consider single elements of structure (beams, columns, etc.) in isolation and ignore the interactions between structural elements that occur in complex three-dimensional buildings, particularly under fire conditions.

The term fire severity is often used, although seldom is any attempt made to quantify it. It has been related to the "fire resistance" that would be required to survive the fire. In analyses of data from a large number of experimental compartment fires, Law established a correlation between t_F, the required fire resistance (minutes), and the fire load (M_F) in kilograms wood equivalent, the ventilation area (A_w in m^2), and the total internal surface area of the compartment (A_T in m^2):

$$t_F = K' \frac{M_F}{(A_w A_T)^{1/2}} \tag{17}$$

where K' is a constant, close to unity. Other similar formulae have been developed which purport to define a "time equivalent" in terms of a wider range of parameters (including shape and the thermal properties of the boundaries of the compartment), but there is still a great deal of uncertainty in how these can best be applied. In general, they do not apply to unprotected steelwork.

VIII. SMOKE PRODUCTION AND MOVEMENT

The fire statistics in the U.K. and the U.S. indicate that over 60% of fire fatalities can be attributed to the inhalation of "smoke and toxic gases" rather than direct contact with flame. Consequently, the rate at which fire products flow from the fire area to invade escape routes and stairways within a building is significant in determining how long people will have to move to a place of safety. Although the term "smoke" refers specifically to a suspension of particles in air, it may conveniently be used here to include the noxious and toxic gases that are formed at the same time.

With no exceptions that are relevant to this discussion, all fires produce smoke. Its rate of production is determined by the rate of burning of the fire, although the density, or opacity, of the smoke will also depend on the nature of the materials and the conditions under which they are burning. As smoke will reduce visibility within a building and make escape more difficult, much effort has been put into devising ways of determining the smoke production characteristics of combustible materials. This has only been partly successful, as the conditions of burning affect the smoke yield quite considerably: for example, restricted ventilation can give rise to a dramatic increase in the amount of smoke produced (in terms of its obscuration potential). Indeed, in order to assess the potential smoke hazard of combustible materials used within a compartment, the likely rate of burning and the fire ventilation conditions should also be considered.

Smoke movement within a building is determined by pressure gradients, which are created by the following factors:

1. Buoyancy created directly by the fire
2. Buoyancy arising from differences between the internal and external temperatures (origin of the stack effect)

3. The effects of external wind and air movement
4. Air-handling systems within the building

The buoyancy created by the fire will be dominant in its vicinity, although this influence will extend as the fire grows in size. Within the compartment of origin, the pressures generated above the neutral plane cause the smoke to move outward along any leakage path including service penetrations which have not been adequately protected. As the smoke moves from the compartment of origin and cools, its volume increases as it becomes more dilute and the buoyancy decreases. It is then that the other forces begin to dominate.

In a tall building, the stack effect created by the temperature differential between the inside and the outside will result in a net flow upward if the inside is warmer (downward if cooler). The former situation can result in a very rapid spread of smoke from a fire on a lower floor of a high rise building to the upper floors (Fig. 20). This is believed to have occurred in the fire at the MGM Grand in Las Vegas in 1981.

External air movement can induce pressure variations on the external faces of a building, thus creating pressure gradients within the building. Smoke movement from the area of origin can be greatly enhanced—or inhibited—by this, depending on the direction of the wind in relation to the location of the fire within the building. With complex building geometries, such as a combined single-story shopping mall and multistory office block, the wind-induced pressure patterns over the surface of the building can be extremely complicated and will vary significantly with wind direction and strength. Such behavior introduces problems in the design of natural smoke-venting systems.

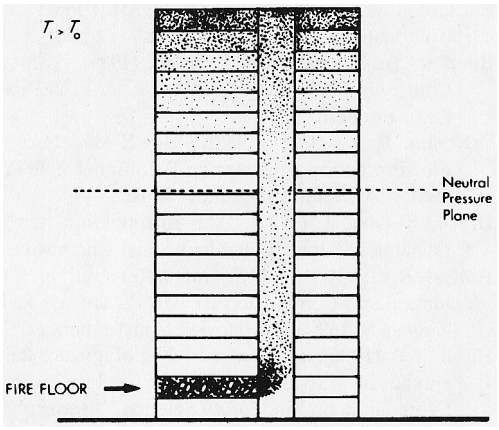


FIGURE 20 Influence of the stack effect on the movement of smoke in a high-rise building. The internal temperature (T_i) is greater than the external temperature (T_o). Below the neutral plane there is a tendency for air to be drawn into the central core, while above it there is a net outward flow. [From Drysdale, D. (1998) "Introduction to Fire Dynamics," John Wiley and Sons, Chichester.]

Much progress has been made in interpreting these various interactions and has led to design procedures for smoke control systems in the U.K. and U.S. However, there are major problems in the design of systems for large open structures such as atria and airport terminal buildings, as there is only limited knowledge on the behavior of smoke in such spaces. It is likely that progress will be made in this area by making use of computational fluid dynamics to generate sophisticated computer models. Although these require significant computer power and CPU time, they will provide a method of testing particular design solutions which will be considerably cheaper than attempting physical modeling. However, some fundamental validation work must first be carried out.

SEE ALSO THE FOLLOWING ARTICLES

COMBUSTION • FUELS • HEAT FLOW • INTERNAL COMBUSTION ENGINES

BIBLIOGRAPHY

American Society for Testing and Materials (1993d). Standard method of test for heat and visible smoke release rates for materials and products using an oxygen consumption calorimeter. ASTM E1354, (ASTM, Philadelphia, PA).

Babrauskas, V., and Williamson, R. B. (1978, 1979). "Post flashover compartment fires, Parts 1 and 2," *Fire Mater.* **2,** 39–53; **3,** 1–7.

Babrauskas, V., and Peacock, R. (1992). "Heat release rate: the single most important variable in fire hazard," *Fire Safety J.* **18,** 255–272.

Barnard, J. A., and Bradley, J. N. (1985). "Flame and Combustion," Chapman and Hall, London.

Butcher, E. G., and Parnell, A. C. (1979). "Smoke Control in Fire Safety Design," E. & F. N. Spon Ltd., London.

Cox, G., ed. (1995). "Combustion Fundamentals of Fire," Academic Press, London.

Di Nenno, P. J. *et al.*, eds. (1995). "SFPE Handbook of Fire Protection Engineering," 2nd Edition, Section 1, Society of Fire Protection Engineers, Boston.

Drysdale, D. (1998). "An Introduction to Fire Dynamics," 2nd Edition, John Wiley and Sons, Chichester.

Hall, A. R. (1973). "Pool Burning: a Review," *In* "Oxidation and Combustion Reviews" (C. F. H. Tipper), ed. **6,** 169–225, Elsevier, Amsterdam.

Hasemi, Y. (1986). "Thermal modeling of upward flame spread," *In* "Proceedings of the First International Symposium on Fire Safety Science," Hemisphere Publishing Corporation, New York, pp. 87–96.

Heselden, A. J. M., and Melinek, S. J. (1975). "The early stages of fire growth in a compartment. A cooperative research programme of the CIB (Commission W14). First Phase," Fire Research Note No. 1029.

Hinkley, P. L., Wraight, H. G. H., and Theobald, C. R. (1984). "The contribution of flames under ceilings to fire spread in compartments," *Fire Safety J.* **7,** 227–242.

Lees, F. P. (1980). "Loss Prevention in the Process Industries," Butterworths, London.

McCaffrey, B. J. (1979). "Purely buoyant diffusion flames: some experimental results," National Bureau of Standards, NBSIR 79-1910.

Malhotra, H. L. (1982). "Design of Fire Resisting Structures," Surrey University Press.

Mitler, H. E. (1985). The Harvard Fire Model. *Fire Safety J.* **9,** 7–16.

National Fire Protection Association (1981). NFPA Handbook 15th Edition.

Nordtest (1986). "Surface Products. Room fire tests in full scale," Nordtest method NT Fire 025.

Palmer, K. N. (1973). "Dust Explosions and Fires," Chapman and Hall, London.

Pettersson, O., Magnusson, S. E., and Thor, J. (1976). "Fire Engineering Design of Structures," Swedish Institute of Steel Construction, Publication 50.

Quintiere, J. G. (1989). "State of fire research and safety," *In* "Proceedings of the Second International Symposium on Fire Safety Science," Hemisphere Publishing Corporation, New York, pp. 15–28.

Strehlow, R. A. (1984). "Combustion Fundamentals," McGraw-Hill, New York.

Tewarson, A. (1982). "Experimental evaluation of flammability parameters of polymeric materials," *In* "Flame Retardant Polymeric Materials," 3. (M. Lewin, S. M. Atlas, and E. M. Pearce, eds.) 97–153. Plenum Press, New York and London.

Zabetakis, M. G. (1965). "Flammability characteristics of combustible gases and vapours," U.S. Bureau of Mines Bulletin No. 627.



Foundations

Fred H. Kulhawy
Cornell University

Anwar Hirany
Electric Power Research Institute

I. Background
II. Types of Foundations
III. Geotechnical Concepts for Foundation Design
IV. Factors Affecting Foundation Performance
V. Basic Theories for Foundation Design
VI. Foundation Construction
VII. Quality Assurance and Control

GLOSSARY

Bearing capacity Capacity of the soil or rock to sustain the applied foundation load. *Ultimate* bearing capacity is the maximum stress sustained by the foundation at shear failure of the soil or rock. *Allowable* bearing capacity is the maximum design stress that can be imposed on the foundation and is equal to the ultimate bearing capacity divided by the *factor of safety*.

Deep foundation Generally, a long, cylindrical foundation having its depth significantly greater than its diameter or width and deriving its capacity from both side and tip resistance. Common types are the *pile*, which is generally between 6 and 18 in. in diameter or width (but can be 5 feet or more for bridges or offshore platforms), and is driven into the ground by mechanical means, and the *drilled shaft*, which is generally between 2 and 5 feet (but can be more than 10 feet) in diameter and is constructed by augering a hole into the ground and backfilling it with reinforced concrete (also called *drilled pier*, *bored pile*, or *drilled caisson*).

Drained and undrained loading Time-dependent loading conditions. *Drained* conditions occur when the excess pore water pressure resulting from the applied load has been dissipated (long-term loading condition in fine-grained soil). *Undrained* conditions occur when excess pore water pressure has developed under the applied load (short-term loading conditions in fine-grained soil).

Effective stress Difference between the total stress and pore water pressure at a point within the soil/rock mass. *Total stress* at a point is defined as the sum of the stress exerted by the total weight (solids plus water) of the soil/rock and the load on the foundation. *Pore water pressure* is the pressure exerted by water in the soil/rock voids. The ratio of the horizontal effective stress to the vertical effective stress at a point is the *coefficient of horizontal soil stress* (K_0).

Excavation support system Generally, a temporary system used to support the walls of the excavation during foundation construction. *Braced* (horizontally supported) vertical timber planks (*sheeting*), steel *sheet*

piles, or a combination of steel H-sections (*soldier piles*) and horizontal timber planks (*lagging*) may be used to support the excavation walls for shallow and mat (slab) foundations. Steel pipe (*casing*) or *drilling slurry* may be used to support the excavation for a drilled shaft.

Inspection Quality assurance and control performed during foundation construction and installation.

Load test Test performed on full-scale foundation to determine its load–displacement characteristics and ultimate bearing capacity.

Settlement (heave) Downward (upward) displacement of the foundation resulting from elastic, consolidation, and secondary compression (expansion) of the soil and/or rock. *Elastic* response results from the initial elastic displacement to the applied load. *Consolidation* (*swell*) response results from dissipation of excess pore water pressure (i.e., flow of water from the pores). *Secondary* response occurs from soil skeleton readjustment to the new effective stress after dissipation of the excess pore water pressure.

Shallow foundation Generally, a large, rectangular foundation embedded to a depth less than or equal to its width and deriving its capacity predominantly from end bearing (tip resistance). Common types are the *spread footing*, which is constructed of cast-in-place reinforced concrete, and the *grillage*, which is constructed of prefabricated steel or treated lumber placed in a crisscross pattern.

A FOUNDATION can be described simply as the support system that rests on or in the ground and transfers to the ground the weight of the supported structure or equipment and the loads acting on them.

I. BACKGROUND

All man-made structures, from buildings to bridges to towers to dams, require a foundation to support them. Foundations also are necessary to support various equipment such as pumps, motors, turbines, generators, etc. They can range from a simple peg driven into the ground for supporting a camping tent to a complex system of very deep foundations for supporting an offshore drilling platform, which will require elaborate design and often innovative construction techniques. Foundations can ultimately affect or even control the performance, safety, and economics of a structure, and it is almost impossible to design or construct these structures without considering the engineering characteristics of the foundation soils and rocks.

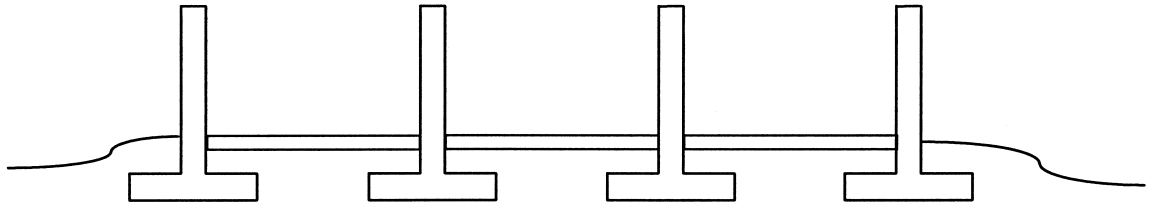
The primary function of a foundation is to distribute the supported load to the soil and rock in a manner that restricts movements (displacement, settlement, heave, strain, rotation, and vibration) to tolerable limits. These movements are associated with the shear failure, compressibility, and/or volume change characteristics of the soil or rock, all of which can be affected adversely by the foundation construction procedures. Therefore, two important requisites for a successful foundation project are adequate engineering knowledge of the natural ground materials—soil and rock—and adequate experience in the art of foundation construction.

Foundations must be designed so that they are stable and do not move excessively. They must resist various loads that could be axial (vertical uplift and compression), lateral (horizontal and moment), inclined, or torsional in direction and can be static (gravity load from a building), transient (wind loads on a high-rise building), cyclic (traffic loads on a bridge), or dynamic (loads from earthquakes or machine vibrations) in nature. For each of these loading combinations, a number of different types of foundations can be used. The final choice commonly is governed by various factors such as the ground conditions at the site (soil or rock characteristics mentioned above), depth of frost penetration, wind and water erosion, type of structure or equipment to be supported, influence on or of adjacent structures, constructibility concerns, and project economics.

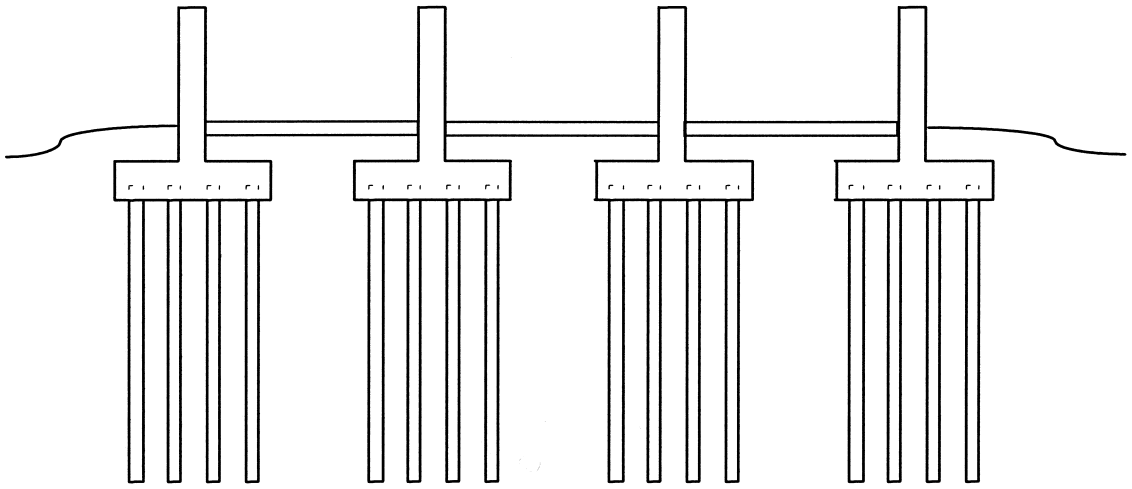
Foundation engineering is a very broad subject. Accordingly, in this article, the emphasis will be only on the basic concepts involved. Details can be found in the bibliography.

II. TYPES OF FOUNDATIONS

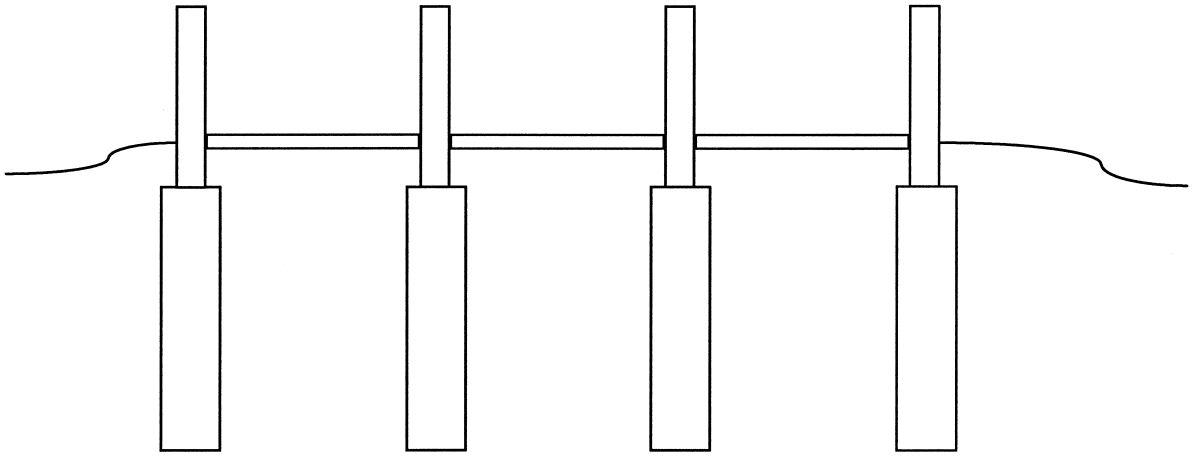
A wide variety of different types of foundations have evolved in response to the many variables affecting foundation design, and these include (see Fig. 1) spread foundations, mat foundations, driven piles, drilled shafts, and other specialty foundations such as direct embedment and anchor foundations. These different foundation types can be grouped under two major categories: shallow and deep. The difference between these two categories is based on their relative depth below the ground level and how the load is distributed to the soil. Generally, if the foundation is embedded in soil to a depth less than or equal to its width, it is considered a shallow foundation. As described later, the applied compressive loads from the structure are resisted by the base of a shallow foundation and by both the base (tip) and sides of a deep foundation. Table I presents a summary of the more commonly used foundations for various soil and loading conditions.



a. Shallow Foundation (spread footing)



b. Deep Foundation (piles and pile cap)



c. Deep Foundation (drilled shaft)

FIGURE 1 Foundation types. (a) Shallow foundation (spread footing). (b) Deep foundation (piles with pile cap). (c) Deep foundation (drilled shaft).

TABLE I Applicability of Foundation Types

Parameter	Applicability for given foundation type[a]					
	Spread	Grillage	Mat	Piles	Drilled shafts	Anchors
Soil/rock condition						
Competent soil/rock near surface	**MF**	**MF**	SC	SC	SC	SC
Expansive/collapsing soil	SC	SC	**MF**	SC	SC	NR
Competent soil/rock at depth	NR	NR	SC	**MF**	**MF**	SC
Swamps	NA	NA	NA	**MF**	SC	SC
Structure type						
Single-family residential	**MF**	SC	**MF**	SC	SC	SC
Low-rise buildings	**MF**	SC	**MF**	SC	SC	SC
High-rise buildings	SC	SC	SC	**MF**	**MF**	SC
River crossings (bridges)	NA	NA	NA	**MF**	**MF**	SC
Highway crossings (overpasses)	**MF**	**MF**	NA	**MF**	**MF**	SC
Equipment/machinery	**MF**	NA	**MF**	SC	SC	NA
Offshore platforms	NA	NA	NA	**MF**	SC	SC
Transmission towers	SC	**MF**	SC	SC	**MF**	**MF**
Loading type						
Axial uplift	**MF**	**MF**	**MF**	**MF**	**MF**	**MF**
Axial compression	**MF**	**MF**	**MF**	**MF**	**MF**	NA
Horizontal/moment	SC	SC	SC	**MF**	**MF**	SC
Vibratory	**MF**	SC	**MF**	SC	SC	SC
Costs	**Low**	**Low**	Medium to high	High	High	Medium to high

[a] MF, Most feasible; NA, not applicable; NR, not recommended; SC, under special conditions.

A. Spread Foundations

This type of foundation generally is used where competent soil or rock is found at shallow depths and the groundwater table is below the bottom of the foundation. It has large, rectangular plan dimensions, is placed in a shallow excavation, and then is backfilled. Effectively, the load from the structure is spread out over a larger area of the soil, thereby reducing both the stress exerted on the soil and the settlement. The two main variations are the footing, which is constructed of cast-in-place concrete, and the grillage, which is prefabricated steel or treated lumber placed in a crisscross pattern. One footing or grillage commonly is used for each column of the structure. Spread foundations are shallow foundations and are used most commonly to support the columns and walls of lightly-loaded structures, such as single-family dwellings or low-rise buildings.

B. Mat Foundations

Mat or slab foundations are similar to spread foundations in that they have large, rectangular plan dimensions, but they basically cover the entire footprint of the structure. Generally, all of the columns and load-bearing walls of the structure are supported by one large mat foundation. These foundations have been used for high-rise buildings where competent soil or rock is found at shallow depths and for single-family dwellings that are constructed on soil susceptible to moisture-related volume change (e.g., expansive clays and collapsible sands). Mat foundations fall under the category of shallow foundations.

C. Driven Piles

Piles are long, slender foundation elements that are driven into the soil by mechanical means. They may consist of steel H-sections, steel pipe, precast concrete, wood, fiber glass, or a composite of these materials. Piles normally are not used singly to constitute a foundation; instead, two or more are driven to create a closely-spaced group, which is then integrated structurally through a cast-in-place concrete pile cap. Piles fall under the category of deep foundations, and they generally are used for high-rise or low-rise buildings and river crossings where the water table is high and/or competent soil or rock is found at depths where excavation for a spread or mat foundation would be uneconomical.

D. Drilled Shafts

Drilled shafts (also known as drilled piers, bored piles, and drilled caissons) are constructed by augering a cylindrical

hole into the ground and backfilling it with reinforced concrete. The bottom of the shaft can be belled in appropriate soil to give a larger bearing surface, although bells are becoming less common. A single shaft normally is used to support a building column, but groups can be used, similar to driven piles. Drilled shafts fall under the category of deep foundations and often are a suitable and economical alternative to pile foundations. They are most commonly used for river and highway crossings, large electricity transmission structures, high-rise and low-rise buildings, and soil-retaining structures.

E. Specialty Foundations

Under certain circumstances, the supporting column of a structure can itself be used as the foundation. A hole is excavated, and the column is placed directly into the hole and then backfilled with the excavated soil. Smaller pole structures for electricity and telephone transmission, sign structures, etc., are supported this way. This type of foundation is called a direct embedment foundation and falls under the category of a deep foundation.

Anchors are long, very slender structural elements, typically of steel, although other options exist. There are several major variations of anchors. Plate anchors are long rods with flat plates at various levels and are placed in excavated holes backfilled with soil or concrete. Screw anchors essentially are augers which are screwed directly into the ground. Grouted anchors are long rods or tendons that are placed in excavated holes and then grouted to fill the annulus. Generally, the tendons and rods are prestressed. Inflated anchors are similar to the grouted type, except that they are expanded to grip the wall of the hole and may or may not be grouted afterward. Anchors generally are used to resist uplift or tension force and fall under the category of deep foundations. They are most commonly used for supporting soil-retaining structures, mobile homes, and structures supported by guy wires.

III. GEOTECHNICAL CONCEPTS FOR FOUNDATION DESIGN

The rate of loading and the geotechnical (or ground engineering) characteristics of the underlying soil govern the foundation response to the applied load, and therefore familiarity with basic geotechnical concepts is necessary to understand the behavior of foundations under load.

A. Consolidation

When a load is applied to any material, it deforms or strains, and a plot of the load (or stress) versus the corresponding deformation defines the load–deformation (load–settlement, stress–strain, or stress–deformation) characteristics of the material. In most elastic materials such as steel, the entire deformation is almost instantaneous, occurring as soon as the load is applied. In saturated particulate materials such as soil, the applied load causes elastic deformation of the soil structure, and it also increases the water pressure in the soil voids.

The soil compresses as the water flows from the loaded region to areas of lower water pressure. This time-dependent compression or deformation of the soil, known as consolidation, is controlled by the rate at which water flows from the pores. As this flow occurs, the soil particles move closer together and the soil mass becomes denser and stronger.

In saturated, fine-grained soils such as clays, this consolidation process takes a long time because of its low coefficient of permeability (hydraulic conductivity), and therefore the deformation can continue to occur for months, years, or even decades after the application of load. On the other hand, coarse-grained soils such as sands and gravels have a relatively high hydraulic conductivity, and the water flow is almost instantaneous. For foundations in competent sand and gravel, most of the deformation occurs during construction, and therefore most of the settlement has taken place by the time the structure is completed.

B. Effective Stress

Soils at the bottom of the ocean or other water body generally are very soft or loose, even though they have been subject to a tremendous amount of water pressure for a long time. This condition implies that the weight or pressure of water does not influence the process of consolidation, otherwise the ocean-bottom soils would be dense or hard. The deeper soils get consolidated under the weight—more specifically, buoyant or submerged weight—of the overlying soils. In soil under water, the reduced buoyant weight means less grain to grain stress in the soil. This grain-to-grain stress is called the effective stress σ' and the water pressure in the pores between the grains is called the pore water pressure u. Mathematically, the effective stress at a point is defined as the difference between the total stress σ and the pore water pressure at that point. Total stress at a given depth below the soil surface is the product of the depth and the total unit weight of the soil.

As stated above, when external stress is applied to saturated soil, the water pressure increases. This increase is above the hydrostatic pressure and is called excess pore water pressure. Theoretically, at the instant the external stress is applied, the excess pore water pressure is equal to the applied stress, and there is no increase in the

grain-to-grain (effective) stress. As consolidation occurs and water flows from the pores, the excess pore water pressure dissipates and the effective stress increases.

C. Overconsolidation Ratio

When a soil is loaded, say by the weight of an ice sheet or any other boundary load, it undergoes consolidation and the ground surface settles under the applied load. Once consolidation has occurred and the ice sheet melts away or the boundary load is removed, the unloaded soil tries to expand. This expansion can occur in the vertical direction only because adjacent soil particles restrict expansion in the horizontal direction. However, the ground surface does not return to its original elevation after unloading because the particles have shifted and realigned themselves during consolidation. The soil is now in a denser or stiffer state than it was prior to the loading. The overconsolidation ratio (OCR) is a qualitative indicator of this densification or stiffening of the soil, and it is defined as the ratio of the maximum overburden stress ever experienced by the soil (i.e., with the ice sheet on top) to the present overburden stress (i.e., without the ice sheet). Soils with an OCR = 1 are called normally consolidated (NC) and unaged (which means the soil layer is young and has not experienced more load than what it supports presently), while an OCR from 1 to 1.3 can still indicate NC soil, but usually soil that is older. Soil with an OCR from 1.3 to 4 is called lightly overconsolidated (LOC), soil with an OCR from 4 to 10 is called moderately overconsolidated (MOC), and soil with an OCR greater than 10 is called heavily overconsolidated (HOC).

The OCR governs the response of the soil to loading. Soils with low OCR (typically NC to LOC) tend to contract when subjected to shearing forces, and expel the water during shear. Soils with high OCR (typically MOC to HOC) tend to expand or dilate when subjected to shearing forces, which causes water to be drawn into the soil.

D. Coefficient of Horizontal Soil Stress

The stress acting at a point within a body of water is the same in all directions, and that is why it is called pressure. If this were true for granular soils, sand dunes would not exist, and the soil would flow like water. Unlike water, soil offers resistance to shear (frictional resistance), which causes the stress to be different in different directions within a soil mass. The ratio of the horizontal stress to the vertical stress at a point within the soil mass is defined as the coefficient of horizontal soil stress K_0 at that point. Depending on the soil OCR, K_0 can be less than, equal to, or greater than one (from as low as $1/4$ or $1/3$ to as much as 3 or 4), and therefore the horizontal stress can be less than, equal to, or greater than the vertical stress, respectively. Generally, the higher the OCR, the higher the K_0.

E. Undrained and Drained Loading

Undrained loading conditions develop when loads are applied relatively rapidly to fine-grained soils and no consolidation takes place because of the low coefficient of hydraulic conductivity. Initially, the pore water supports the entire applied load, and the soil volume, water content, and effective stress remain unchanged. With time, as water flows from the pores, the soil consolidates under the applied load, its volume and water content decrease, and the applied load is supported gradually by the soil particles, thereby increasing the effective stress. This limit state is referred to as the drained loading condition. In coarse-grained soil with high hydraulic conductivity, the drained loading condition occurs almost instantaneously after the application of load. Depending on the OCR, the drained strength may be greater or less than the undrained strength, and this fact must be considered in the design. In moderately to heavily overconsolidated soil, the drained strength generally is less than the undrained strength, which implies that long-term conditions generally will govern for this type of soil. The opposite is true for normally consolidated and lightly overconsolidated soil.

IV. FACTORS AFFECTING FOUNDATION PERFORMANCE

Foundations have to be designed for a variety of loading conditions on the basis of site-specific characteristics. Many factors must be considered to design a foundation that will perform satisfactorily, including geologic environment, loading characteristics, structure/equipment characteristics, foundation characteristics, foundation response, soil or rock response factors, and construction procedure.

A. Geologic Environment

Geologic environment provides the starting point for the foundation selection and design and includes site geology, soil and rock conditions, *in situ* state of stress, and groundwater conditions. Some soils are dense or hard, some are loose and soft, and some are hard and later become soft. When exposed to water, some soils expand and lift the foundation, whereas other soils collapse and lower the foundation. Among other factors that need to be considered are frost susceptibility and chemical composition of the soil and rock, seasonal groundwater fluctuations, erosion, and formation of solution cavities in rock. Details

of the local geologic environment are quite important in both design and construction.

B. Loading Characteristics

Loading characteristics include the load combinations (axial, lateral, moment, inclined, torsional) mentioned earlier. The key variables for foundation design are the magnitude, direction, and line of action of the load and how fast (rate of loading) and how often (frequency of loading) it is applied. Depending on the magnitude of the load, soil and rock response associated with a gradually applied static load can be significantly different if the same magnitude load is applied instantaneously or repeatedly.

C. Foundation Characteristics

Foundation characteristics include the weight, size, shape, and material of the foundation. The weight of the foundation helps in resisting uplift and horizontal force but reduces the amount of compression (downward) force that can be applied. The size and shape (geometry) of the foundation determine the volume of soil that will be affected by the applied loads. Larger foundations have a greater zone of influence both horizontally and vertically. Under certain load combinations, the frictional resistance between the foundation material and soil or rock will affect the load-carrying capacity of the foundation. The foundation material can be affected by the chemistry of the soil and rock or by groundwater fluctuations. Certain chemicals present in the ground can attack concrete and steel, and timber can deteriorate with wet and dry cycles caused by groundwater fluctuations.

D. Structure or Equipment Characteristics

The supported structure or equipment dictates the limits of tolerable movement of the foundation. Generally, it is the differential movement between adjacent foundations of a structure or interconnected equipment that is of greater concern than the absolute movement. Differential movement imposes additional load on the structural members (beams and columns) or misaligns the interconnected equipment so that they cannot function properly. The amount of differential movement that can be tolerated depends primarily on the foundation spacing (the greater the spacing, the greater the tolerance) and on the functional requirements of the structure or equipment. However, if all of the individually supported columns of a building settle by the same amount simultaneously, no additional loads will be imposed on the structure. The structure will be safe, but it could lose its functionality if the settlement is large.

E. Foundation Response

Foundation response describes how the foundation reacts to the applied loads, which defines the method of analysis. Foundation response is a function of its relative stiffness (or flexibility) compared to the surrounding soil and rock. Stiffer foundations generally tend to have a rigid-body movement (that is, they do not flex under the applied load) and transfer the load to the soil more evenly than flexible foundations. For example, if a lightly-loaded column is supported concentrically by a thick, heavily-reinforced concrete foundation having an area slightly greater than the cross-sectional area of the column, the load will be distributed evenly below the foundation. If this foundation is replaced by a very thin plate of steel having a significantly large area, more of the load will be transferred to the soil immediately below the column than below the edges of the plate. This flexibility also influences the development of potential failure surfaces and how they vary with depth. These surfaces can take various forms and determine how the load will be transferred to the soil or rock in end bearing (tip resistance) or side resistance. The actual load distribution will determine the deformation response, which is also a function of the soil and foundation material characteristics.

F. Soil and Rock Response Factors

Soil or rock response factors determine how the subsurface material will respond under the applied loads. Fine-grained soils, such as clays and silts, respond differently than coarse-grained soils, such as sands and gravels, when subjected to different types of loading. This response differs mostly because of the very different coefficients of permeability. Also, soil response is governed by its water content, and therefore the same soil will behave differently under dry, moist, and saturated conditions. A typical example of this can be found at the beach. The dry sand near the dunes is very loose and difficult to walk or drive on. The saturated soil close to the water (or under water) is comparatively stronger, yet sinks under weight. However, the moist soil on the main portion of the beach between the dunes and the water generally is strong enough to drive on.

G. Construction Procedures

Construction procedures influence the ultimate performance of the foundation system because they can alter the properties of the soil and rock. Sloppy construction practices will adversely affect the foundation performance. For example, poor backfilling will reduce the uplift capacity of shallow foundations or lateral capacity of direct embedment foundations. Often, a number of different methods can be used to construct a particular type of

foundation, and a contractor is given the freedom to choose a construction procedure that will allow a competitive bid for the project. This freedom keeps the project costs down and encourages innovation, but care must be exercised that the procedure used does not adversely affect the soil and rock conditions, foundation performance, and adjacent structures or property.

V. BASIC THEORIES FOR FOUNDATION DESIGN

Soil and rock deform under load, causing the supported structure to move. The foundation will fail if the applied load is greater than the strength of the underlying soil or rock (bearing capacity failure) or if the soil and rock deformation results in excessive damage to the structure. Foundations must be designed to avoid both types of unsatisfactory behavior. Generally, a significant amount of foundation displacement is required to mobilize the ultimate bearing capacity (the maximum load that can be applied to the foundation without causing a shear failure of the soil or rock) of a foundation. This displacement may be much greater than the tolerable limits of the structure or equipment.

The description and evaluation of bearing capacity and displacement given below pertain to axial compression loads only. Procedures for other loading modes such as uplift, horizontal, and dynamic loads can be found in the bibliography.

A. General Description of Bearing Capacity

When you observe your footprint on a moist portion of the beach, you notice that along its edges the sand has moved or bulged upward. The grains immediately below the foot are forced downward because of your weight, resulting in the depression of the footprint. To make room for this downward movement, the adjacent soil grains are pushed outward and upward. The greater the resistance offered by the soil to this outward and upward movement, the greater the bearing capacity of the soil and the shallower the depth of the footprint. The frictional resistance of the sand and the confining or vertical stress around the footprint offer resistance to this outward and upward movement. This type of bearing capacity failure is called general shear failure. It develops in soils that exhibit a brittle-type load-settlement behavior (i.e., the failure occurs suddenly), and it is catastrophic under most circumstances.

In saturated soft clays, the edges of the footprint will not show any bulging because the weight causes the foot to simply "punch" through the soil and there is no outward movement of the soil particles. Only the soil directly below the footprint is involved in resisting the load. The greater the resistance to punching, the greater the bearing capacity of the soil and the shallower the depth of the footprint. Resistance to punching in clays is mostly offered by the undrained shear strength of the soil. This type of bearing capacity failure is called punching shear failure, and it develops in soils that exhibit very plastic-type load-settlement behavior (i.e., the failure is progressive, with continuous downward movement or punching of the foundation).

For deep foundations in medium-stiff clays or medium-dense sands, the bearing capacity (tip resistance) failure mode often lies between the two extreme cases of general shear and punching shear failure. In this case, there is both a punching of the foundation into the ground and some lateral and upward movement of the soil particles. Because of the high confining stress at the tip, the upward movement of the particles is restricted and does not manifest itself at the ground surface. This type of failure is called local shear failure, and it develops in soils that exhibit somewhat plastic-type load-settlement behavior (i.e., the failure occurs gradually with significant downward movement).

Bearing capacity of the foundation is governed by the hydraulic conductivity and shear strength of the soil and by the foundation geometry and location below the surface. Bearing capacity failure results when the shear strength of the soil is exceeded. For undrained loading, the undrained shear strength of the soil s_u and the total unit weight of the soil are required to evaluate bearing capacity. For drained loading, the effective stress friction angle, submerged unit weight, and coefficient of horizontal soil stress are required to evaluate bearing capacity.

Deep foundations develop their capacity from end bearing or tip resistance (like shallow foundations) and from side resistance. The side resistance results from the frictional force between the foundation material and soil, and therefore it is a function of the normal (horizontal in case of vertical foundations) force on the foundation and the coefficient of friction between the soil and foundation material. The horizontal force on the foundation normally increases with depth and is a function of the buoyant or effective unit weight of the soil and the coefficient of horizontal soil stress.

B. General Description of Settlement

Foundation geometry influences the volume of soil that will be affected by the foundation load. When shallow foundations are loaded, there is an increase in stress (bearing or contact stress) immediately below the foundation. As one moves farther away below the center of the foundation, the load is spread over a larger area (for a square foundation, this increase in area with depth can be visualized approximately as a truncated pyramid), and the stress gradually decreases. At a depth equal to about twice the

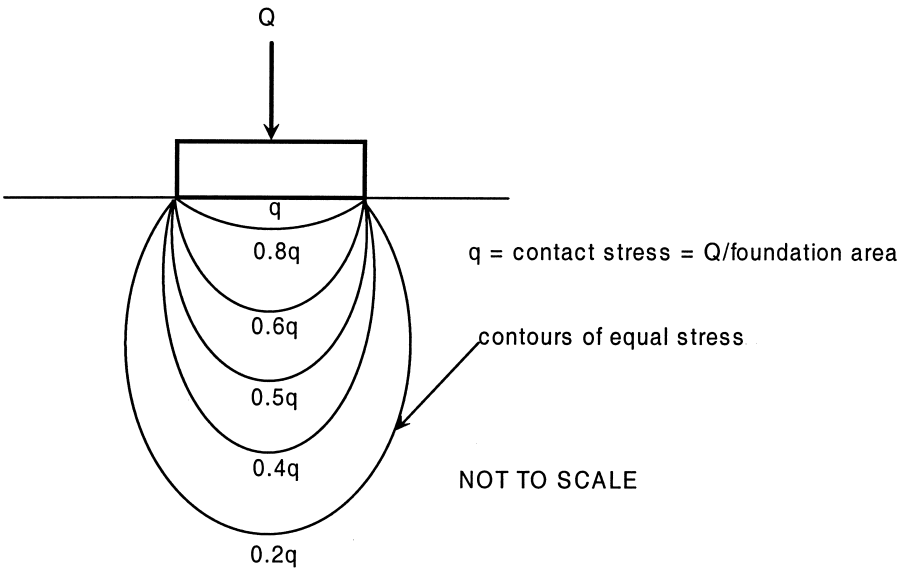


FIGURE 2 Stress bulb.

width of the foundation, the stress decreases to about 10% of the contact stress. Figure 2 shows a contour plot of stress increase below a foundation. These contours also are called pressure bulbs or envelopes.

Foundation settlement results from the magnitude of the stress increase experienced by the soil. Because the zone of influence of wider foundations is greater than that of narrow foundations, they cause a stress increase in deeper layers of soil than do narrow foundations, and therefore, for the same bearing stress, wider foundations will experience more settlement than narrow ones. However, for the same bearing stress, wider foundations will carry more load than narrow ones. If the load is the same on both foundations, the wider one will have a lower contact stress. This contact stress may not be enough to cause significant increase in stress, even at depths equal to half the width of the foundation, and therefore the settlement may be less than for a narrower foundation carrying the same load.

When a load is applied to a foundation, three distinct phases of settlement occur: elastic, consolidation, and secondary settlement. The elastic settlement occurs immediately after load application because of initial stress–displacement response. Concurrently, excess pore water pressure develops in the soil, and there is no increase in effective stress. As this excess pore water pressure dissipates with time, the load is transferred to the soil grain-to-grain skeleton, and the effective stress increases. This transfer process results in consolidation settlement. Further settlement continues after consolidation because the soil skeleton gradually readjusts to the new level of effective stress. This readjustment leads to secondary settlement.

In fine-grained soils such as clays, all three settlement components are distinct. The elastic response occurs as soon as the load is applied; the consolidation settlement is time-dependent and can take years to complete, because of the low hydraulic conductivity of the soil. The secondary response also is time-dependent and continues indefinitely. In coarse-grained soils such as sands, these three components are rarely distinct. Because of the high hydraulic conductivity of sands, the elastic and consolidation settlement occur simultaneously and immediately after load application. The secondary settlements in coarse-grained soils usually are very small and are often disregarded in sands and gravels of sound mineralogy.

In addition to the three settlement components described above, two additional components need to be considered. The first is the elastic deformation of the structural material of the foundation, which normally is minor for shallow foundations but can be significant for deep foundations. The second is the broad category of special cases that has to be evaluated on a case-by-case basis. These include moisture-susceptible soils such as expansive clays and collapsible sands, biochemical decay in landfills and other organic soils, groundwater chemistry that can adversely affect the foundation material, earthquake loading that could lead to liquefaction, and other special local problems.

C. Shallow Foundations

1. Evaluation of Bearing Capacity

The boundaries between the general shear, punching shear, and local shear failure modes described above are not distinct, primarily because the changes in soil behavior modes

are gradual. However, for the relatively low stress levels associated with shallow foundations, general shear failure normally occurs in the medium to denser sands and in most clays during undrained loading.

Different methods of analysis are used for the different failure modes. For the general shear mode, a rational approach based on limiting states of equilibrium is employed. The approach is based on theory and has been confirmed in principle by laboratory and field testing. For the local and punching shear modes, a variety of approaches have been suggested, but none is strictly correct. A more rational approach was suggested by Vesic (Winterkorn and Fang, 1975), which incorporates the stress–deformation characteristics of the soil and is therefore applicable over a wide range of soil behavior. This method is described below.

The ultimate bearing capacity for the general shear mode is computed utilizing the shear surface illustrated in Fig. 3, assuming that the foundation is of infinite length. The soil within the shear surface is assumed to behave as a rigid-plastic medium and is idealized as three zones: an active Rankine zone (I), a radial Prandtl zone (II), and a passive Rankine zone (III). The soil above the foundation is modeled as an equivalent surcharge (i.e., additional weight only), as shown in Fig. 3.

The general solution is known as the Buisman–Terzaghi equation and is given as follows:

$$\begin{aligned}\text{Ultimate bearing capacity} &= q_{\text{ult}} \\ &= cN_c + 0.5B\gamma N_\gamma + qN_q,\end{aligned} \tag{1}$$

where c is the soil cohesion, B is the foundation width, q is the surcharge ($=\gamma D$), D is the foundation depth, γ is the soil unit weight, and N_c, N_γ, and N_q are dimensionless bearing capacity factors that depend upon the friction angle of the soil.

The Buisman–Terzaghi equation is used in either of two derivative forms, which depend upon the soil type and rate of loading. For undrained loading, which develops when loads are applied relatively rapidly to fine-grained soils such as clays, the friction angle $\phi = 0$ and $N_c = 5.14$, $N_\gamma = 0$, and $N_q = 1$. The resulting equation is

$$q_{\text{ult}} = 5.14c + q. \tag{2}$$

In this case the cohesion term c is actually the undrained shear strength of the soil, more commonly denoted as s_{u}.

For drained loading, which develops under most loading conditions in coarse-grained soils such as sands and for long-term sustained loading of fine-grained soils, $c = 0$ and the resulting equation is

$$q_{\text{ult}} = 0.5B\gamma N_\gamma + qN_q. \tag{3}$$

The above two equations apply very well to stiffer or denser soils, where the deformation under load is small, and this corresponds well with the assumed rigid-plastic soil behavior for general shear. For softer or looser soils, the deformation under load becomes large, which is at variance with the assumption of rigid-plastic soil behavior. To account for this variance, Vesic introduced the concept of rigidity index, which accounts for soil deformability. In this method, the rigidity index I_{r} of the soil is compared to the critical rigidity index I_{rc} of the soil–foundation system. The rigidity index is a function of the shear strength parameters c and ϕ, the elastic modulus, and Poisson's ratio of the soil. Fundamentally, it is defined as the ratio of the shear modulus to strength, as modified for volume changes. The theoretically-derived critical rigidity index is a function of the foundation geometry (aspect ratio) and the friction angle ϕ of the soil. When the rigidity index is greater than the critical rigidity index, the soil behaves as a rigid-plastic material, and therefore the general shear failure equations given above can be used. When the rigidity

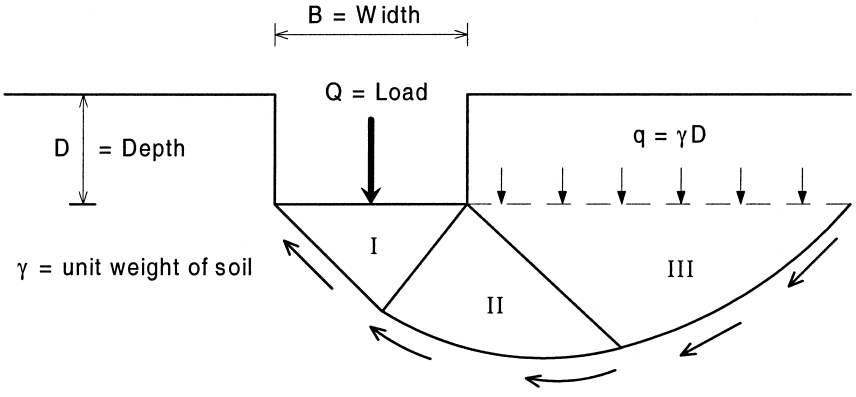


FIGURE 3 General description of bearing capacity.

index is less than the critical, the equation is modifed as follows:

$$\text{Ultimate bearing capacity} = q_{\text{ult}} = cN_c\zeta_{cr} + 0.5B\gamma N_\gamma \zeta_{\gamma r} + qN_q\zeta_{qr}. \quad (4)$$

The ζ_{xr} terms are the rigidity factors and are functions of the soil friction angle, soil rigidity index, and foundation aspect ratio.

The bearing capacity equations developed above are based on the following assumptions:

- Horizontal ground surface
- Horizontal, infinitely long strip foundation at ground surface
- Vertical loading, concentrically applied

Taking into account the variations that occur in practice, the basic bearing capacity equation is modified by a number of factors that separately treat each variation:

$$q_{\text{ult}} = cN_c\zeta_{cr}\zeta_{cs}\zeta_{ci}\zeta_{ct}\zeta_{cg}\zeta_{cd} + 0.5B\gamma N_\gamma \zeta_{\gamma r}\zeta_{\gamma s}\zeta_{\gamma i}\zeta_{\gamma t}\zeta_{\gamma g}\zeta_{\gamma d} + qN_q\zeta_{qr}\zeta_{qs}\zeta_{qi}\zeta_{qt}\zeta_{qg}\zeta_{qd}. \quad (5)$$

Each ζ factor is doubly indexed to indicate the term to which it applies (c, γ, or q) and the phenomenon it describes (r for rigidity of the soil, s for shape of the foundation, i for inclination of the load, t for tilt of the foundation base, g for ground surface inclination, and d for depth of foundation). Although this equation seems formidable at first, it really only represents the inclusion of simple modification factors that depend upon the soil shear strength parameters and foundation geometry. The above equation does not account for the influences of the groundwater table and soil layering, but procedures exist to incorporate them in the design of shallow foundations.

2. Evaluation of Displacements

As stated above, when a load is applied to the foundation, three distinct phases of settlement occur: immediate or elastic (ρ_e), consolidation (ρ_c), and secondary settlement (ρ_s):

$$\text{Total settlement} = \rho_t = \rho_e + \rho_c + \rho_s. \quad (6)$$

In coarse-grained soil, the elastic and consolidation settlements basically occur immediately and simultaneously, and the secondary settlement usually is very small.

The elastic settlement of a foundation is determined by assuming that the soil behaves as an elastic material. The basic approach employed is to use the traditional Boussinesq solution for stress changes in an isotropic, elastic medium subjected to a point load applied normal to the surface of a half-space (the ground is a half-space because it does not exist above the ground surface). The elastic settlement of the foundation is given by

$$\text{Elastic settlement} = \rho_e = [\Delta\sigma(1-\nu^2)A^{0.5}]/(E\beta_z), \quad (7)$$

where $\Delta\sigma$ is the applied stress (Q_c/A, with Q_c the applied compressive load), A is the foundation area, ν is Poisson's ratio of the soil, E is the elastic modulus of the soil, and β_z is the shape and rigidity factor, ranging from about 1.04 for flexible circular foundations to 1.41 for rigid, rectangular foundations with a length-to-width ratio of 10. Real foundations are neither perfectly rigid nor perfectly flexible, but the usual reinforced concrete footings are nearer to rigid, while the usual grillages are nearer to flexible. The above solution is based on a homogeneous soil with constant elastic modulus with depth, both of which are incorrect for real situations. However, procedures exist for modifying the results to account for this variation.

For fine-grained soils, the consolidation settlement often is the major component of the total settlement. As noted previously, consolidation settlement is the time-dependent volume change that occurs as the excess pore water pressure, developed during load application, gradually dissipates and is transferred to the soil skeleton. The consolidation settlement is given by

$$\text{Consolidation settlement} = \rho_c = (\Delta e \cdot H_c)/(1+e_0), \quad (8)$$

where e_0 is the initial void ratio (=volume of voids/volume of solids) of the soil (which can be determined by laboratory tests on undisturbed soil samples), Δe is the change in void ratio caused by the increase in stress from the foundation, and H_c is the height (thickness) of the compressible soil (which can be determined by soil exploratory borings). Although this is a simple expression, the computation of the consolidation settlement requires detailed knowledge of the stress changes occurring in the soil and both the volume change characteristics and the stress history of the soil. The change in void ratio caused by the increase in stress from the foundation is determined by laboratory consolidation tests, the results of which give the stress–deformation (volume change) characteristics of the soil and the amount of time it would take for that volume change to occur. Details of these tests can be found in many soil mechanics (geotechnical engineering) textbooks.

Secondary settlement continues after consolidation settlement is complete, as the soil structure gradually readjusts to the new level of effective stress. This process continues indefinitely. The prediction of secondary settlement is of considerable importance with clays of high plasticity and with micaceous and organic soils, but it typically is of minor concern with coarse-grained soils

and overconsolidated, low plasticity, inorganic clays. Secondary settlement often is computed from the relation

$$\rho_s = C_\alpha H_c \log(t_2/t_1), \tag{9}$$

where C_α is the coefficient of secondary settlement (determined from laboratory consolidation tests), H_c is the thickness of the soil layer, t_1 is the time when secondary settlement begins to be computed, often defined as the time for 90% consolidation settlement, and t_2 is an arbitrary time to which the secondary settlement is computed. Values for C_α typically are very low for overconsolidated soil, but they can be high for soft, normally consolidated soil.

D. Deep Foundations

As the depth of embedment of a foundation increases, the applied compressive loads are resisted by a combination of tip resistance (end bearing) and side resistance, as shown in Fig. 4. When a deep foundation is loaded, the foundation moves downward with respect to the surrounding soil, mobilizing side resistance in the soil. For small loads, small displacements occur, and most of the load is supported by the side resistance. As the load increases, larger displacements occur, and the full side resistance is reached. From this point on, additional applied loads are resisted by the tip. Typically, the displacement needed to mobilize side resistance is in the range of 0.2–0.6 in. (5–15 mm), and it is relatively independent of the foundation dimensions or soil type. On the other hand, the displacement necessary to mobilize the tip resistance is a function of foundation size and can range from about 4% to 10% of the foundation diameter, depending on the foundation type and construction method.

The side resistance is analyzed as a problem in determining the shearing resistance of the foundation–soil interface acting over the surface of the foundation. The tip resistance is treated as an ultimate bearing capacity problem, similar to that for a shallow foundation.

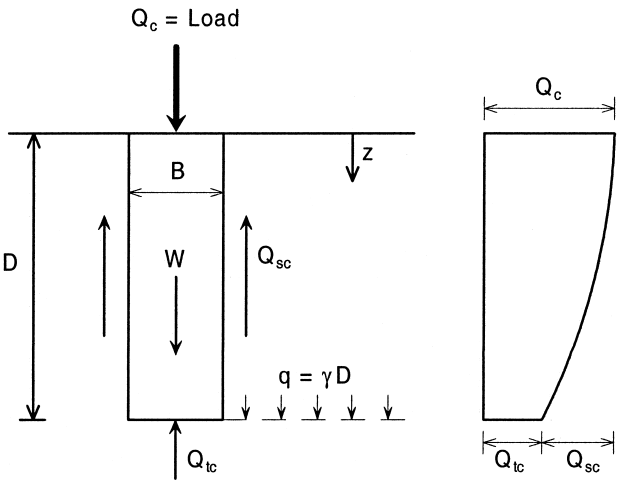


FIGURE 4 General description of deep foundation.

1. Evaluation of Tip Resistance

The tip resistance of a deep foundation is a bearing capacity problem. The maximum tip resistance is given by

$$Q_{tc} = q_{ult} A_{tip}, \tag{10}$$

where q_{ult} is the maximum bearing capacity at the tip (given above by the modified equation for shallow foundations) and A_{tip} is the area of the foundation tip.

2. Evaluation of Side Resistance

The side resistance of a deep foundation is given by

$$Q_{sc} = \int_{surface} \tau(z)\, dz, \tag{11}$$

where $\tau(z)$ is the shearing resistance with depth along the foundation side, and the integral represents a summation with depth of $\tau(z)$ times the foundation surface area.

For drained loading, which develops under most loading conditions in coarse-grained soils such as sands and for long-term sustained loading of fine-grained soils, the cohesion $c = 0$, and the side resistance is evaluated by

$$Q_{sc} = \pi B \int \sigma'_h(z) \tan\delta(z)\, dz, \tag{12}$$

where $\sigma'_h(z)$ is the horizontal effective stress at depth z, which acts as a normal stress on the soil–foundation interface, and $\delta(z)$ is the effective stress angle of friction for the soil–foundation interface at depth z. Expressing the horizontal stress in terms of the overburden stress $\sigma'_v(=\gamma' z)$, we obtain for the above equation

$$Q_{sc} = \pi B \int \gamma' z K(z) \tan\delta(z)\, dz, \tag{13}$$

where γ' is the effective unit weight of the soil (submerged below the water table; dry or moist above) and K is the operative coefficient of horizontal soil stress $(=\sigma'_h/\sigma'_v)$.

The most complex term to evaluate in the above equation is K, which is a function of the original *in situ* horizontal stress coefficient K_0 and the stress changes caused in response to construction, loading, and time. Values of K can range from as low as 0.1 to over 5.

For undrained loading conditions, which develop when loads are applied relatively rapidly to fine-grained soils such as clays, $\phi = 0$ and the above equation can be expressed as

$$Q_{sc} = \pi B \alpha \int s_u(z)\, dz, \tag{14}$$

where $s_u(z)$ is the undrained shear strength, which varies with depth z, and α is the adhesion or reduction factor applied to the average s_u over the shaft depth D to account for construction effects, disturbance, etc. The adhesion factor is a purely empirical factor that has been used to correlate the results of full-scale load tests with the average soil undrained shear strength.

3. Evaluation of Displacements

The settlement analysis of deep, shaft-type foundations is somewhat different from that of shallow, spread-type foundations because of two primary reasons. First, the installation process for deep foundations (driving or "pounding" the pile into the ground or excavating the hole and keeping it open for drilled shafts) causes soil disturbance and alters the state of stress in the soil surrounding the foundation. Second, the load is transferred to the soil in both tip resistance and in side resistance along the foundation shaft. This load transfer process is complex and varies with load intensity. The problem is complicated further by the installation of adjacent foundations and possible group action.

A number of procedures exist for evaluating the settlement of deep foundations, and each of them has its advantages and shortcomings. The procedure recommended by Vesic (Winterkorn and Fang, 1975) is a simple method that gives reasonable results. This procedure is described below.

The total settlement ρ_t of deep, shaft-type foundations is given by

$$\rho_t = \rho_f + \rho_{tt} + \rho_{ts}, \tag{15}$$

where which ρ_f is the settlement caused by axial deformation of the foundation structural material, ρ_{tt} is the settlement of the tip from the tip load, and ρ_{ts} is the settlement of the tip from the shaft side resistance load. Both ρ_{tt} and ρ_{ts} include elastic, consolidation, and secondary settlement components. For coarse-grained and overconsolidated fine-grained soils, the elastic settlement normally constitutes at least 80% of the total settlement, and secondary effects usually are minimal. However, if a compressible soil underlies the foundation, the situation is different. For this case, a settlement analysis similar to that for shallow foundations is performed to evaluate the settlement of the underlying stratum.

The value of ρ_f is determined from conventional formulas for the deformation of a structural material and is given by

$$\rho_f = D(Q_t + \alpha_s Q_s)/(AE_m), \tag{16}$$

where Q_t is the load transmitted at the foundation tip, Q_s is the load transmitted in side resistance along the foundation shaft, D is the foundation depth, A is the foundation cross-sectional area, E_m is the foundation elastic modulus, and α_s is a coefficient that depends on the distribution of side resistance along the foundation depth. For a linearly increasing side resistance with depth, $\alpha_s = 2/3$.

The tip settlement equations for the simplified Vesic solutions employ elastic theory and empirical correlations, resulting in

$$\rho_{tt} = (C_t Q_t)/(Bq_{ult}), \tag{17}$$

$$\rho_{ts} = (C_s Q_s)/(Dq_{ult}), \tag{18}$$

where q_{ult} is the maximum bearing capacity at the tip and C_t and C_s are empirical coefficients, with C_t varying from 0.02 to 0.18 and $C_s = [0.93 + 0.16(D/B)^{0.5}]C_t$.

E. Safety Assessment

1. Factor of Safety

The primary purpose of foundation design is to ensure that the foundation performs satisfactorily during a specified time period. Economic constraints often make it impossible to design for all possible circumstances (e.g., for a major hurricane or nuclear blast). However, the designer does need to consider other factors that can have an adverse effect on the foundation, such as variation in load, unrecognized loads, variation in material strength, unforeseen *in situ* conditions, inaccuracies in design equations, errors from poorly supervised construction, change in system function from original intent, etc.

Traditional design practice does not require a designer to go through the process of considering each of these factors separately and in explicit detail. The designer is required to have a global appreciation of these factors, which ultimately leads to a single factor of safety in design, which is defined as

$$\text{FS} = \text{ultimate bearing capacity}/\text{allowable bearing capacity}, \tag{19}$$

in which the ultimate bearing capacity is evaluated from equations similar to those given above, and allowable bearing capacity is equal to the maximum design stress that can be applied on the foundation. In conventional foundation design, factors of safety between two and three generally are considered to be acceptable when site conditions are reasonably uniform and soil and rock properties are known with some confidence.

The single factor of safety used in foundation design suffers from a major flaw in that it is not unique. It depends upon how it is defined, and it can vary significantly over a wide range. For example, in uplift loading, in which the foundation weight contributes to the uplift capacity of the foundation, certain designers choose not to apply the factor of safety to the foundation weight because it can be calculated fairly accurately by the foundation geometry

and material density. Others choose to include the foundation weight when defining the factor of safety. Also, since the factor of safety is not defined within a consistent and common framework, engineers cannot communicate and share their experiences effectively. This approach also does not distinguish between model (design equation) and parameter (soil and rock property) uncertainty, and therefore makes it difficult to justify any reduction in the safety level if there is additional information about the soil and rock properties or if there are advances in the state of the art in foundation design equations.

2. Reliability-Based Design

An alternative design philosophy is the concept of reliability-based design (RBD), which is also known as LRFD or load and resistance factor design. This approach is well-advanced for structural design, but it is still largely in the development stage for foundation design. In contrast to the single factor of safety approach, factors greater than or equal to one are applied to the various loads, while factors less than one are applied to the soil or rock resistances. These various factors should be established by rigorous calibrations using reliability theory, with an in-depth knowledge of the uncertainties and their variations for both the loads and resistances. Foundation design is slowly progressing toward RBD.

VI. FOUNDATION CONSTRUCTION

It is important for the foundation designer to be familiar with the various equipment and methods used during foundation construction because they can adversely affect the foundation and soil response, project economics, and overall effectiveness of the design. Generally, the contractor chooses the method of foundation construction, and the designer is unaware of the exact details of the method until the time of construction. There may be occasions when the designer specifies a particular method of construction, for example, to protect existing structures in close proximity to the new foundation, but this will almost certainly add to the cost of construction. Whatever method of construction is used, the key to ensuring that the design methods are appropriate is to ensure that the construction is done properly for that method. This approach requires that the specifications describe clearly the proper construction techniques that result in quality construction and also allow the contractor as much freedom as possible.

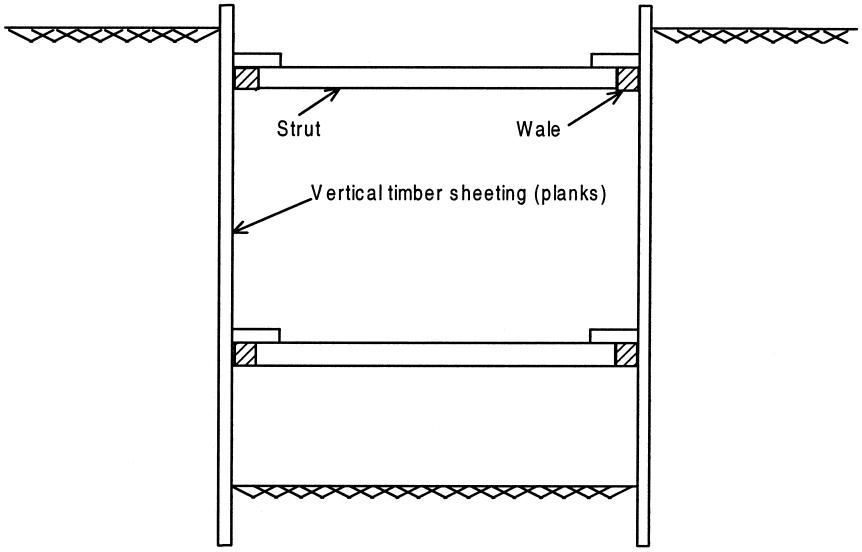
A. Shallow Foundations

Construction of spread footings involves excavation of the hole, placement of the required forms, reinforcing rods, and concrete, and then backfilling of the hole with the excavated soil. The forms are used to contain the concrete within the specified dimensions. Care must be exercised that the hole is not overexcavated and that the bottom of the hole is subjected to the minimum amount of disturbance during the placement of the forms and reinforcing rods. If the hole is overexcavated too deeply beyond that called for in design, it should be backfilled with concrete and not with the excavated soil. This approach will prevent additional settlement from occurring.

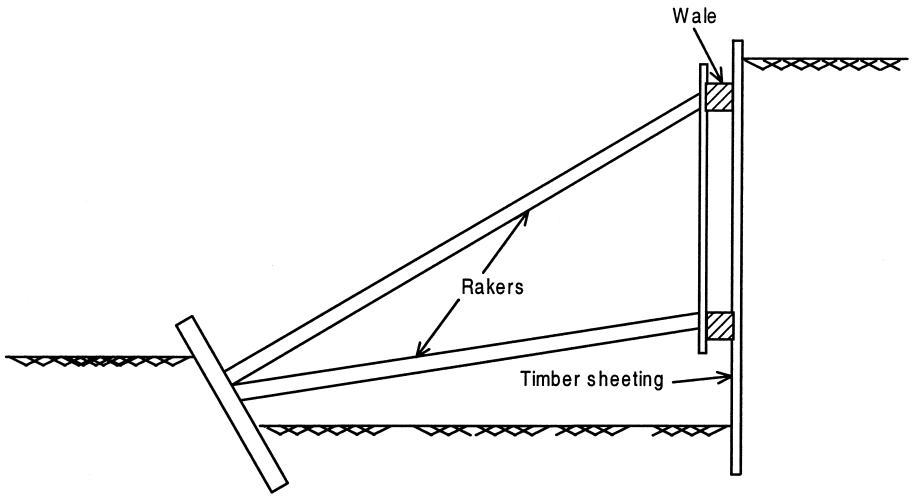
Excavations for shallow foundations that are up to 5 ft (1.5 m) deep generally do not require support for the excavation walls. For these shallow depths, vertical cuts can be made in moist silts, clays, and sands containing little amounts of silt or clay. In dry sands and gravels, the walls can be sloped back for stability. Because of worker safety issues, excavations deeper than about 5 ft (1.5 m) either have to be sloped back or require wall support even in clays, where vertical cuts can be made up to 20 ft (6.0 m) or more. The steepness of the slopes depends upon the soil or rock characteristics, depth of the excavation, amount of time the excavation is to remain open, and weather conditions.

When the excavation is close to the property line or adjacent to existing buildings, the sides of the excavation must be made vertical and have to be structurally supported. For individual spread foundations, the excavation generally is less than about 10 ft (3.0 m) deep, but if a single-story or multistory basement (as for an underground parking facility) is to be constructed, the excavation can be more than 15 ft (4.5 m) deep. For shallow (less than 10–12 ft or 3.0–3.7 m in depth) vertical cuts, vertical timber planks (sheeting) can be used to support the excavation walls. The plank is driven with its wider side against the soil face to a few feet (up to 1 m) below the surface. As excavation proceeds, the sheets are driven farther into the ground so that they are always a few feet below the bottom of the excavation. Horizontal beams (wales or walers) are placed close to the top and bottom of the hole to support the sheeting. In narrow excavations, horizontal timber or pipe struts that run from side to side, as shown in Fig. 5a, support the wales. In wider excavations (such as for basements), inclined struts (rakes or rakers) can support the wales, as shown in Fig. 5b.

For excavations deeper than 12–15 ft (3.7–4.5 m), timber sheeting becomes uneconomical, and other alternatives such as soldier pile and lagging or steel sheet piles can be used. If a vertical face of about 30–50 ft^2 (2.78–4.65 m^2) of soil can be left unsupported for a short period without collapsing, a system of soldier piles and lagging can be used to support the vertical cuts. Steel H-sections (soldier piles) are driven into the ground with their flanges (flat part) parallel to the sides of the excavation to a depth a



a. Horizontal bracing for narrow shallow excavations



b. Inclined bracing for wide shallow excavations

FIGURE 5 Bracings for shallow excavations. (a) Horizontal bracing for narrow, shallow excavations. (b) Inclined bracing for wide, shallow excavations.

few feet below the maximum depth of the excavation and at a spacing of about 4–8 ft (1.2–2.4 m). As excavation proceeds, horizontal timber boards with their wide sides against the soil are wedged behind the exposed flanges of the soldier piles, as shown in Fig. 6. Wales and struts then are placed at regular intervals to support the soldier piles. For wide excavations where horizontal struts may not be feasible, inclined braces have to be used. Under these circumstances, it may be possible to excavate the central portion to the maximum depth, leave a bench (berm) of soil around the perimeter to support the soldier piles, and construct a portion of the foundation to support the inclined braces. Once the braces are in place to support the soldier piles, the berms can be excavated. Under certain conditions, it may be feasible to use an anchor system (also called tiebacks) to support the soldier pile wall (see Fig. 7)

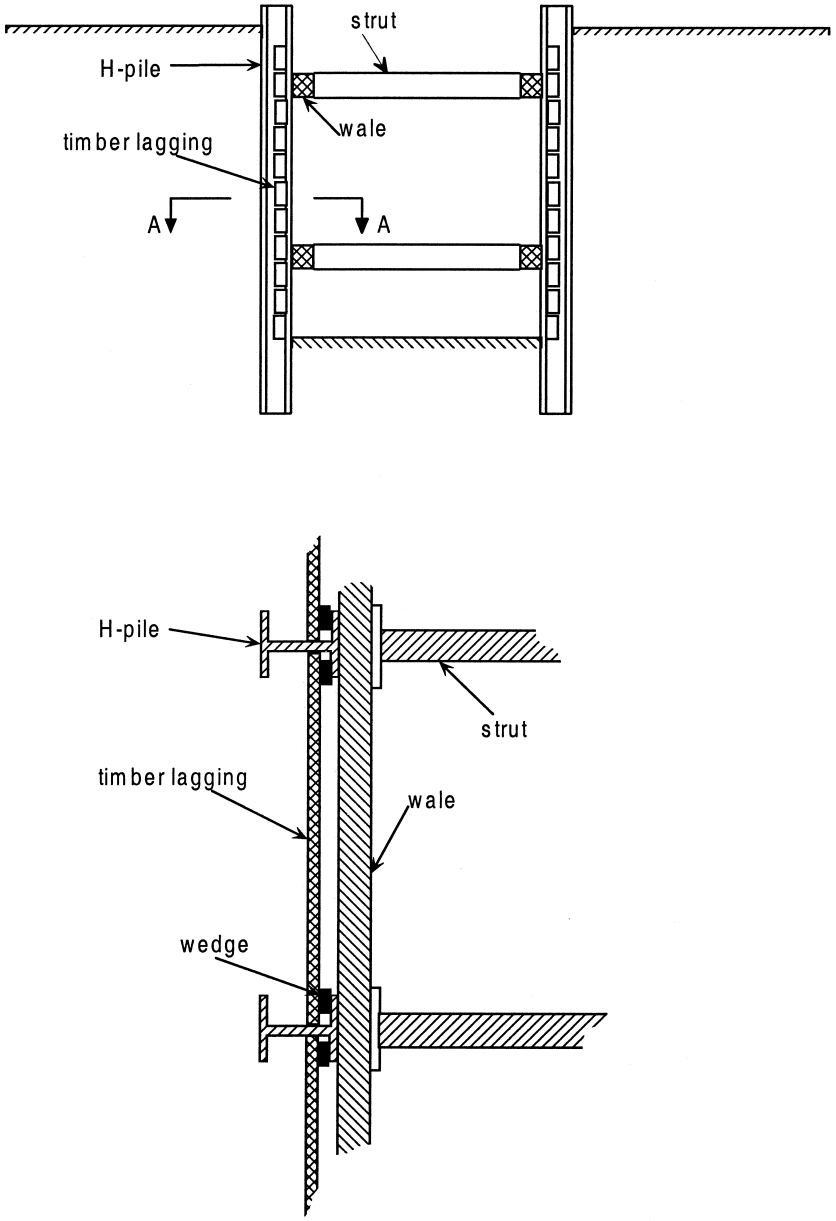


FIGURE 6 Soldier pile and lagging.

instead of the inclined braces or rakers. These tiebacks are constructed by drilling an inclined hole into the soil, placing a reinforcing rod in the hole, and then backfilling the hole with concrete. Once the concrete has set, the reinforcing rods generally are prestressed before excavation proceeds.

If the soil conditions are such that a soldier pile system cannot be used (e.g., clean, dry sand that will run as soon as it is excavated or sand below the water table), steel sheet piles can be driven to support the vertical sides of the excavation. Interlocking steel sections are driven into the ground around the perimeter of the excavation to a depth a few feet below the final depth of the excavation. Wales and struts (or tiebacks) then are placed at regular intervals, as described above for the soldier pile and lagging system.

B. Deep Foundation

1. Pile Foundations

a. General. Wooden pegs for supporting a camping tent generally are driven into the ground by a hand-held hammer. If the ground is hard, the hammer blows can

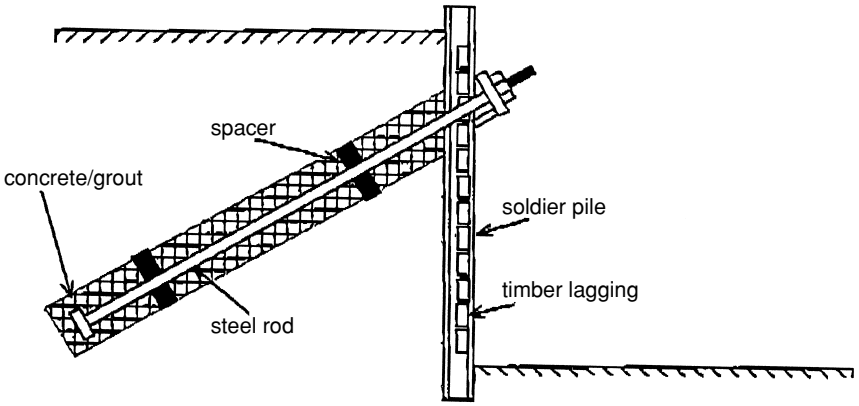


FIGURE 7 Tieback system.

damage the top of the peg, and the bottom of the peg can be crushed or split because of the ground resistance. Similarly, pile foundations are driven into the ground by impact or vibratory-type hammers. If the ends of the pile are not protected adequately, the hammer blows can damage the top or butt of the piles, and the resistance offered by the ground can damage the tip of the pile. Special attention also must be paid to pile storage and handling, placement tolerances, driving, and extraction to ensure a properly constructed pile foundation.

Piles can be damaged during storage and handling, which is especially true for precast concrete piles that can crack or develop a camber (permanent bending) if they are supported or lifted at locations away from the designated pickup points. Excessive camber can cause the pile to be driven out of plumb and be overstressed when loaded. Special care also must be exercised when handling piles with protective coatings, and the damaged areas must be repaired prior to installation. Generally, all piles are visually examined at the driving location by a qualified inspector immediately prior to installation.

Piles have to be driven close to the location and inclination designated on the construction plans; otherwise, they will be unable to support the design load with an adequate margin of safety. However, it is impractical to drive a pile exactly at the given location and inclination, and the engineer must allow for reasonable placement tolerances and account for these deviations in the design. These tolerances need to be designated in the construction specifications.

Perhaps the single most important aspect of pile construction is the process of driving the pile into the ground. Pile driving should not result in permanent deformation of steel H-piles or steel pipe piles, crushing or spalling of concrete piles, or splitting or brooming of wood and fiber glass piles. Pile damage can be avoided if the proper equipment is utilized for the prevailing subsurface conditions and the type and capacity of pile being used. Even if proper equipment is used, piles can still be damaged if the pile and hammer are misaligned.

Sometimes, when piles are driven through hard layers of dense sand, cobbles, boulders, or other obstructions, it may be necessary to supplement pile driving with the help of pile shoes, jetting, or preboring to prevent damage to the pile from overdriving. These methods usually result in reduced pile capacity and therefore should be used with caution.

Specially designed pile shoes (tip protector/reinforcer) improve driveability and also provide increased cutting ability and pile tip protection. These shoes usually are fabricated from cast steel. Jetting is accomplished by pumping water under high pressure through pipes attached to the pile. The water loosens and removes the soil and thereby allows the pile to be driven more easily. When hard layers of soil are encountered, drilling a pilot hole or preboring can reduce the driving resistance. Preboring can also be used when passing through materials such as cobbles and boulders that can deflect the pile.

When pile damage is suspected during installation, it may be necessary to extract the pile for inspection. Extraction also may be required if excessive drift occurs during driving and the pile location and inclination exceed the specified placement tolerances. Pile extraction is an expensive and difficult process that requires experienced personnel. Care should be taken that the tensile strength of the piles, especially concrete piles, is not exceeded during the extraction process.

b. Pile-driving equipment. Typical pile-driving equipment includes a crawler-mounted crane (see Fig. 8) with a boom, leads, hammer, driving caps, and other accessories all connected to form a unit. This unit must be

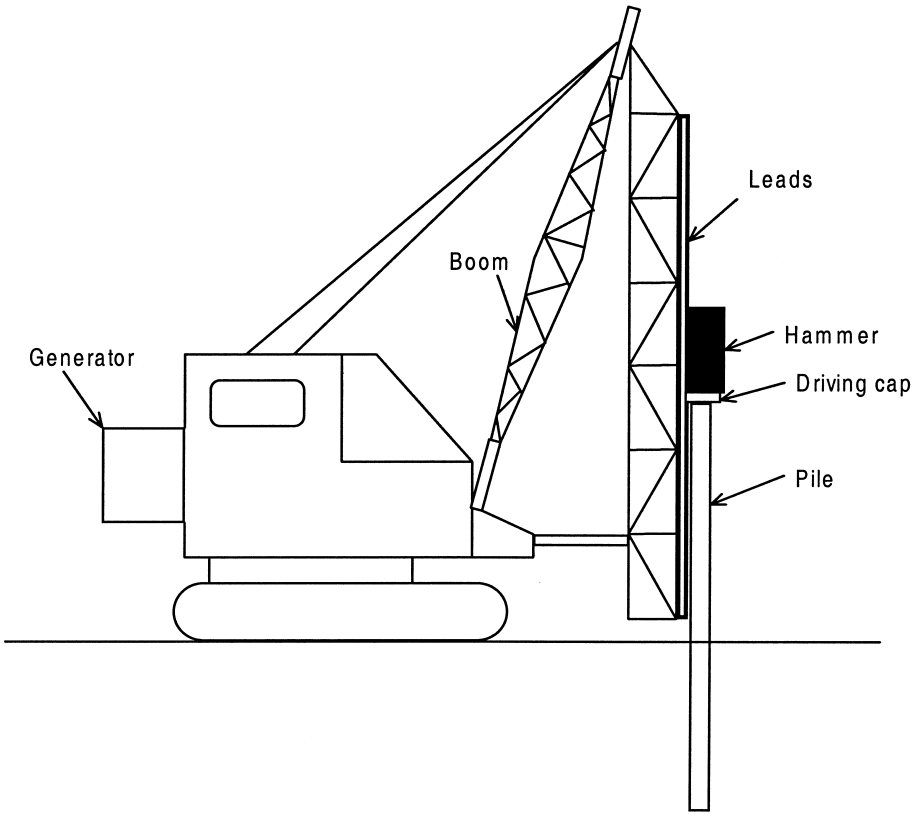


FIGURE 8 Pile-driving rig.

strong enough to guide and drive each pile accurately to its final position and withstand safely all loads imposed during the pile driving process.

Pile hammers can be grouped under impact (drop hammer, steam or air hammer, diesel hammer) or vibratory types. Drop hammers are the simplest and oldest type of impact hammers. They consist of a guide weight or ram that is lifted up to a required height by a hoist line and then released. The drop hammer is best suited for small projects that require relatively low capacity and employ lightweight timber or steel piles. Other types of impact hammers include the steam or air hammer, in which steam or compressed air pressure is used to raise the ram. The ram falls freely and strikes the drive cap when the pressure is released automatically. These hammers can be used for a variety of different pile–soil combinations, but they are best suited for timber and H-piles in sand.

Diesel hammers are self-contained, lightweight, economical, and easy to service. Initially, the ram is raised by mechanical means, and the fuel is injected into the cylinder while the ram drops. When the ram strikes the anvil, the fuel is ignited, explodes, and forces the anvil down against the pile and the ram upward. The sequence repeats itself automatically, and the pile is driven into the ground by the ram impact and the explosion of the fuel. Diesel hammers are best suited for hard driving conditions and do not work well in soft soil because of the driving resistance required for compression and ignition of the fuel.

Vibratory hammers utilize eccentric weights that are rotated with the help of electric or hydraulic motors to produce vertical vibrations. The vibrations reduce the frictional soil resistance along the sides of the pile and also loosen the soil at the tip . This type of hammer is best suited for steel piles in sand. Clay soils dampen the vibrations of the hammer and therefore reduce pile penetration. These hammers are not effective in penetrating soils with large obstacles and boulders and are not suitable for concrete or timber piles.

Pile-driving leads are used to align concentrically the pile head and hammer, provide lateral support for the pile, and maintain proper pile alignment and position during driving. If the hammer is not aligned properly with the pile head, eccentric loading can cause excessive stresses that result in structural damage of the pile. Also, the driving energy transferred to the pile will be reduced. Leads or leaders are usually fabricated of steel.

Driving caps are placed between the pile head and the hammer and transfer the driving energy from the hammer

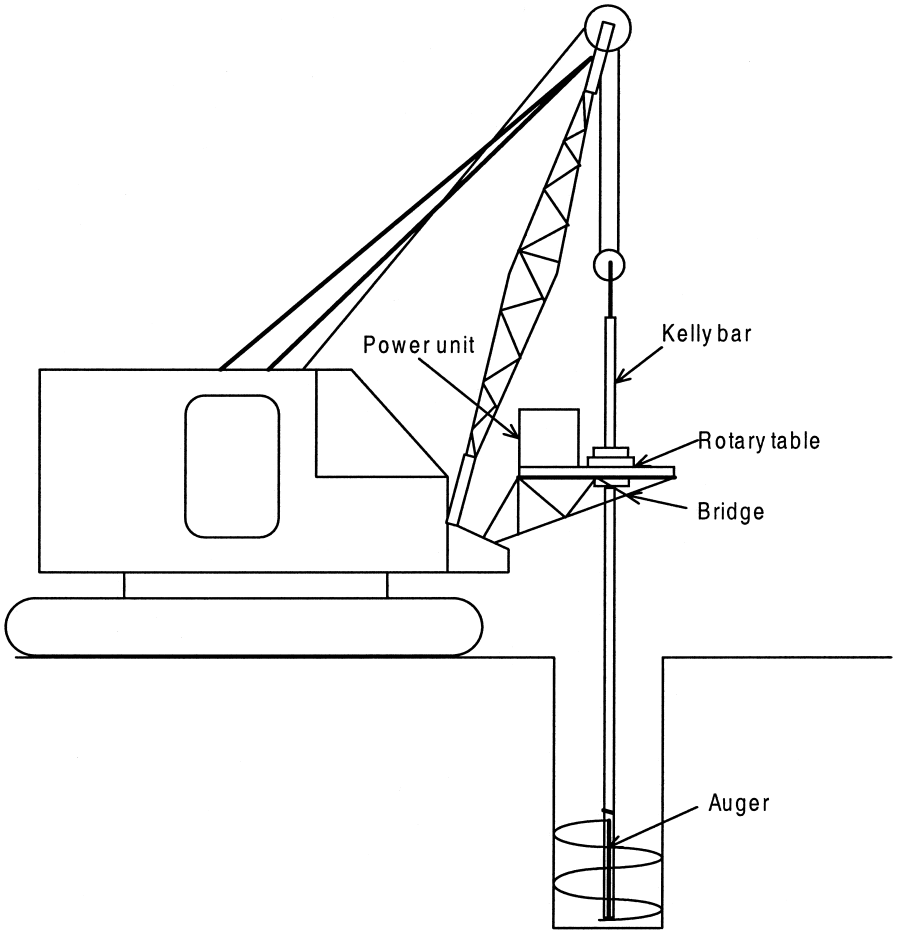


FIGURE 9 Drilling rig for drilled shafts.

to the pile without causing distress to the pile. Driving caps consist of a pile cushion placed directly on top of the pile butt, a helmet placed on top of the pile cushion, and an anvil placed on top of the helmet. Sometimes an additional cushion is placed between the helmet and the hammer ram to protect the hammer from the shock wave reflected back to it. Plywood and oak boards normally are used for pile cushions, and thick blocks of hardwood are used for the hammer cushion.

2. Drilled Shafts

a. General. Drilled shaft construction generally involves excavating a cylindrical hole in the ground, providing a suitable method of keeping the hole from caving in, placing tied steel reinforcing rods (also called a rebar cage) in the excavated hole, and then filling the hole with concrete. Drilled shaft construction can be classified in three broad categories, dry, casing, and wet, according to the method used for keeping the hole open during excavation. The method selected for construction depends upon the soil in which the shaft is being constructed. The performance of a drilled shaft depends on the construction method, and therefore it is included as part of the design process.

The most common excavation method used for drilled shafts is by rotary drilling, although in some developing countries hand excavation also is used. The rotary drilling machines are either truck-mounted or crane-mounted (see Fig. 9) and can be equipped with a number of different types of drilling tools, such as a drilling bucket, flight auger, rock auger, core barrel, or shot barrel. The type of drilling tool used depends upon the type of soil or rock being excavated.

b. Construction method. Drilled shafts generally are constructed by the dry, casing, or wet method or a combination of these methods, depending upon the type of ground encountered during excavation. The dry method can be used above the water table for soil and rock that will not cave or slump when the hole is drilled. Above the water table, stiffer clay and sands mixed with some

cohesive material are best suited for dry construction. This method sometimes can be used for low-permeability soils below the water table if the concrete is placed immediately after the hole has been excavated. A short piece of casing commonly is used near the surface during dry construction to prevent the loose surface soil from caving into the hole. Generally, this casing is corrugated steel pipe and extends some distance above the ground surface. The casing also serves as a guide for drilling tools, safety barrier for the workers, and means for preventing material falling into the hole. After the hole is excavated to the required depth, the bottom of the hole is cleaned to remove the loose material. Hand cleaning sometimes is performed, but this should be avoided as much as possible because of safety considerations. Special regulations have to be followed if a person is lowered into the hole for hand cleaning or inspection. If the bottom of the hole has to be enlarged, or belled, to form a larger bearing area, an underreaming (belling) tool can be used.

The casing method is used for drilled shaft construction when excessive caving and deformation of the soil or rock will occur during excavation. A casing consists of a simple steel pipe or tube that is either driven or vibrated into the ground, either before drilling or as the drilling proceeds. Casings also are used when the drilled shaft extends through water-bearing soil into rock. The casing is driven or vibrated through the soil to form a seal at the soil–rock boundary that will prevent water from flowing into the excavated hole. When large obstructions such as boulders are anticipated, the casings can be telescoped. Telescoping involves placing a casing with a larger diameter than that required for the drilled shaft to a certain depth below the surface. A smaller diameter borehole is then drilled and cased below the bottom of the first casing. This process then can be repeated to the desired depth of the drilled shaft.

Casings can be temporary or permanent, depending on the type of soils involved. Temporary casings are withdrawn as the concrete is placed in the shaft, but they may have to be left in place if it is difficult to extract them. This smooth-wall casing can reduce the side resistance of the shaft and therefore its overall load-carrying capacity. Permanent casings often are used for very soft and loose soil conditions and are left in the ground after the concrete is placed. Construction with permanent casing is more expensive than for temporary casing.

An alternative to the casing method is the wet or slurry-displacement method, which can be used in any of the soil conditions described for the casing method. This method generally involves use of slurry to keep the excavated hole stable for the total depth of the excavation. Either mineral slurry or polymer slurry can be used for this purpose. The wet method is more feasible than the casing method for very deep holes and when it becomes difficult to seal casing in rock or impermeable soil below water-bearing granular soils. This method can be adapted for construction of drilled shafts where the surface of the ground is under water, as in a lake or river.

When minerals such as bentonite are mixed with water, they form a suspension of platelike solids in the water that is called slurry. It also can be formed with certain polymers. When the fluid pressure of the slurry exceeds that of the ground water in permeable soil, the slurry displaces the water, penetrates the soil, and forms a membrane on the walls of the hole. This membrane or mudcake helps to keep the hole stable. The effectiveness of the slurry depends on the proper amount of bentonite or polymer used and the chemistry of the water with which it is mixed. The use of slurry involves a significant additional cost resulting from the materials, handling, mixing, placing, recovering, cleaning, and testing of the slurry. For mineral slurry, there is an additional cost for properly disposing of the slurry because of environmental concerns.

Basically, the wet method either can be a static process or a circulation process. In the static process, the cuttings are lifted up by the drilling tool, and in the circulation process the cuttings are brought to the surface with the circulating slurry. Once excavation has been completed to the required depth, it is important to check that too much soil material is not in suspension within the slurry; otherwise, the particles can settle to the bottom and form a loose bearing surface for the drilled shaft.

c. Drilling rigs and tools. Excavations for drilled shafts are performed most commonly by rotary methods using some form of a rotary drilling unit or rig. These drilling units are mounted on trucks, cranes, crawler tractors, or skids. The type of drilling rig that is best suited for the job depends on the soil and rock conditions at the site and the size of the foundation. The capacity of the rig is determined by the amount of downward force (crowd) and the twisting force (torque) it can apply. A steel drive shaft called a kelly bar is turned by the power unit and transfers the crowd and torque to the drilling tools. The crowd can be provided by the weight of the kelly bar and the drilling tool or by other mechanical or hydraulic means. Although the crowd and the torque are important factors for achieving the best drilling rate, progress can be hindered if the proper drilling tool is not used.

Tools for rotary drilling include drilling buckets and flight augers for soil excavation and rock augers, core barrels, shot barrels, and full-faced excavators for rock excavation. To form an enlarged base, underreamers or belling buckets can be used. Drilling buckets commonly are used for excavation in granular soils and are equipped with cutting teeth at the bottom of the bucket. Two openings at the

bottom of the bucket allow the excavated soil to enter the bucket. Flaps inside the bucket prevent the excavated soil from falling out of the bucket. When the bucket is full, it is raised and withdrawn from the hole, and the hinged bottom is opened to remove the excavated soil. These buckets also can be used with drilling slurries.

Flight augers are best suited for cohesive soils and soft rock and look like enlarged versions of a carpenter's drill bit. A cutting edge is fixed to the bottom of the auger that drills through the soil or rock. The cuttings are brought to the surface by traveling up the flights of the auger in much the same fashion as the wood shavings are brought to the surface when a hole is made in wood with a drilling bit. The cutting faces on the augers are aligned to result in a flat base at the bottom of the excavation. The bottoms of the comparable rock augers are equipped with tungsten carbide conical teeth that are used to break up hard rock. The flights of the rock augers are made of thicker metal than soil augers. Augers generally are not well suited for slurry construction, and care must be exercised when lowering or removing the auger from a slurried hole. Otherwise, differential pressures can be developed above and below the auger that can cause instability of the walls of the shaft and result in a cave-in.

Sometimes the rock is too hard to be excavated by a rock auger. Under these conditions, a core barrel can be used for excavating the hole. One of the simplest types of core barrels consists of a cylindrical steel tube equipped with hard metal teeth at the cutting edge. The tube cores into the rock, and the rock core is brought to the surface by lifting the core barrel to the surface. Friction from the cuttings holds the core in place inside the barrel. Rock coring is a slow process. If the hardness of the rock makes rock coring ineffective and only a minor penetration into rock is required, a shot barrel can be used. This barrel is similar to the core barrel but has a plain bottom. Hard steel shot are fed below the base of the barrel and grind the rock as the barrel is rotated.

d. Concrete placement. Concrete can be segregated and lose its strength if it is not placed properly inside the excavated hole. In dry holes, concrete generally is allowed to fall freely inside the hole, but it can be segregated if it strikes an obstruction such as a rebar cage or the sides of the hole. To prevent segregation of free-falling concrete, a flexible hose generally is used to direct it toward the center of the hole. Some studies have shown that the impact of the free-falling concrete at the bottom of the hole has a beneficial effect on the strength of the concrete because it drives air out of the voids, making it denser and therefore stronger.

When the excavated hole is not completely dry and is partially or fully filled with water or slurry, it is not feasible to place the concrete with the free-fall method. Also, the free-fall method is not recommended when the shafts are not vertical. Under these circumstances, concrete is placed through a tremie pipe. This pipe is made of steel and is equipped with a flap or plug at its bottom end to prevent the concrete from being contaminated by the slurry or water as the tremie is lowered into the hole. The concrete is allowed to flow through the tremie by opening the flap or removing the plug only when it reaches the bottom of the hole. Because of the large difference between the unit weights of the concrete and the slurry, the slurry is displaced by the concrete and floats up above the concrete, as long as the head of concrete is larger than the slurry head (head at a point is the unit weight of the material times its height above the point). As the column of concrete rises in the hole, the tremie pipe is raised gradually to allow reasonable flow of concrete through the tremie. Care must be exercised that the bottom of the tremie stays well below the top of the column of concrete; otherwise, pockets of slurry can get trapped within the concrete and weaken the shaft.

e. Effects of construction method. Construction procedures have a significant influence on the performance of the drilled shaft. Even the length of time the hole is allowed to remain open before concrete is placed can have a negative impact on the load-carrying capacity of the drilled shaft. For dry hole construction, the soil immediately surrounding the shaft expands and gets weaker with time. For slurry construction, if the slurry is allowed to remain standing in the hole without circulation, it can form a cake on the sides of the excavation and reduce the frictional side resistance of the shaft.

Frictional resistance between smooth steel casing and soil is lower than the frictional resistance between rough concrete and soil. Therefore, when temporary steel casings are left in place after concrete placement, the side resistance of the shaft will be decreased. All of these factors need to be considered in the design process.

VII. QUALITY ASSURANCE AND CONTROL

The success of a foundation system depends on the care exercised during its construction, and therefore quality assurance and quality control in the field are extremely important. Qualified personnel (inspectors) should be present on the site for verifying the soil conditions and inspecting the construction materials and procedures. On large projects, it might be feasible to perform full-scale load tests on the foundations to verify the design assumptions.

Verification of subsurface conditions is important because the foundation capacity and construction procedures

depend on the soil and rock properties. Soil conditions can be verified easily during construction of shallow foundations because the soils are exposed during construction. For drilled shafts and anchors, soil conditions can be verified by the cuttings brought to the surface. For pile foundations, the soil conditions cannot be determined by visual examination but they can be inferred from the performance of the pile as it is being driven into the ground.

Materials used for foundation construction, such as reinforcing rods, concrete, piles, etc., are inspected prior to use to ensure that they meet the required specifications. The type, number, and size of both the longitudinal and transverse reinforcement are verified prior to placement in the excavated hole. Tests on concrete are performed for verifying the temperature, air content, and slump (amount of water) of the concrete, and cylinder samples are prepared for compression testing in the laboratory. The field tests are performed for verification only and not for control. Control tests on the concrete are performed at the concrete batch plants. Piles are inspected prior to installation to verify the condition, dimensions, and straightness of the pile. The pile is rejected if it does not meet the required specifications.

Full-scale load testing of a foundation is expensive and often is feasible only for large projects or critical structures. The success of the load test depends on the experience and qualifications of both the contractor and the designer and, if proper care is not exercised during the conduct of the load test, the results would be meaningless. Generally, foundation load tests are conducted for construction control or for verifying that the design capacity is comparable to the capacity of the as-built foundation. These tests are called "proof" tests, probably because they are envisioned as the "proof of the pudding." Load tests also can be performed during the design stage to obtain design parameters or for verifying the design assumptions.

A number of different procedures are used for foundation load testing, and the results obtained are influenced by the testing procedure. The simplest form of load test consists in applying a specified load at the top of the foundation and measuring its corresponding displacement. However, load tests have been performed with sophisticated instruments to measure the load (or stress) and displacements (or strains) along the length and at the tip of deep foundations and in the surrounding soil or rock. It should be noted that the results of the load tests are only applicable to the particular foundation being tested and the prevailing soil or rock conditions at the time of the test. Extrapolation of the load test results to other foundations on the project must be done with caution, especially if there is a change in construction procedure or soil and rock conditions. Also, load tests are performed over a short time duration compared to the life of the foundation. Extrapolation to long-term response should be done with care because soil conditions can change with time (e.g., change in ground water elevation) and because long-term response is governed by soil parameters different from those that govern short-term performance.

SEE ALSO THE FOLLOWING ARTICLES

CONCRETE, REINFORCED • CONSTITUTIVE MODELS FOR ENGINEERING MATERIALS • DAMS, DIKES, AND LEVEES • EARTHQUAKE ENGINEERING • GEOENVIRONMENTAL ENGINEERING • MASONRY • MECHANICS OF STRUCTURES • ROCK MECHANICS • SOIL MECHANICS

BIBLIOGRAPHY

Bell, F. G. (1987). "Ground Engineering Reference Book," Butterworths, London.

Broms, B. B. (1998). "Foundation Engineering," www.geoforum.com/knowledge/texts/broms/index.asp.

Canadian Geotechnical Society. (1992). "Canadian Foundation Engineering Manual," 3rd ed., BiTech, Richmond, BC, Canada.

Fang, H.-Y. (1991). "Foundation Engineering Handbook," 2nd ed.,Van Nostrand Reinhold, New York.

Handy, R. L. (1995). "The Day the House Fell," ASCE Press, New York.

Macaulay, D. (1976). "Underground," Houghton Mifflin, Boston.

Peck, R. B., Hanson, W. E., and Thornburn, T. H. (1974). "Foundation Engineering," 2nd ed., Wiley, New York.

Sowers, G. F. (1979). "Introductory Soil Mechanics and Foundations; Geotechnical Engineering," 4th ed., Macmillan, New York.

Winterkorn, H. F., and Fang, H.-Y. (1975). "Foundation Engineering Handbook," Van Nostrand Reinhold, New York.

Witherspoon, W. T. (2000). "Residential Foundation Performance," ADSC, Dallas, TX.



Light Sources

Graeme G. Lister
Osram Sylvania Inc.

John F. Waymouth
Consultant

I. Introduction
II. Overview and Historical Perspective
III. Electromagnetic Radiation
IV. Light
V. Incandescent Lamps
VI. Tungsten-Halogen (T-H) Lamps
VII. High-Intensity Discharge (HID) Lamps
VIII. Fluorescent Lamps
IX. Low-Pressure, Non-Mercury Lamp
X. Luminescent Light Sources
XI. Summary

GLOSSARY

Arc Name applied to an electric discharge in a gaseous medium, originally derived from the bowed shape assumed by a horizontal, unconfined discharge, as a result of upward convective displacement of the center while ends remain anchored to electrodes.

Ballast A device that, by means of reactance or resistance, singly or in combination, limits the current of a lamp to the required value for proper operation and, where necessary, provides adequate starting voltage; and in the case of rapid start circuits, provides for low voltage lamp filament heating.

Cold spot Coldest spot on the wall of a discharge lamp. This determines the vapor pressure of the principal radiating species.

Color rendering index (CRI) Measure of the capability of a light source to illuminate colored objects in such a way that the colors are perceived as "normal" (see Section IV).

Color temperature Measure of the color of the light emitted from a light source, being the temperature of the blackbody radiator having the same color when evaluated by a CIE (Comission Internationale de l'Éclairage) observer. If an exact color match is not possible, the closest match is designated *correlated color temperature* (CCT). This distinction is frequently ignored in practice.

Dumet Composite material used for lead-in wires in glass-to-metal seals for lamps using soda-lime glass envelopes, comprising a core of low thermal expansion nickel-iron and a cladding of high-expansion copper, in such proportions that the net expansion matches that of the glass.

Efficacy Measure of performance of light sources and lighting systems, equal to the ratio of light output to electrical power input, normally given in lumens per watt (lm/W). Since this ratio is not dimensionless, the word *efficiency* is avoided for this quantity.

Getter Active chemical compound that reacts and removes impurities such as water, oxygen, and hydrogen from lamps.

Hard glass Generic jargon term for glasses having softening and working temperature greater than soda-lime glass, but less than fused silica (quartz), e.g., borosilicate glasses.

Lead-in wires Component of a lamp through which electric current is introduced, through a glass-to-metal seal, into the interior of a hermetically sealed envelope. In industry jargon, these are termed lead wires, but that term is not used here to avoid confusion with the element lead.

Luminance Broadly speaking, a measure of the intensity of light. Specifically, it is the luminous evaluation of power associated with a pencil of rays, equal numerically to the luminous flux in lumens per area, per solid angle (formerly, brightness or photometric brightness).

Polycrystalline alumina (PCA) Translucent, ceramic material used for discharge tubes of high-pressure sodium lamps and some newer metal halide lamps (see Section VII).

Press seal Glass-to-metal seal in a lamp, fabricated by mechanically pressing softened glass to and around the electrode lead-in wires. In fused silica (quartz), used in tungsten-halogen lamps (see Section VI) and high-intensity discharge lamps (see Section VII), the lead-in wires take the form of thin foils of molybdenum.

Rated life Usually the burning hours of a large test batch of lamps when 50% have failed.

Resonance radiation Radiation emitted in an atomic transition, in which the lower state of the transition pair is the lowest energy state (ground state) of the atom. A large fraction of the atoms of any species in a lighting discharge are in the ground state.

Stem Component of a lamp incorporating the glass-to–metal seals of the lead-in-wires, together with mating surfaces of the glass, ready for flame-fusion sealing to the glass envelope. The stem usually includes a tube, through which air from the envelope interior can be exhausted after the stem is sealed to the envelope.

Wall-loading Measure of the power input to a discharge lamp, in watts per unit area of the inner wall surface. An important design parameter for most discharge lamps.

NEXT TO THE INVENTION of the wheel, humanity's oldest and most important invention is the artificial light source, a means or device for converting some other form of energy into light. From the dawn of human history until about 200 years ago, such devices could convert only chemical energy into light, through the medium of fire. For more than 100 years, commercially useful electric light sources have been available—a technological accomplishment of stupendous magnitude—that has transformed our lives. Electric lamps produce the light that makes it possible to carry on all forms of human activity far outside the sunrise-to-sunset time limitation faced by earlier societies, greatly increasing our productivity. Light sources make our streets and highways safer after dark, and they decorate and illuminate all manner of public places. Finally, they do so at a cost to us of one-quarter percent of our personal product, individually, or of the gross national product collectively.

I. INTRODUCTION

Mankind learned to use fire, rather than flee from it, more than half a million years ago, and the first portable light source was developed by stone age man, burning animal and vegetable oils in a bowl fitted with a wick. This was the precursor of the candle, probably invented in Roman times. Light was produced by such primitive means until the advent of the Industrial Revolution, around 1780, which led to a century of development in lighting technology based on the oil lamp. Together with developments in the use of the burning coal gas as a light source, vast improvements in quality of life were achieved through the production of cheaper and brighter light sources. But the development of electric lamps, beginning in the second half of the nineteenth century, promoted a revolution in lighting that continues today and influences every aspect of our daily lives. The major types of light sources used for electric lamps are *incandescent light sources*, such as the ubiquitous household "light bulb," and *gas discharge light sources*, of which the most common and best known are fluorescent lamps and high-pressure discharge lamps. Approximately 5 billion incandescent lamp, 2 billion fluorescent lamps, and 200 million high-pressure discharge lamps are in service around the world. They consume about 25% of the worldwide electric energy production. Electric energy production, in turn, accounts for 25% of worldwide energy consumption.

The main focus of this article is to describe the scientific and technological bases of the major electric light sources for general lighting, with particular emphasis on the market in the United States. Section II is intended as

an overview of the principles of each of the light sources to be discussed, together with a guide to the historical development of these sources. Sections III and IV provide brief introductions to electromagnetic radiation and the properties of light as they relate to electric lamps. Sections V–VIII are concerned with the technical discussions of the major light sources (incandescent, tungsten-halogen, and high-pressure discharge and fluorescent lamps). Each section describes the mechanisms of radiation, the energy balance, and the factors that set the limits for performance and efficiency; the fabrication technology and the principal deviations from ideality; representative examples of commercially available products; and development toward improved future products. Some discharge lamps used for special applications are described in Section IX, and a brief review of the relatively new field of electroluminescence applied to light sources is given in Section X.

II. OVERVIEW AND HISTORICAL PERSPECTIVE

A. Incandescent Light Sources

Incandescent light sources produce radiation by passing an electric current through a filament wire, which radiates in both the visible and infrared (IR) regions of the spectrum. The first demonstration of this effect was in England in 1802, by Humphrey Davey, who passed an electric current through metal strips, but it was not until 1878 that Thomas Edison in the United States and Joseph Swan in England separately developed an incandescent lamp using a filament made from carbon thread. This achievement was possible by enclosing the filament in a bulb maintained at high vacuum to prevent rapid burning of the filament. These first lamps burned for a total of 40 hr. Over the next half century, improvements in efficacy and lamp life were obtained, principally by use of improved filament materials and design, culminating in the development of coiled-coil tungsten filaments, which are used in incandescent lamps to the present day (see Section V). The most significant development in incandescent lamp technology since the 1930s has been the *tungsten-halogen* (*T-H*) *lamp* (see Section VI). The operating principle of this lamp is to include a small amount of halogen (bromine or iodine) into the bulb, which combines with evaporated filament material and prevents it from blackening the bulb wall. This principle was first patented by E. A. Schribner in the United States in 1882, but the first practicable T-H lamps were not developed until 1959, as a result of research by E. G. Zuber and F. A. Mosby, also from the United States. The halogen cycle allows tungsten filaments to operate at much higher temperatures, and therefore with higher light output, than standard incandescent lamps.

B. Gas Discharge Light Sources

1. Classification of Gas Discharges for Lighting

Gas discharge light sources are produced by passing an electric current through a gaseous medium, such that a fraction of the constituents of the gas is broken down into ion-electron pairs, forming a conductive plasma. Electrical power is then converted into heating of the gas, walls, and electrodes of the discharge, or into radiation which escapes from the discharge. Radiation is emitted by atoms and molecules of the gas which have been excited into electronic states above the ground state, and then lose their energy by radiative decay to lower electronic states. The fraction of electrical power converted to visible radiation (either directly or by use of a phosphor) determines the efficacy of the discharge as a light source.

Gas discharges are used for a variety of lighting applications and cover a wide range of plasma parameters. Examples can be found in each of the major classifications of industrial plasmas; "thermal" or low thermal equilibrium (LTE) plasmas (*HID* lamps), low-pressure or non-LTE plasmas (e.g., *fluorescent lamps*) and "nonthermal" plasmas (*barrier discharge lamps*). The principle of the electric arc was discovered by Humphrey Davey around 1802, at the same time as his discovery of incandescence induced by electricity in metals. In 1808, he demonstrated a continuous luminous arc, passing a current in air between two pieces of charcoal. This led to the development of the carbon arc, which had limited success in the nineteenth century for street and stage lighting. The carbon arc was formed by connecting two carbon electrodes to the terminals of a current source, bringing the two carbon electrodes momentarily together and then separating them to form an arc. The bright light formed was due to incandescence of the carbon electrodes, and in this sense the carbon arc is distinguished from the discharge lamps used today. Research into discharges in evacuated glass tubes containing small amounts of gases, such as nitrogen and carbon dioxide, dates back to the middle of the nineteenth century. However, discharges containing gases for lighting purposes were not introduced into the market until around 1900, and truly successful application of gas discharge light sources was not achieved until the 1930s with the development of high-pressure mercury, low-pressure sodium and fluorescent lamps.

High-pressure mercury lamps were the first high-intensity discharge (HID) lamps (see Section VII). First produced in 1906, these lamps are high temperature arc

discharges operating at pressures ~1 atm and above, which produce visible light, mainly in the blue and green parts of the spectrum. Since mercury also radiates in the UV, these lamps are usually coated with phosphor to provide an acceptable color rendition. Addition of sodium metal to the mercury lamp has resulted in *high-pressure sodium lamps* (HPS), which are highly efficient lamps with relatively poor color rendition, suitable for outdoor applications. The addition of metal halide salts to the mercury lamp has resulted in the development of *M-H lamps* with very high efficacy and excellent color rendition. High-intensity discharge lamps are primarily used in outdoor floodlighting, street and highway lighting, and indoor industrial and commercial applications. High-intensity discharge lamps operating at very high pressure (150–300 atm) are used for photo-optics applications.

Fluorescent lamps (see Section VIII) are low-pressure discharges, which contain a small amount of mercury vapor and a few torr or less of rare gas. The mercury atoms emit principally UV radiation, which is converted to visible light by a phosphor coating on the lamp. Mercury has the highest vapor pressure of any metallic radiator at room temperature, at which these lamps operate. Fluorescent lamps are used for indoor commercial, industrial, institutional, and retail applications. They were first introduced into the market in 1938, and today fluorescent lamps light over 50% of the building floor space in North America.

Low-pressure sodium (LPS) lamps (see Section IX) contain a small amount of sodium vapor, which emits visible radiation in the yellow region of the spectrum, and a few torr of neon. Low-pressure sodium lamps are the most efficient light sources currently available, but because of their very poor color rendition (CRI), their use is restricted to outdoor applications. *Rare gas discharges* (see Section IX) emit principally in the VUV (70–170 nm), and conversion by a phosphor to visible light is inefficient. Visible emission can also be used for particular purposes, for example, the red line from neon is used for stop lights on some automobiles. Rare gas discharges have been used in special light sources such as street signs and lamps for photocopiers, as well as plasma displays.

Barrier dielectric discharges (or *silent discharges*) (see Section IX) are produced by applying alternating current (AC) high voltages between electrodes, one or both of which are covered with a thin dielectric barrier, such as glass. These discharges contain thousands of "microdischarges" consisting of very energetic electrons, and are classified as nonthermal plasmas. Barrier discharges have been used since 1850 as a source of ozone production, and they are also used for large-scale UV production. More recently, this technology has been used in large flat panel xenon discharges, with a phosphor, for such applications as backlighting of LCD screens.

2. Electrical Circuit (Ballast)

Gases are normally insulators; therefore, to establish current flow through a gas requires generation of free electrons and ions. This is generally accomplished in discharge lamps by application of a sufficiently high potential across the discharge tube terminals (electrodes) so that a single initial electron may gain enough energy to ionize a gas atom. This creates an ion-electron pair, the electron of which gains enough energy to create a new ion-electron pair, etc., in an avalanche, which eventually results in sufficient free electrons in the gas to carry the required discharge current at the desired potential drop. The necessary starting (ignition) and breakdown potential is usually higher than the ultimate operating voltage. Therefore, the electrical circuit must provide a high open circuit voltage for ignition, but must include means to limit the current to the value in the lamp's electrical specification once the discharge is ignited. If the circuit had no current-limiting impedance, the ignition voltage would continue to be applied, the electron avalanche would continue to grow unabated, (runaway) and lamp (arc) current would rapidly rise to destructive levels. The electric circuit device that performs these functions is called a *ballast*. During steady-state operation, the ballast must maintain the correct electrical characteristics of the lamp, with minimum consumption of electric power and no adverse effect on lamp life.

In principle, the current-limiting impedance provided by the ballast could be resistive; however, the power dissipated in the resistance R due to the electric current I is I^2R, which adds to power consumption of the circuit without contributing any light, thus reducing system efficacy. The majority of discharge lamps for general lighting operate on AC, and the necessary impedance in their ballasts is reactive—either pure inductive or inductive plus capacitive, minimizing power dissipation in the ballast. A *conventional magnetic ballast* consists of a coil of wire wound on a core, and operates at the supply line frequency; it may be either a transformer type or simple reactance type. An *electronic ballast* is a ballast involving high-frequency switching that is controlled by active components (transistors, thyristors, etc.), and with the lamp-ballasting impedance provided by a series capacitive or inductive reactance appropriate for the high switching frequency.

C. Electroluminescent Light Sources

Electroluminescence is the emission of light caused by the interaction of an electric field with a solid. Electroluminescent panels became commercially available in the 1950s and are used for back lighting and graphic displays. Light

is produced from a layer of electrolumniscent phosphor powder or phosphor film, activated by the electric field produced between two parallel plate electrodes, one of which is transparent. The most common lighting application of electroluminescence, however, is in light-emitting diodes (LEDs), which are widely used as indicator lamps and in various displays. The principle that radiation can be emitted at a biased *p-n* junction in semiconductors was first demonstrated by Round in 1907. However, it was not until the discovery in 1962 that a gallium arsenide junction, biased in the forward direction, was an efficient emitter of radiation, that serious advances were made in the development of LEDs as light sources. Electroluminescent *p-n* junctions have identical electrical characteristics to conventional silicon and germanium diodes. In common with other developments in the semiconductor industry, rapid advances are being made in this area and the potential for application of LEDs to general lighting is becoming apparent. A detailed evaluation of applications of electroluminescence for lighting is beyond the scope of this article, but a brief summary of the current state of the art is given in Section X.

III. ELECTROMAGNETIC RADIATION

A. Emission and Absorption of Radiation by Matter

A ray of electromagnetic radiation passing through a material may be characterized in terms of its *spectral radiance*, which is the number of photons of frequency ν crossing unit area per second, per unit solid angle, per hertz. As it traverses the substance, it may be attenuated as a result of absorption, or it may be augmented as a result of spontaneous and stimulated emission. Stimulated emission results in photons being added to the beam, in phase and in the identical direction as the stimulating photon (i.e., within the same solid angle as the incident ray). Spontaneously emitted photons are emitted (in the absence of nuclear orientation effects not usually encountered in light sources) with equal probability in all directions. That is, the fraction of the spontaneously emitted photons emitted within a given solid angle $d\Omega$ is equal to $d\Omega/4\pi$.

1. Radiation Transport Differential Equation

These considerations may be quantified in the radiation transport equation, Eq. (1). For simplicity, we consider the variation in spectral radiance $\Gamma(\nu, s)$ of a ray with frequency ν along a "line of sight" s in the medium, such that all the plasma parameters depend only on position s along the line:

$$\frac{d\Gamma(\nu, s)}{ds} = \varepsilon(\nu, s) - \alpha'(\nu, s)\Gamma(\nu, s) \tag{1}$$

where $\varepsilon(\nu, s)$ is the spectral emission coefficient, the spontaneous transition rate per unit volume, per second, per hertz of bandwidth, at frequency ν, and $\alpha'(\nu, s)$ is the effective spectral absorption coefficient. Note that $\alpha'(\nu, s)$ includes the negative contribution from stimulated emission.

For an ensemble of atomic radiators in state u, with number density $n_u(s)$ at point s along the ray:

$$\alpha'(\nu, s) = \alpha(\nu, s) - \kappa(\nu, s)n_u(s) \tag{2}$$

where $\alpha(\nu, s)$ and $\kappa(\nu, s)$ are the spectral absorption coefficient and the stimulated emission coefficient, respectively. Einstein first demonstrated that absorption, spontaneous emission, and stimulated emission coefficients are related through:

$$\begin{aligned} \kappa(\nu, s) &= \frac{g_u}{g_l}\alpha(\nu, s)\, n_l(s) \\ \alpha(\nu, s) &= \frac{\varepsilon(\nu, s)g_u}{2\nu^3 g_l}\, c^2 n_l(s) \end{aligned} \tag{3}$$

where $n_l(s)$ is the density of particles in the lower state of the transition, l, g_u, g_l are the statistical weights of upper and lower states, respectively, and c represents the velocity of light. Incorporating these relationships in the radiation equation gives the following result:

$$\begin{aligned} \frac{d\Gamma(\nu, s)}{ds} = \alpha(\nu, s)\Bigg\{ &\frac{2\nu^3}{c^2}\frac{n_u(s)g_l}{n_l(s)g_u} \\ &+ \left(\frac{n_u(s)g_l}{n_l(s)g_u} - 1\right)\Gamma(\nu, s)\Bigg\} \end{aligned} \tag{4}$$

A light source comprises many such atomic transitions, and the RHS of Eq. (4) must contain the sum of all relevant transitions. The source of the energy input to the material that maintains a population of upper states in the face of depletion by radiation does not concern us in this section.

2. Local Thermal Equilibrium

For a system locally in thermal equilibrium, at local temperature $T(s)$, the velocity distribution function for particles of every species in the plasma, including the electrons, must follows a Maxwell-Boltzmann distribution

$$f(v_j) = \frac{4v_j^2}{\sqrt{\pi}}\left(\frac{m_j}{2kT(s)}\right)^{3/2} \exp\left(-\frac{m_j v_j^2}{kT(s)}\right) \tag{5}$$

where v_j is the velocity of particles j of mass m_j and k is Boltmann's constant. The number densities of upper and lower states are related through a Boltzmann factor:

$$n_u(s) = n_l(s)\frac{g_u}{g_l}\exp(-h\nu/kT(s)) \tag{6}$$

For systems not in local thermal equilibrium (LTE) we may still define a radiation temperature from Eq. (6) for the transition in question, with the stipulation that the radiation temperature will not necessarily be the same for all possible transition pairs.

3. Blackbody Radiation

In an infinite material medium in which the temperature is independent of position, global equilibrium exists, and $d\Gamma(\nu, s)/ds \approx 0$. Solving Eq. (4) for $\Gamma(\nu)$ yields the blackbody, or ideal radiator law, expressed in the units of photons per second, per unit area, per steradian, per hertz of bandpass

$$\Gamma_{BB}(\nu) = \frac{2\nu^3}{c^2}\left(\exp\left(\frac{h\nu}{kT}\right) - 1\right)^{-1} \tag{7}$$

Therefore the radiation transport equation can be rewritten for systems in LTE. in terms of the local temperature and the local blackbody spectral radiance:

$$\frac{d\Gamma(\nu, s)}{ds} = \alpha(\nu, s)\left(1 - \exp\left(\frac{h\nu}{kT}\right)\right)[\Gamma_{BB}(\nu, s) - \Gamma(\nu, s)] \tag{8}$$

Equations (1) and (8) imply Kirchoff's law for the relation between absorption and emission coefficients

$$\epsilon(\nu, s) = \Gamma_{BB}(\nu)\alpha'(\nu, s) \tag{9}$$

For most light sources, $h\nu/kT \geq 3$ and the term $\exp(-h\nu/kT) \ll 1$, so it is usually neglected (i.e., stimulated emission is negligible and $\alpha'(\nu, s) \approx \alpha(\nu, s)$).

Note that in Eq. (8) the temperature may vary with position and therefore the corresponding spectral radiance as well. Note also that the units of spectral radiance are those chosen for the blackbody spectral radiance. Although Eq. (8) is derived in terms of photons per second, it is much more general and can be used with whatever units for blackbody radiance are convenient and familiar. Note also that the growth and decay of a ray of radiation in transiting a material medium is proportional to the absorption coefficient. Where it is less than the blackbody level, spectral radiance will increase only at those frequencies for which the absorption coefficient is greater than zero. Material substances cannot emit electromagnetic radiation at any frequency at which they do not also absorb. Therefore, determination of the radiant emission spectrum of a material substance requires a determination of its absorption spectrum.

B. Sources of Absorption

The processes that contribute to absorption and emission of radiation in the UV, visible, and IR regions of the electromagnetic spectrum are electronic transitions. Many electric discharge lamps take advantage of emission and absorption by electronic transitions in isolated atoms. Absorption and emission "lines" in the radiation spectrum from such transitions are usually broadened by a variety of perturbing influences, such as: (1) collisional or pressure broadening due to collisions between radiating atoms and perturbing atoms; (2) Stark broadening from the electric fields of nearby charged particles, electrons, and ions, in the plasma of a discharge lamp; (3) Doppler broadening as a result of thermal motions of the radiating atoms; and (4) resonance broadening as a result of perturbation by identical atoms in the lower state of the transition pair. In addition to atomic line emission, emission as a result of electronic transitions in molecules, as well as vibrational and rotational transitions, contribute to absorption and emission in some media. Free-free and free-bound transitions of free electrons, which result in absorption and emission of radiation over wide ranges of frequency (continuum emission) can also be important; this process is the most characteristic feature of emission from heated metals.

Equation (8) is quite general for all radiative transitions in a system in LTE. It is beyond the scope of this article to deal with the *a priori* calculations of the frequency-dependent absorption coefficient required for Eq. (8). However, it is clear from the foregoing that the totality of absorption processes must be analyzed with full knowledge of transition probabilities and broadening parameters applicable to the system in question. In many cases of interest in the science and technology of electric discharge light sources, necessary basic data are unknown; they must be either estimated, or used as variable parameters in an empirical fitting to observational data. Research and development of light sources therefore remains an empirical science in many cases, but with continual improvement in numerical models to guide the empirical program.

C. Spectral Radiance Versus Position and Wavelength

For a material medium in LTE, if the value of absorption coefficient is known as a function of frequency and position, and temperature is known as a function of position, spectral radiance may be calculated as a function of position by integration of Eq. (8). For systems not in LTE, integration of Eq. (4) is required.

The optical depth of the medium is defined as

$$\tau(\nu) = \int_0^D \alpha'(\nu, s')\, ds' \tag{10}$$

where D is the distance traversed by the ray. In two limiting cases, optically thick, $\tau(\nu) \gg 1$ and optically thin, $\tau(\nu) \ll 1$, particularly simple results can be obtained.

1. Constant Temperature

For the optically thin case, $\Gamma(\nu, s) \ll \Gamma_{BB}(\nu, s)$ everywhere and it may be neglected in comparison. Radiance increases monotonically with distance through the medium:

$$\frac{d\Gamma(\nu, s)}{ds} = \alpha(\nu, s)\Gamma_{BB}(\nu, s) = \varepsilon(\nu, s)\left(\frac{h\nu}{kT} \geq 3\right) \quad (11)$$

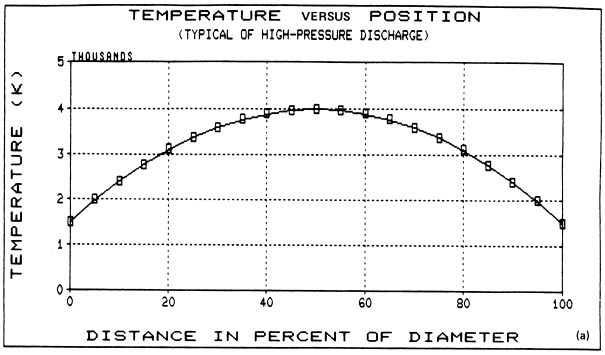
For the case of constant temperature where radiation is not optically thin, integration of Eq. (8) gives:

$$\Gamma(\nu, s) = \Gamma_{BB}(\nu, s)(1 - \exp(-\alpha(\nu)D)) \quad (12)$$

For $\tau \geq 3$, radiance approaches the blackbody. Wherever absorption coefficients are sufficiently large that the substance is optically thick, it is not necessary to know the absorption coefficient exactly; radiance is everywhere equal to the local blackbody value. Because of the large absorption coefficient for free-free transitions of electrons in metals, due to the large free electron density, and because the temperature inside a heated wire is essentially independent of position, this case corresponds reasonably well to the situation inside the metal filament of an incandescent lamp. The internal radiance essentially equals the blackbody level from the UV to the IR.

2. Non-uniform Temperature: Self-Reversed Lines

A common situation in electric discharge light sources is optically thick emission of atomic line radiation from a discharge in LTE, with a position-dependent temperature, as shown in Fig. 1. Although the absorption coefficient at the line's center may be high, leading to a large optical depth, the absorption coefficient decreases with increasing frequency shift from line center because of the broadening processes. Depending on the wavelength shift from line center, optical depth may range from $\tau \gg 1$ at line center to $\tau \ll 1$ in the line wings. Integration of Eq. (8) for the cases $\tau \ll 1$, $\tau \sim 1$, and $\tau \gg 1$ leads to the result shown in Fig. 1. At the line center, where absorption coefficient is high, spectral radiance is essentially equal to the local blackbody value everywhere, rising from a low level at $s = 0$ to a maximum coincident with maximum temperature, and decreasing to a very low value at $s = D$. In the far wings of the line, where optical depth is low, there is essentially no absorption of any radiation emitted into the ray, and radiance increases monotonically to a maximum at $s = D$, but only reaches a relatively low level. In the vicinity of $\tau = 1$, however, spectral radiance



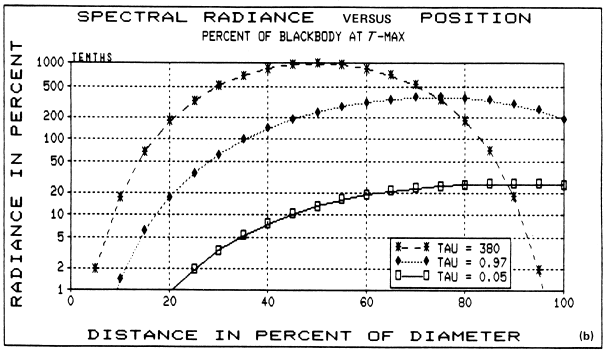


FIGURE 1 (a) Typical profile of temperature along a diameter through an electric discharge in local thermal equilibrium (LTE). (b) Positively directed spectral radiance along a diameter through the LTE discharge of (a) for three different values of the optical depth τ between temperature maximum and discharge wall. Radiance for $\tau = 380$ is indistinguishable from blackbody radiance at the temperature corresponding to each position.

continues to increase for all those values of s for which it is less than local blackbody radiance, and decreases for those values of s for which the converse is true. The distance scale over which these changes can take place, however, is $1/\alpha$. Therefore, the value of emergent radiance at $s = D$ is much greater than that of a blackbody at the temperature found at $s = D$, and typically reaches 10–20% of blackbody spectral radiance at the maximum temperature.

The frequency or wavelength distribution in the emitted line (see Fig. 2) shows a minimum (reversal) at the center wavelength of the atomic line, with maxima in emission on either side. Such emission lines are commonly referred to as self-reversed. The optical depth between temperature maximum and escape point is typically about unity at wavelength of maximum emission.

D. Reflection at Bounding Surfaces

The integration of the radiation transport equation does not of itself permit calculation of the radiation emitted by a bounded material medium. There is an interface at

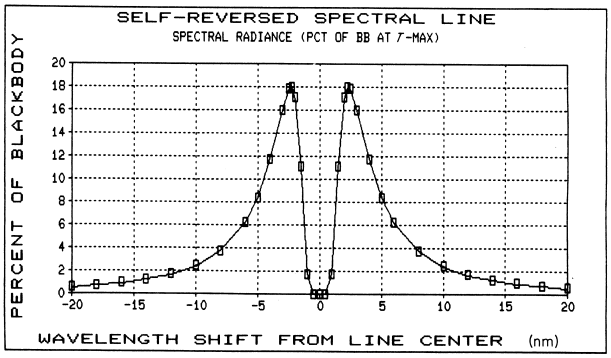


FIGURE 2 Spectral distribution of emitted atomic resonance line radiation for temperature profile of Fig. 1a, illustrating self-reversal, calculated for sodium resonance line (589.0–589.6 nm) for 0.1 atm pressure in a tube 1.0 cm in diameter. Line center (reversal minimum) corresponds to $\tau = 380$ in Fig. 1b; maxima correspond to $\tau = 0.97$; 10-nm shift from line center (either side) corresponds to $\tau = 0.05$.

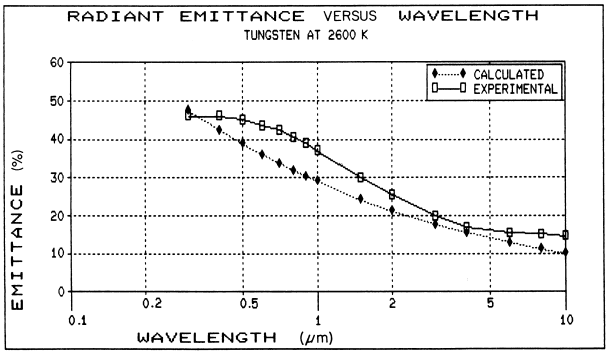


FIGURE 3 Radiant emittance of tungsten as a function of wavelength, experimentally measured and calculated from Eq. (13).

the boundary where the radiating medium meets the external environment, and in general there is a mismatch of indices of refraction. As a result of the difference in the indices of reflection, there is a reflection loss, so that a portion of the internally generated radiation does not escape. For discharge lamps in which the radiating plasma is contained in a transparent vessel, the reflection losses are relatively small, approximately 4–5% at each of several glass surfaces. It is more severe in discharges contained in translucent vessels, which embody many scattering centers within the vessel wall that result in significant diffuse reflection of escaping radiation. For incandescent metal sources, such as tungsten-filament lamps, however, it is extremely significant because of the large index of refraction of the metal, especially in the far IR range.

The escaping flux is equal to $\varepsilon\Gamma_{BB}$, where $\varepsilon = 1 - r$ is called the emittance. The reflectance, r, of a metal surface can be approximately calculated by treating the free electrons of the metal as a free-electron gas, having a plasma frequency greater than the radiation frequency. Provided the collision frequency of electrons in the metal is small in comparison to radiation frequency, a simplified Drude formula gives for ε:

$$\varepsilon = 0.365\left(\frac{\rho}{\lambda}\right)^{1/2} - 0.0464\left(\frac{\rho}{\lambda}\right) \qquad (13)$$

where $\rho\Omega$ cm is the electrical resistivity and λ cm is the wavelength.

Figure 3 shows the observed spectral emittance of tungsten as a function of wavelength, together with the values calculated from Eq. (13). The simplified model is a fairly accurate representation of the IR emittance. It is clear that reflection at the surface drastically reduces IR radiation escape, by as much as 90% at 10 μm, for example. Therefore, even though spectral radiance internal to a heated tungsten body reaches the blackbody level over the entire range of optical frequencies, escaping radiance is at most ~50% of the blackbody value in the visible spectrum, declining toward 10% at 10 μm. This has important consequences for the efficacy of tungsten-filament lamps, as is shown in Section V.

IV. LIGHT

A. Electromagnetic Radiation and the Eye

The proper measure of the "goodness" of artificial light sources is not how well they emit electromagnetic radiation, but how well the emitted radiation can be detected by the eye—that is, light. Therefore, the concern in light source development is not only radiant power, but also the radiant power weighted by the biological action spectra of vision. These action spectra correspond to two distinct perceptions of light; luminance (formerly called brightness) and color. There are two distinct action spectra for luminance: scotopic, or dark-adapted, and photopic, or light-adapted. Under the vast majority of situations involved in application of artificial light sources, the luminance level is high enough that the eye is light-adapted, and the photopic action spectra is the one to use. In photopic vision, there are three sets of photoreceptors, called cones, sensitive to red, green, and blue radiation, respectively. In combination, these are sensitive to the wavelength band of the electromagnetic spectrum between approximately 380 and 780 nm.

Signal processing in the eye-brain system is such that luminance information comes from the sum of the responses of red (long wavelength) cones and green (intermediate wavelength) cones. Color information comes from a combination of the luminance signal and the differences

between the responses of the red, green, and blue (short wavelength) cones.

The relative spectral sensitivity functions of these responses have been determined from an extended series of luminance and color matching experiments by a number of observers, who have been selected to eliminate various pathologies from consideration. The average values of these functions have been agreed on by deliberations of the *Commission Internationale de L'Eclairage* (International Commission on Illumination), known as the CIE.

B. Luminance and Lumens

The spectral luminous efficiency function for photopic vision is designated as $V(\lambda)$, and the total light output of a light source in lumens is determined from the convolution integral over the wavelength interval (or range) of radiated spectral power distribution $P(\lambda)$, weighted by the $V(\lambda)$ function:

$$L = k \int V(\lambda) P(\lambda)\, d\lambda \tag{14}$$

The luminous efficacy of a light source is determined by dividing its total lumen output by the input power required to maintain it in a constant light-emitting state, and is measured in lumens per watt (lm/W). The coefficient k has been arbitrarily set at 683 lm/W, so that the lumen output value determined according to Eq. (14) should agree with comparison photometry based on older so-called standard candles. In this older definition, one standard candle emits 1 lm/sr.

C. Color

The color of light can be expressed in terms of two indices, the chromaticity coordinates x and y, determined from three other functions, X, Y, and Z, called tristimulus values:

$$x = \frac{X}{X+Y+Z} \qquad y = \frac{Y}{X+Y+Z} \tag{15}$$

The functions X, Y, Z are themselves determined from convolution integrals over the spectral power distribution, weighted according to tristimulus functions $\bar{x}(\lambda)$ (red), $\bar{y}(\lambda)$ (green), and $\bar{z}(\lambda)$ (blue). The $\bar{y}(\lambda)$ function is the same as $V(\lambda)$ used for luminance calculations. All three functions are shown in Fig. 4. The x and y values may be measured directly by photometers with suitably calibrated filters, which are in effect analog computers for calculating the convolution integrals.

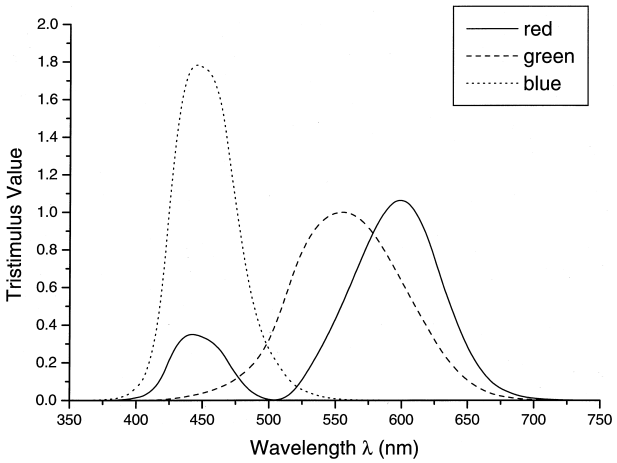


FIGURE 4 Values of the red ($\bar{x}(\lambda)$), green ($\bar{y}(\lambda)$), and blue ($\bar{z}(\lambda)$) tristimulus functions for computing color properties of light sources. Note that although these are ultimately dependent on the spectral responses of the three different types of retinal cones, they are not identically equal to those response functions. The $\bar{y}(\lambda)$ bar (green) function is the same as $V(\lambda)$, the photopic action spectrum for luminance.

Figure 5 shows the locus of monochromatic colors on the $x - y$ plane, forming the 1931 chromaticity diagram ($x - y$ diagram). Color points x, y for nonmonochromatic radiation sources lie in the space enclosed by the locus of spectrum colors. The x, y value for the sum of two sources of color x', y', and x'', y'' are found along a line joining the two x, y coordinates, which is a useful additive property. The curved line extending through the center of the diagram is the locus of colors of blackbody sources of the indicated temperatures. The color points for daylight also fall on or near the blackbody locus, toward the high-temperature end. For reasons probably rooted in evolution, the preferred color domain for most artificial light sources is also on or near the blackbody locus; x, y points above the locus are perceived as too green, those below as gloomy and purple. Therefore a common, although far from exact, representation of the color of a light source is its correlated color temperature (or just color temperature): the temperature of the blackbody source having the closest color match. The $x - y$ diagram, although useful for representing colors of light sources, suffers from the fact that equal perceptible color differences are not represented by equal vector distances between x, y points in different parts of the diagram. Therefore, other systems for color representation are also in use, but these do not concern us here.

D. Color Rendering

Color rendition is another property of light sources. Color rendering is a measure of the degree of change in the

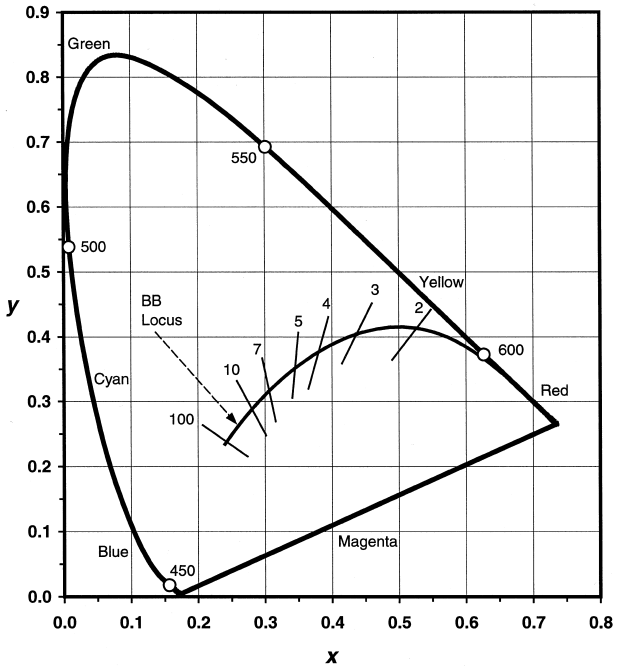


FIGURE 5 CIE chromaticity ($x-y$) diagram: the outer curve is the locus of monochromatic spectrum colors, labeled at selected wavelengths (nm); all real colors are contained within the bounded region. The inner curve is the locus of colors of blackbody radiators ranging from low temperatures at the right to high temperatures at the left: selected temperatures are labeled in 1000 K. Light sources perceived as white have $x-y$ values near the blackbody (BB) locus between 2500 and 5000 K.

apparent colors of colored pigments perceived when they are illuminated by the source in question, in comparison to their apparent colors when they are illuminated by a standard source. For example, a source emitting monochromatic blue light and monochromatic red light, in such proportions that the composite color lies on the blackbody locus at a color temperature of 3000 K, will cause a green pigment to appear black or purple in comparison to the blackbody of the same color temperature—although the colors of the light emitted from the two sources are indistinguishable. The tristimulus values of the light reflected from a pigment can be calculated using formulas identical with those of Eq. (14), except that they also include the spectral reflectance of the pigment as a weighting factor.

The CRI of a light source is computed by calculating the shifts, in a uniform color space, of the color coordinates in a series of standard pigments illuminated by the light source, from the computed color coordinates for the same pigments illuminated by a standard source of the same color temperature. For color temperatures less than 5000 K, the reference source is a blackbody; for color temperatures greater than 5000 K, it is a standard daylight. Continuous spectrum sources, such as incandescent lamps, have high CRIs, near 100, since their spectra are nearly blackbody. Multiline spectrum sources, such as high-pressure discharge lamps and fluorescent lamps, can have CRIs of up to about 90. Color rendering index values less than 40 are generally not desirable for illuminating spaces occupied by humans.

V. INCANDESCENT LAMPS

A. Introduction

The most common form of electric light source is the incandescent lamp, which consists of a solid filament, heated by the passage of electric current, in a hermetically sealed glass envelope, from which all traces of oxidizing gases have been removed. Figure 6 is a schematic diagram illustrating the principal components of a screw-based incandescent lamp of the type manufactured in the United States. Present-day incandescent lamps use a tungsten filament in a coiled or multiply-coiled helix. Tungsten is chosen because of its high melting point (3683 K) and because it is a selective emitter, with the lowest evaporation rate of any known conductive material. Most types use an inert gas filling in the bulb to retard evaporation of the filament, thereby permitting higher filament temperatures for a given design life.

B. Radiation Mechanism

The radiation mechanism in incandescent lamps is the continuous spectrum resulting from deflection and unquanitized radiation of valence electrons (traveling freely through the conduction band of the solid) by collisions with nuclei, each other, and lattice phonons. The optical depth is great enough that the internal

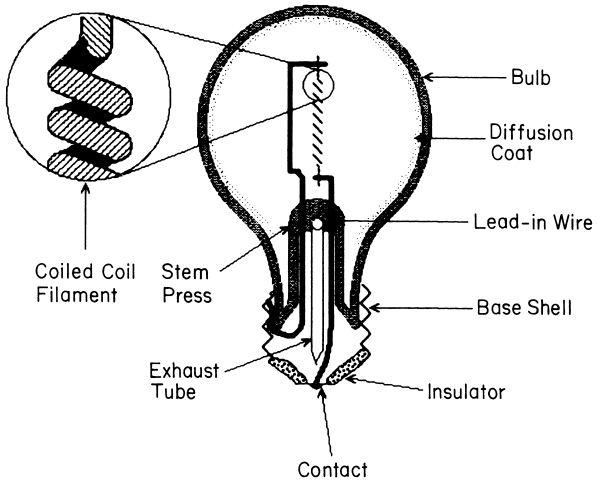


FIGURE 6 Schematic diagram, in section, illustrating components of a typical incandescent lamp.

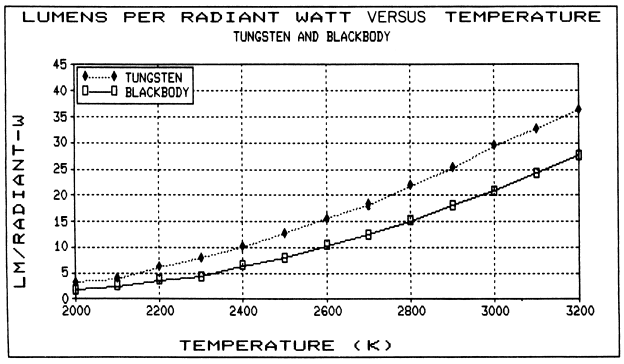


FIGURE 7 Plots of luminous efficacy versus temperature for tungsten and blackbody incandescent radiators. The higher luminous efficacy for tungsten is the result of its higher radiant emittance in the visible in comparison to the IR range (see Fig. 3).

radiation flux reaches the blackbody level throughout the UV-visible-IR spectrum. A large index of refraction causes reflection losses in the escape of this radiation from the surface, yielding an effective emittance of ~0.45 in the visible, decreasing to approximately 0.10 to 0.15 at 10 μm wavelength in the IR. Thus, tungsten emits a larger fraction of total radiation into the visible (giving approximately 30–40% higher luminous efficacy) than a blackbody at the same temperature, with a comparable increase in this visible fraction (and luminous efficacy) with increasing temperature (see Fig. 7). For efficient incandescent lamps, the temperature of the radiator must be as high as possible, yet consistent with constraints on service life. Visible spectral power distribution is blackbody-like, with a color temperature of 50 to 100 K higher than true: the CRI is in the high 90s.

C. Energy Balance

Ohmic heating by the energizing current raises the filament temperature to dissipate power by radiation, convective cooling by the gas fill, and solid conduction to filament supports and electric lead-ins. The power consumption W of the lamp at a given input voltage V is determined by the filament resistance R, a function of the length and cross-sectional area of the wire and resistivity of the wire material, according to $W = V^2/R$ (or I^2R, where I is the lamp current). The resistivity of tungsten increases strongly with temperature, and is 15-fold less at room temperature than at operating temperature; inrush current at initial switch-on is correspondingly higher than operating current. At normal operating temperatures (~2750 K), resistivity $\rho \approx 80\ \mu\Omega$ cm. For a wire of length l and cross-section A, the resistance $R = \rho l/A$. A 100-W lamp operating from a 120-V line has a resistance of 144 Ω, which for total filament lengths of ~0.5 m requires wires of a few hundredths of a millimeter in diameter. Even more slender filaments are required for operation at lower wattages or from 240 V power lines. For a given filament design, electric power input and filament temperature increase with increasing line voltage. The gaseous convection loss from the filament is minimized by coiling the filament into a helix or multiple helix, because of the existence at the hot surface of a stationary boundary layer of gas, with thickness comparable to the diameter of the helices. Thus, the effective surface area cooled by the gaseous convection is that of the cylindrical envelope of the helix, much less than that of the filament wire itself. The effective radiating surface is also smaller than that of the filament wire surface as a result of the coiling, but not by so great a factor. This is because radiation emitted into the interior of the helix does have a reasonable probability of escape between the turns, after one to several reflections, absorptions, and re-emissions. In typical household incandescent lamps, 10–15% of the input power is lost in gaseous convection, 80–85% in radiation, and the balance in conduction to the supports and electric lead-ins. The reduction in evaporation rate resulting from the gas fill allows the filament to operate at ~300 K higher temperatures than in a vacuum, for the same lamp life. This results in as much as 80% increase in luminous efficacy of the emitted radiation, which more than justifies a 10–15% reduction due to gaseous convection loss.

D. Material Limitations Influencing Performance

Maximizing luminous efficacy of incandescent lamps requires as high a radiator temperature as possible. Evaporation of the filament material and the maintenance of mechanical strength at elevated temperatures limit the permissible temperature, so that tungsten is the material of choice.

1. Vaporization

The principal determinant of lamp life is the rate of vaporization of tungsten, which varies as $\exp(-105{,}000/T)$; at 2750 K, this equates to a 4-fold increase per 100 degrees. The end-of-life failure of most incandescent lamps is caused by the formation of "hot spots" on the filament. A thin spot along the length of the filament has higher resistance and higher I^2R heating, because the same total current must pass through every cross-sectional area. It consequently operates at a higher temperature and vaporizes faster than adjacent spots, becoming thinner faster. Failure generally occurs when the thinnest point along the length of the filament is reduced in diameter by ~50%, when less than 1% of the total filament mass has evaporated. High

inrush current at switch-on frequently causes failure at the thinnest spot. However, the same mechanism could cause failure under steady burning conditions within a few hours. The lives of most incandescent lamps are not significantly dependent on the frequency of switching. Thin spots, or hot spots, may originate from nonuniformity in wire diameter or nonuniformity in coil turn spacing (caused by the coiling process it self), distortions in mounting of the filament, or shifts during operation at the elevated temperatures ("squirm" and "sag").

Both luminous efficacy and evaporation rate depend on filament temperature, reflected in an empirical "LPW/DL" law: design life $L_D \propto \eta^{-n}$, where η lm/W is the efficacy and n is typically 7 or 8. The real cost of a lamp to produce a given illumination is the sum of the initial cost plus cost of electricity usage over life. Consequently, the optimum design life is that for which this cost is a minimum. If the design life is too short, although efficacy will be greater and energy cost lower, the cost of bulbs and replacement bulbs will be higher, and if design life is too long, although lamp cost is minimized, energy cost increases. In household use (no labor cost for bulb replacement) the bulb life of ~1000 hr gives the minimum total cost of light. In commercial or industrial use, replacement labor cost may exceed the cost of the bulb itself: thus, longer design lives are needed.

Life and luminous efficacy of an incandescent lamp are strongly affected by operation at service voltages other than the design value. Operation at 10% greater than design voltage will shorten lamp life by a factor of 3, but it will give 35% more light. Operation at 10% less than design voltage will lengthen life 4-fold, at the expense of 28% reduction in light output.

Incorporating an inert gas filling in the bulb obstructs the flow of vapor from the filament, such that net tungsten evaporation is determined by diffusion instead of primary vaporization at the coil. The dependence on filament temperature remains, however, since the density gradient is determined by the equilibrium pressure of tungsten vapor at the filament surface, which has exactly the same temperature dependence as the primary evaporation rate.

The loss rate of filament material is dependent on the diffusion coefficient of tungsten in the gas; life increases approximately linearly with pressure, and increases with atomic mass number of the gas. Household lamps are generally filled with argon, plus a few percent of nitrogen at ~1 atm of operating pressure. The nitrogen prevents the establishment of an electric discharge in the gas between the lead-in wires. Krypton fill is used in certain premium types, for higher efficiency; the lower diffusion coefficient of tungsten in krypton permits slightly higher filament temperature for equal life, and the convective losses in krypton are smaller.

2. Mechanical Strength

Wire for lamp filaments must initially be ductile enough to coil to a mandrel diameter comparable with the wire diameter, without splitting or "green-stick" fracture, but the filaments must become rigid and nonductile in operation at 75% of melting temperature, so that the turns of the coil do not shift, and the wire does not stretch. Control of the microstructure, not only of the wire as it is drawn, but also after recrystallization at elevated temperature, produces this astonishing result.

Nonsag tungsten for lamp use incorporates approximately 30–50 ppm of elemental potassium, introduced in compound form in the powder preparation stage of a powder-metallurgy process. Controlled wire-drawing in multiple stages, with intermediate anneals, results in a line-grained fibrous-looking microstructure of elongated grains parallel to the wire axis. The potassium, which is insoluble in tungsten, is dispersed as uniformly as possible throughout the wire and is concentrated as minute globules along grain boundaries. The fibrous microstructure imparts a flexibility to the wire analogous to rope, so that the wire can be bent easily without fracture.

At temperatures between 1500 and 2000 K (much less than the service temperature) the tungsten recrystallizes, and the small fibrous grains grow to sizes comparable with the wire diameter. Without potassium dopant, the preferred crystal structure is an equi-axed hexagonal (see Fig. 8) with many grain boundaries approximately perpendicular to the wire axis, along which slippage or sliding occurs at elevated temperatures, so that the wire sags or squirms. With potassium-doped wire, however,

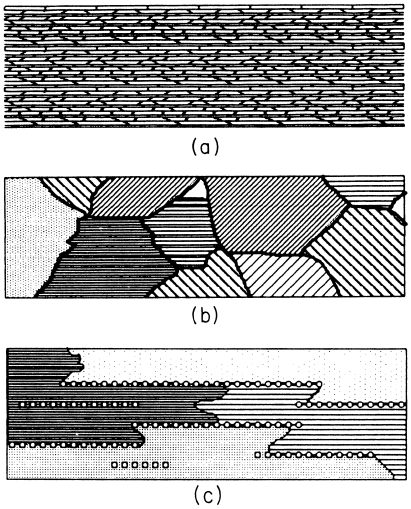


FIGURE 8 Schematic diagram illustrating tungsten metal microstructure: (a) as-drawn, fibrous; (b) recrystallized, undoped wire, equi-axed; (c) recrystallized, potassium-doped wire, interlocking.

the recrystallization is quite different. Since the potassium globules are insoluble and volatile at the recrystallization temperature, they vaporize and create voids distributed in long rows parallel to the wire axis. The rows of voids act as barriers to grain boundary migration, so that propagation of grain boundaries across the wire is greatly inhibited in comparison to propagation along the wire axis. The resulting recrystallized structure, also shown in Fig. 8, then comprises elongated interlocking grains (with greater aspect ratio than 10:1 in high-quality wire) oriented along the wire axis, with few if any perpendicular grain boundaries. The long-grain boundaries exhibit much greater resistance to sliding than short perpendicular boundaries, which greatly increases the hot strength of the wire.

E. Practical Considerations

1. Fabrication Technology

Assembly and processing of incandescent lamps takes place on highly automated machine groups producing 10,000 lamps per hour, essentially without operator intervention. Such manufacturing methods are not only necessary to maintain low costs, but also to cope with the sheer volume of production required to fill the demand of ~1 billion lamps per year in the United States alone.

Lamp manufacture begins with fabrication of tungsten wire coils, typically wound on molybdenum or steel wire mandrels, which are acid-dissolved after stress-relieving. Primary coil winding is done fully automatically, at speeds in excess of 10,000 rpm, with precision of pitch better than 1%. Secondary coil-winding is slower but equally automatic and precise. Glass bulbs are made of soda-lime glass, blown into rotating molds at speeds of 1700 parts per hour.

Stem presses are fabricated by automatically sealing lead-in wires into a press, fusing glass to metal with gas-oxygen and gas-air fires. The lead-in wires themselves are a controlled-thermal expansion composite, comprising a low-expansion nickel-iron core and a high-expansion cladding of copper (Dumet) in such proportions as to provide an expansion match to the glass. For safety reasons, a fuse is attached to the lead-in wires. On failure, an arc may be drawn between the broken filament ends, which transfers to the lead-in wires, causing a high instantaneous current. The fuse is included to break the current and protect the lamp and circuitry from excessive energy dissipation when this occurs. After glass working, the filament supports are automatically bent, and the coiled filaments are automatically fed and mounted. Meanwhile, a light-diffusing particulate coating is applied electrostatically to the inside of the bulb, and both bulb and mount are automatically transferred to the sealing machine, where the glass of the bulb is fused to the glass of the stem with gas-oxygen fires. The sealed bulbs are automatically transferred to an exhaust machine, exhausted and back-filled with gas through an exhaust tube integral with the stem, after which the exhaust tube is fused closed and cut off by gas oxygen fires. The evacuated bulb is transferred to a basing machine, where the external leads are threaded to and affixed to the base, and the filaments are "flashed" to recrystallize the wire in its high temperature microstructure. Automatic testing and packaging follow.

2. Influence of Impurities

Certain impurities substantially degrade the life and lumen maintenance performance of incandescent lamps, even when their filament temperatures are exactly as specified. An important object of the process and material control is to minimize these impurities. The elements hydrogen, oxygen, and carbon, and the transition metals iron, nickel, and chromium have significant deleterious effects.

Carbon can cause formation of tungsten carbides, resulting in brittleness of the filament and susceptibility to fracture under mechanical shock. Since graphite is the preferred lubricant for the wire-drawing process, it is plain that cleaning processes to remove carbon must be under very careful control. Transition metals result in exaggerated grain growth and tend to defeat the purpose of the potassium doping in the recrystallization process. Many of the tools and jigs and fixtures involved in the assembly process are made of steel, and the lead-in wires themselves are typically made of nickel-plated steel. Contamination of filaments and coils from these sources must be monitored and controlled to a minimum.

Oxygen in gaseous form is a particularly damaging impurity, since it reacts with the tungsten filament, thus forming volatile tungsten oxides. As little as 10 parts per billion of oxygen in the gas will remove tungsten from the surface of the filament at 2750 K as fast as it vaporizes. That is, the oxygen will cut the life of the lamp in half. Water vapor is a particularly troublesome form of volatile oxygen, since the reaction to tungsten oxide at the filament surface releases hydrogen. The tungsten oxide evaporates to the cooler bulb wall, at a temperature where the reaction reverses and the evolved hydrogen reduces the tungsten oxide to tungsten, thus reforming water to return to the filament to react again. To minimize the degree of attack on the tungsten filament, present day incandescent lamps contain one or more getters, which are active chemical compounds that react with and remove water, oxygen, hydrogen, or all three, from the gaseous atmosphere of the lamp. Common getters include phosphorus and its compounds and active metals such as zirconium, aluminum, and their alloys. Specific formulations are

TABLE I Representative Incandescent Lamps

Power (W)	Type	Length (in.)	Line (V)	Rated life (hr)	Light output (lm)	Efficacy (lm/W)	CCT (K)
10	10S14/F[a]	$3\frac{1}{2}$	120	1,500	76	7.6	2370
40	40A/W	$4\frac{1}{16}$	120	1,500	445	11.1	2770
100	100A/W	$4\frac{1}{16}$	120	750	1710	17.1	2870
500	PS35/F	$9\frac{3}{8}$	120	1,000	10,000	20	2960
1,500	PS52/CL	$13\frac{1}{16}$	120	1000	33,620	21.1	3070

S, spherical; A, common household lamp shape; PS, pear shaped; F, inside frost; CL, clear; and W, soft white (diffusing-powder-coated-bulbs). CRI $\approx$ 100 for all types.

[a] Vacuum lamp.

proprietary to several manufacturers. Most manufacturers use proprietary, automated, on-line spectroscopic detection of oxygen and water vapor to cull out leakers and contaminated lamps that would be very short-lived in service.

F. Commercially Available Products and Performance Characteristics

Table I lists the performance characteristics of a representative sample of commercially available products, while Fig. 9 is a photograph illustrating a sampling of the variety of incandescent lamp types that are available. The most common bulbs for household applications are the A-line series, with powers of 40–200 W.

Subminiature incandescent lamps, with bulb diameters $<$6 mm, are used for a number of applications where space and weight is critical, such as in aircraft. Bulbs of 3 mm diameter are used in medical apparatus, where the lamps are required to last the life of the equipment, and hence are designed with very low filament temperatures, implying low efficacy.



FIGURE 9 Sample of the variety of types of incandescent lamps.

G. Research and Development

The major deficiency of incandescent lamps is that less than 10% of the emitted radiation is in the visible spectrum, resulting in poorer luminous efficacy than other important light sources. The search for materials with greater selectivity of emittance, or of selective emitter surface coatings compatible with high-temperature tungsten, is a low-level concern in many laboratories.

VI. TUNGSTEN-HALOGEN (T-H) LAMPS

A. Similarities to Incandescent Lamps

Tungsten-halogen lamps, comprising a tungsten-filament radiator in a transparent glass envelope with a gas fill, are substantially the same as incandescent lamps in all aspects of their radiation generation and energy balance. Shared characteristics include dependence of efficacy and life on filament temperature, and many of the limitations on performance resulting from material limitations.

B. Unique Features of T-H Lamps

Tungsten-halogen lamps use much higher gas filling pressures than do their incandescent counterparts, typically 5–10 atm in operation, depending on type. Tungsten-halogen lamp filaments can therefore operate at higher temperatures than incandescent lamps, resulting in about 20% higher efficacy for equal life. To safely contain internal pressures greater than 1 atm requires small-diameter, heavy-walled vessels, fabricated of refractory glasses operating at much higher temperatures than the bulb of an incandescent lamp. To prevent early blackening of the bulb by evaporated tungsten due to its close proximity to the filament, the T-H lamp makes use of a chemical cleaning cycle in which gaseous halogens react with evaporated tungsten, forming volatile halides that do not condense on the hot bulb walls, thereby keeping the walls clean. The tungsten halides decompose on the hot filament, redepositing the tungsten there. Unfortunately, the temperature dependence of the redeposition is not the same as that of evaporation, and tungsten is not redeposited on the hottest spots from which it primarily evaporates. Thus the halogen cycle is not a regenerative one that of itself lengthens the life of the filament.

1. Principal Construction Features

Figure 10 shows the main construction features of the two principal types of T-H lamps: single-ended and double-ended. The most common glass used for such lamps is either pure fused silica (quartz) or Vycor glass containing

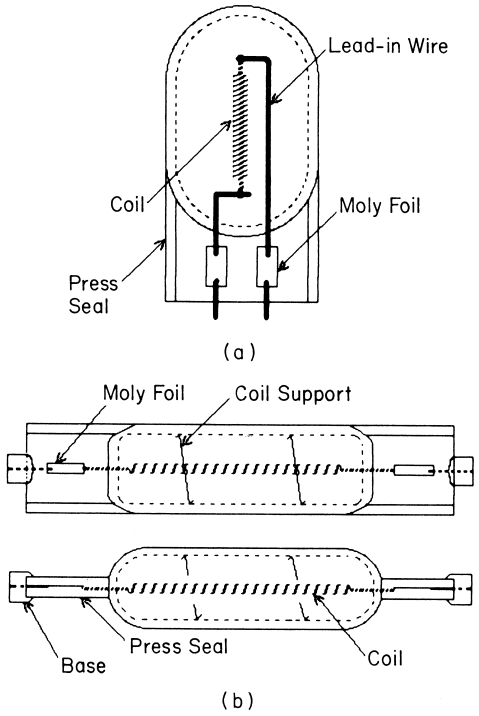


FIGURE 10 Schematic diagram, illustrating components of typical T-H lamps. (a) Single-ended type, with press seal incorporating two molybdenum-foil feedthroughs. (b) Double-ended, shown in top and side views to illustrate geometry of press-seal; the coil support is usually a spiral tantalum or molybdenum wire loop affixed to the coil and loosely contacting the quartz wall, to center and support the elongated coiled filament.

>90% fused silica (Corning Glass Works), both of which have softening temperatures near 1500°C and safe service temperatures of 900–1000°C. Glass-to-metal seals for electric lead-ins for these glasses use a very thin (0.025 mm) molybdenum foil, of thickness-to-width ratio of 1/80 or less, with a very shallow taper at the edge, to less than 25% of the thickness. This configuration permits the molybdenum to yield like plastic to accommodate the factor of 10–12 difference in thermal expansion between quartz and molybdenum, without developing sufficient stress in the surrounding glass to cause it to fail. The lead-in wires (usually molybdenum) and the filament tail assembly are welded to the foil. The structure is then inserted into the end of the quartz tube, which is softened at about 2000°C and then pinched onto the foil (press seal). Such seals are susceptible to oxidation failure when operated at high temperature in air. For each type of T-H lamp, a maximum seal temperature is specified to ensure that seal failure does not occur before normal filament burnout.

For some types of T-H lamps (typically <100 W), in which relatively high-production volumes are needed, alumino-silicate glasses (hard glass) are used. These glasses are also virtually free from alkalis, which would leach out and disrupt the halogen cycle, and they have low UV transmission. The sealing temperature of these glasses is 500–600°C less than quartz or Vycor, and their thermal expansion coefficient is suitable for making a matching seal with molybdenum wire. Use of hard glass allows lower material cost, simpler processing at lower temperatures, and simpler mount assemblies than quartz or Vycor lamps.

2. Chemical Cleaning Cycle

Tungsten-halogen lamps use reversible reactions of the form

$$W + X \leftrightarrow WX(g)$$

$$W + X + O \leftrightarrow WOX(g)$$

in which X is a halogen (usually iodine or bromine). The reactions proceed to the right at low temperatures, to react with evaporated tungsten, forming volatile products that do not condense on the wall of the bulb, which therefore remains clean despite its close proximity to the filament. The maintenance of light output during the life of the lamp is therefore close to 100%. The halides and oxyhalides decompose in contact with the tungsten filament at elevated temperatures, redepositing tungsten on the filament and releasing halogen. Careful control of the cycle is needed to ensure that cleanup is sufficiently rapid to prevent wall-blackening, without being so rapid as to erode the low-temperature tungsten filament legs before normal burnout occurs. This is accomplished by (1) selection of halogen and halogen pressure and (2) control of the oxygen pressure.

C. Practical Problems

1. Fabrication Technology

The fabrication technology of T-H lamps is different from that for incandescent lamps in three respects. Tungsten-halogen lamps require (1) higher glass working temperatures, (2) longer sealing times, and (3) pre-flushing of the vessel with inert gas before sealing, to prevent oxidation of the filament during sealing. Filaments are recrystallized (stabilized) by firing in hydrogen furnaces before assembly to improve accuracy of filament positioning, especially in lamps for use in optical projection systems. Filling the vessel to pressures greater than atmospheric poses another unique problem in T-H lamp fabrication, solved by condensing out the fill gas into liquid form in the bulb with liquid nitrogen coolant before sealing, so that flame sealing of the exhaust tube can be done in the conventional way, with subatmospheric internal pressure. A final major technological difference is the necessity to automatically dispense accurate amounts of corrosive halogens in the lamp on the exhaust machines.

As a result of these differences, the manufacture of many types of T-H lamps is not nearly as automated as it is for incandescent lamps. Only in the manufacture of types produced in quantities of millions per year, such as automotive headlamps, are fully integrated automatic machine groups for mount assembly, sealing, exhausting and filling, and basing used. For other types, of lower production volume, automated and semi-automated, highly flexible machines (capable of rapid type change) are used for the individual processes, and the parts are manually transferred from machine to machine.

2. Influence of Accidental Impurities on Chemistry

The cyclic halogen clean-up reactions in a T-H lamp are critically dependent on oxygen content, since the oxyhalide reaction occurs in parallel with the halide reaction. Oxygen is present as an impurity in the tungsten itself, as well as in lead-in wires and supports. Moreover, variations in the degree of oxidation protection at sealing lead to variations in oxygen content of the lamp. Carbon, again an unavoidable impurity in the tungsten, serves to getter oxygen as carbon monoxide, in which form it is substantially inert. Hydrogen forms water from oxygen, providing water-cycle bulb blackening in parallel with evaporation. Thus, the role of the oxygen may be accelerated cleanup (and filament-leg erosion), neutral, or accelerated blackening, depending on the relative amounts of carbon, hydrogen, and oxygen present. The fact that substantial fractions of all of these contaminants are present due to the normal impurity contents of the lamp materials means that extremely close material control and compensatory adjustment of processes are required to produce a uniform product.

3. Mechanical Strength Problems

All the mechanical strength problems of incandescent lamps are present in T-H lamps, to an exaggerated degree because of the higher operating temperatures of T-H filaments. Most manufacturers establish more stringent specifications for wire in T-H lamps than in incandescent lamps to avoid sag and squirm during operation. The problem of filament strength, as well as leg erosion, hindered development of low-wattage 120 V lamps because of the extremely fine filament wire they require. Improved materials, processes, and process control now permit manufacture of 120 V lamps operating as low as 35 W.

D. Commercially Available Products and Performance Characteristics

The principal applications of T-H lamps have been: retail merchandise lighting; outdoor floodlighting; and projection-type applications, photographic or audio-visual. Single-ended quartz or Vycor T-H lamps have been widely used for automotive headlamps. Hard glass halogen capsules in sealed-beam-type outer envelopes have become globally popular, since they provide equal low-beam performance at two-thirds the electric power, and brighter high beam performance at equal power in comparison to the incandescent product. Use in studio, theater, and television (STTV) applications, although small in absolute numbers, has revolutionized illumination technology. Table II summarizes the characteristics of some T-H lamps, both double-ended and single-ended.

Current use of alumino-silicate (hard glass) T-H capsules in sealed reflector (PAR) lamps has resulted in cost-effective products that compete with traditional

TABLE II Representative T-H Lamps

Power (W)	Type	Length (in.)	Line (V)	Rated life (hr)	Light output (lm)	Efficacy (lm/W)	CCT (K)
Double-ended							
100	100T3Q/CL	$3\frac{1}{8}$	120	2000	1,600	16	3000
500	500T3Q/CL	$4\frac{11}{16}$	130	2000	8,750	17.5	3000
1,500	1500T3Q/CL	$10\frac{1}{16}$	240	2000	33000	22	3000
Single-ended							
75	75Q/CL	$2\frac{3}{8}$	120	2000	1,400	18.7	3000
750	750Q/CL	$3\frac{3}{4}$	120	500	16,500	22	3000
250	FBV[a]	$3\frac{3}{8}$	30	6	7,000	28	
1,000	BTR[a]	$4\frac{3}{8}$	120	200	27,500	27.5	3200
10,000	DTY[a]	15	120	350	290,500	29.1	3200

T, tubular; Q, quartz; and CL, clear.
CRI $\approx$ 100 for all types.
[a] Lamps for studio applications.

TABLE III Comparison of Performance of Pressed-Aluminum-Reflector (PAR) Lamps: Standard Incandescent versus PAR38 with Internal Hard Glass Halogen (HGH) Capsule

Power (W)	Type	Center beam intensity (candela[a])	Beam angle[b] (degrees)	Field angle[c] (degrees)
90	PAR-38 narrow spot	19,500	9	18
150	Incandescent PAR spot	17,000	8	18
45	PAR-38 narrow spot	10000	9	28
75	Incandescent PAR spot	4,800	13	28
45	PAR-38-/FL	1600	28	40
75	Incandescent PAR flood	1,800	30	48
60	PAR-38/IR/FLd[d]	3600	30	50

[a] One candela = 1 lm/sr.
[b] Angular deviation between center of beam and $\frac{1}{2}$-maximum intensity.
[c] Angular deviation between center of beam and $\frac{1}{10}$-maximum intensity.
[d] Infrared-reflecting coating on bulb.

incandescent applications, providing equal illumination and life, with very substantial energy savings. Halogen lamps are available in a wide variety of sizes, wattages, and beam spreads to replace almost any standard incandescent lamp. The halogen lamps offer greater performance and design options than the incandescent lamps they replace. The Energy Policy Act that went into effect in 1995 effectively eliminated incandescent PAR lamps in favor of halogen PAR lamps, due to the increased energy efficiency of the halogen sources. Halogen PAR lamps are now widely used for both retail and residential lighting needs. Table III provides a comparison of performance of hard glass T-H PAR lamps with their incandescent predecessors. Figure 11 is a photograph illustrating a sampling of some of the various types.



FIGURE 11 Sample of T-H PAR reflector lamp types.

Increased efficacies are realized in IR tungsten halogen lamps by deposition of a multilayer interference filter on the outside surface of the single-ended-glass lamp capsule. The multilayer coating is designed to transmit visible light while reflecting infrared radiation back onto the filament, thereby increasing the lamp efficacy. The reflecting filter consists of alternating layers of tanatala and silica and results in an efficacy increase of ∼35%. This technique requires exact mechanical precision of the envelope geometry and of coil location in the center at manufacture, as well as limiting subsequent movement of the coil, both due to mechanical shock in shipping and sag or squirm in operation.

E. Research and Development

As with incandescent lamps, research into new filament materials is continuing. Optimization of performance through selection of gas-fill, pressure, and halogen dose is also a concern and numerical models may be able to assist in this development. In this context, a true "regenerative" cycle which could "heal" filament hot spots would make a major impact on lamp life. Another area where major improvement in efficiency could be achieved is the development of anti-Stokes phosphors, which could convert IR to visible radiation.

VII. HIGH-INTENSITY DISCHARGE (HID) LAMPS

A. Introduction

High-intensity discharge lamps are gas discharges which operate at pressures of order 1 atm or greater. Much higher radiating temperatures can be achieved than in any incandescent solid; appropriate selection of the gaseous medium results in a higher fraction of radiated power emitted in the visible spectrum, with a much smaller fraction of IR radiation. As a result, such sources are very bright, and are 3–10 times more efficient than incandescent lamps. These discharges are classified as "thermal plasmas," since all species in the discharge (atoms, molecules, ions, and electrons) are assumed to have approximately the same high temperatures (1000 K–7000 K). There are three principal types of HID lamps, distinguished by their radiating species: high-pressure mercury; high-pressure sodium (HPS); and metal halide (M-H).

Figures 12a and 12b show schematic diagrams, illustrating the principal components of a typical HPS lamp and M-H lamp, respectively. Each comprises an inner discharge tube (arc tube) containing the high-pressure gas or vapor enclosed in a hermetically sealed outer envelope (jacket). The outer jacket is required for protection of the

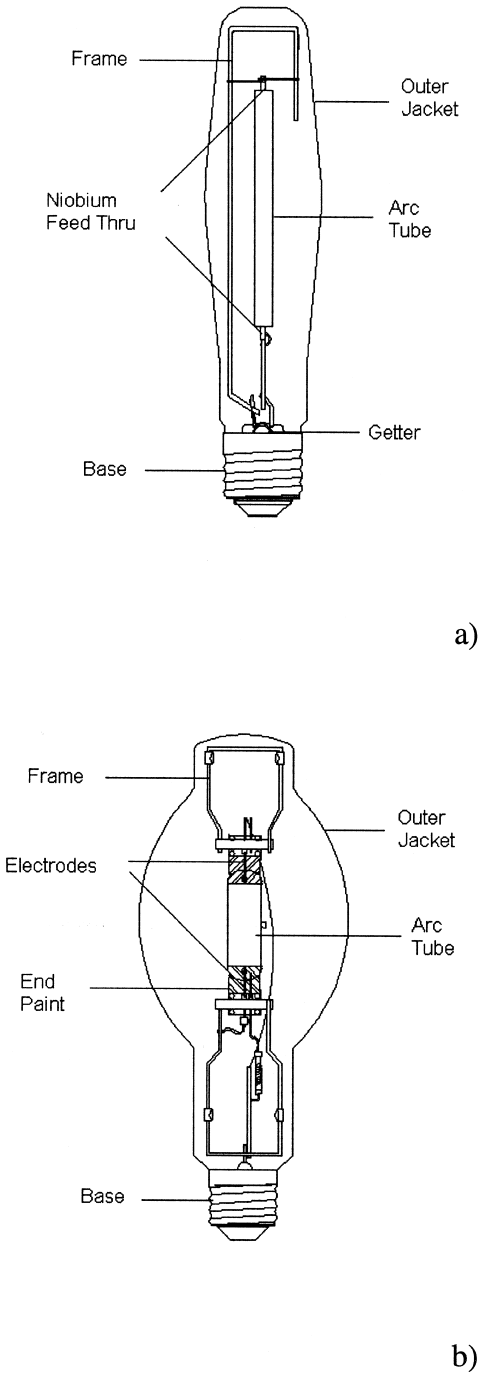


FIGURE 12 Schematic diagram showing principal features of (a) 400-W HPS lamp and (b) 400-W M-H lamp.

arc tube seals from oxidation, and absorption of any short wave length UV rays that may be emitted from the arc tube. Arc tubes for mercury and most M-H lamps are quartz, similar to T-H lamp capsules; arc tubes for HPS are fabricated from translucent PCA, a ceramic material able to withstand corrosion by hot molten and gaseous sodium. Ceramic PCA M-H lamps are also being introduced into the market, because of their increased resistance to chemical reactions and ability to operate at higher wall temperatures (~1100°C, compared to ~900°C for quartz). The arc tubes in these lamps are filled at room temperature to about 100 torr of inert gas. The high pressure during operation is produced in all of these lamps by vaporization of mercury, introduced into the lamps as a droplet or an amalgam.

B. Radiation Mechanism

In HID lamps, the fraction of electrical power dissipated in heating the gas produces very high temperatures (4000–7000 K) near the axis. Unlike a metallic conductor, the gas has very low thermal conductivity; there is a large temperature gradient between the center axis and the walls of the arc tube, which remains at modest temperatures (1000–l500 K, depending on lamp type). Rapid collisions sharing energy among electrons, ions, gas atoms, and excited gas atoms promote LTE, and the radiating properties are described by Eq. (8) of Section III, with absorption coefficients determined by the optical transitions of the gaseous medium. Such a discharge is called a high-pressure arc discharge, or arc.

The emitted radiation consists primarily of the line spectra of the elements present in the gas: mercury in mercury lamps and sodium in HPS lamps. Although mercury vapor is also present in the gas in HPS lamps, the temperature is insufficient to excite mercury to states that contribute to emission in the visible wavelengths, and mercury lines are essentially absent from the spectrum.

Metal halide lamps contain atomic metal radiators that either do not vaporize at wall temperatures accessible to quartz or react vigorously with quartz at those temperatures. They are introduced into the arc tube in the form of metal iodides, which are quite volatile at ~1000 K and relatively nonreactive to quartz. The introduction of ceramic technology into M-H lamps has allowed these lamps to operate at 150–200°C higher wall temperatures, and therefore higher vapor pressures of the salts, with a consequent gain in efficacy. The metal iodides participate in a halogen cycle, similar to that in T-H lamps. Iodide molecules evaporate from the tube wall and enter the high-temperature core of the arc, where they dissociate, freeing metal atoms to contribute their optical transitions to the radiation of the gas mixture. Free metal and iodine atoms diffuse, or are convected, back into the low temperature regions of the discharge and chemically recombine to metal iodide, preventing condensation on, or reaction with, the arc tube wall by metal atoms. The emitted radiation is dominated by the line spectrum of the added metals, with contributions from mercury (which is also present) and minor radiation from metal iodide molecules. Arc temperatures

at the core are too low to excite iodine significantly, and its emission lines are barely detectable.

In all three types of HID lamps, the absorption spectra exhibit broadening of the natural line widths by resonance, van der Waals, and Stark broadening (see Section III.B). Calculation of the absorption coefficient as a function of wavelength requires knowledge of the broadening mechanisms, as well as of the density of absorbing atoms, and of the probabilities of their optical transitions. Calculation of the spectral radiance at the arc tube wall requires knowledge of the spatial temperature profile. The temperature profile is itself modified by the radiation transport; radiation emitted at one point may be absorbed at another point, resulting in a net transfer of energy. In all HID lamps, optical depths range from very thin to very thick. Because of the large temperature gradients, self-reversed lines are common. Some success has been achieved in modeling radiation transport in HPS lamps, but modeling of M-H lamps remains a formidable problem. Spectral radiances at emission line maxima approach 50% or more of the blackbody level at the arc axis temperature for strong nonresonance lines, and 25% or more of that level for resonance lines. Infrared transitions in most atoms, between upper energy levels near the ionization limit, are not strongly excited at the arc temperatures. Their contribution to the absorption spectrum is weak, as are corresponding emission lines. By design, electron density is maintained at a sufficiently low level to minimize free-free transitions of electrons colliding with each other and with ions, since these processes lead predominantly to emission in the IR. The consequent weak emission of HID lamps in the IR contributes strongly to their high efficacy. Proper choice of radiating atoms can favor emission in the visible over the UV wavelengths as well. Metal halide lamps permit the choice of radiating atoms that provide for emission of all colors, so that the CRI is also high. Figure 13 shows spectral power distributions of the radiation emitted from the three types of HID lamps. These spectra are taken on instruments with relatively wide spectral bandwidth, and do not reveal the self-reversed nature of the spectral lines except in the case of the sodium resonance lines at 589 nm in the HPS lamp.

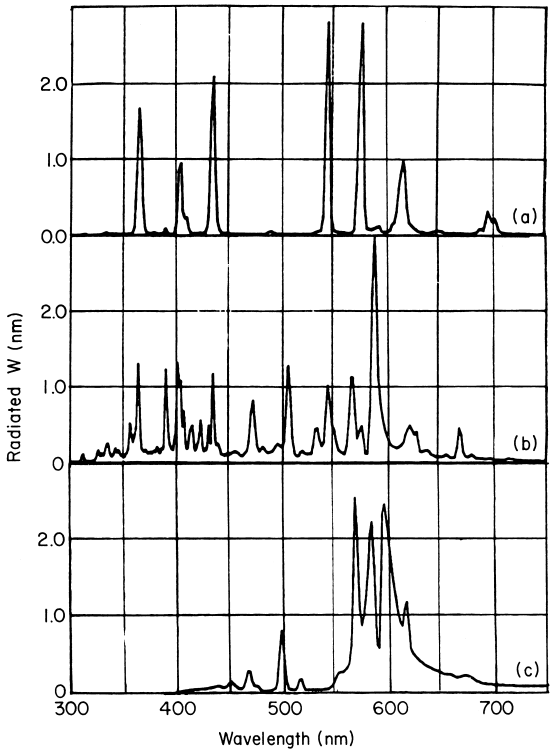


FIGURE 13 Spectral distribution of radiated power (W/nm) of HID lamp types in use in the United States, all at 400-W input power. (A) Mercury lamp with phosphor-coated outer jacket; strong peaks at 365, 405, 436, 546, and 578 nm are atomic lines of mercury, while the peak at 615 nm is red emission from the phosphor, excited by UV from the arc tube: CCT = 4000 K, CRI = 45. (B) Sodium-scandium M-H lamp; the mercury lines are visible in the spectrum. The strong line at 590 nm is due to sodium, and the balance is due to groups of lines emitted by scandium atoms (unresolved at the spectrometer bandpass used here): CCT = 4000 K, CRI = 65. (C) HPS lamp; all radiation is from sodium, the major yellow peak with a minimum at 590 being the self-reversed resonance line of sodium (see Fig. 2): CCT = 2100 K, CRI = 22.

C. Energy Balance

1. Arc Discharge

Electric power is dissipated in the arc tube in three modes: heating electrodes to electron emitting temperatures, heat conduction through the radial temperature gradient in the gas to the walls, and radiation from the high temperature gaseous medium. Table IV shows the approximate percentages of power dissipation in these modes for 400-W mercury, HPS, and M-H lamps. Electrode loss is

TABLE IV Approximate Percentages of Power Dissipation in High-Intensity Discharge Lamps

Lamp type	Electrode loss	Heat conduction plus wall absorption	Radiation		
			IR	UV	Visible
Mercury[a]	12	38	10	24	16
Metal halide	12	25	17	19	21
High pressure sodium[b]	4	37	32	0	37

[a] Low wall loading in mercury lamps results in a large conduction loss, which, together with unfavorable fraction of radiation in UV results in low efficiency of visible radiation in comparison to other types.

[b] High thermal conductance of sodium vapor results in large heat conduction loss; strong IR emission lines of sodium also reduce fraction of visible radiation. Visible radiation is predominantly near the maximum of the photopic response curve; therefore, luminous efficacy is high.

minimized by designing the lamp to operate at as low a current and as high a voltage drop, for a given power, as is practical. Voltage drop between electrodes increases with increasing interelectrode spacing (arc length) and with increasing gas pressure. Heat conduction loss is approximately a constant loss per unit length of arc. The percentage lost in heat conduction is minimized by operation at as high a power input per unit length as possible, i.e., operation with as short an arc length for a given power input as possible. Maximizing voltage drop while minimizing arc length requires high gas pressures, ~1–10 atm, depending on type. In most commercially available HPS and M-H lamps, mercury is used as a buffer gas because, for the same gas temperature, gaseous mercury has a lower electrical conductivity (and therefore higher resistance) than any credible alternative, meaning that higher voltages can be achieved for a given current at modest pressures. Further, the thermal conductance of mercury vapor is very low, minimizing heat conduction loss.

Convection plays a role in determining lamp performance in M-H lamps. Convective flows due to thermal gradients can result in nonuniform arc tube wall temperatures. For ideal performance, the body of the inner arc tube wall should be maintained at a near-uniform temperature, close to that required for optimum vapor pressure of the constituent gases. Large temperature differences along the wall may result in the cold spot temperature being too low to provide the required vapor pressure, or the maximum wall temperature being too high, resulting in deterioration of the arc tube material. If required, wall temperatures may be increased by a coating of an absorbing end paint at one or both ends of the arc tube. Thermal management of the arc tube can also be modified by surrounding it with a a glass or quartz shroud, which also protects the outer jacket in the event of destructive failure, and provides a positive electric potential to prevent sodium loss (see Section VII.D). Convective flows can be reduced by reducing the tube radius and choice of the "aspect ratio" (the ratio between tube length and diameter) for optimum thermal performance is an important aspect of lamp design. The development of numerical models to describe thermal behavior of HID lamps is providing an increasingly useful lamp design tool.

Metal halide lamp performance is significantly affected by orientation during operation. The arc discharge received its name from the upward bowing of the discharge during horizontal operation. Bowing is due to the buoyancy forces resulting from the balance of convective flows and gravitational forces and may produce higher temperatures at the top of the arc tube. In some products, this is remedied by curving the arc tube to follow the bowing. Bowing also causes difficulties in design of optical systems for some lamps. Straightening of the arc may be achieved by application of a correcting magnetic field or operation at high frequencies. The combination of diffusion and natural convection in vertically burning M-H lamps leads to chemical separation with a concentration of lighter metallic species near the lower electrode.

2. Lamp-Electrical-Circuit System

The majority of HID lamps operate on alternating current and the necessary impedance in their ballasts is reactive: either pure inductive or inductive plus capacitive, minimizing power dissipation in the ballast. The ignition voltage requirements of many HID lamps exceed the magnitude of the AC input voltage connected to the ballast primary. Ballasts for such HID lamps employ either a voltage step-up transformer or a supplemental igniter circuit to increase the ballast open circuit voltage in order to properly start the lamp.

In mercury lamps and older M-H systems, starting is achieved using an auxiliary electrode (or probe) placed a few millimeters from one of the main electrodes. When the supply is switched on, a glow discharge forms between the main electrode and the probe. Current is controlled by a high temperature resistor in series with the probe and situated between the outer wall of the arc tube and the outer jacket. The low current glow discharge provides sufficient ionization to initiate breakdown between the two main electrodes of the arc tube. After breakdown, the probe plays no role during run-up or steady-state lamp operation and must be effectively insulated to avoid electrolysis. Another aid to starting is UV irradiation of the discharge from a small capacitively coupled glow discharge lamp, located in the outer jacket. High-pressure sodium lamps and some M-H lamps require supplemental "igniter circuits" in addition to starting probes and/or other types of UV starting aids to ensure proper lamp ignition. Igniter circuits interface with a portion of the ballast's output circuit, such that pulse voltages, with magnitude of several kilovolts, are applied across the parallel connection of the arc tube and its probes/starting aids until breakdown is achieved, after which the igniter circuit plays no further role. Recent trends in M-H lamp technology have led to the replacement of the probe by 2–4.5-kV pulsed start circuits, with improved lumen maintenance, for most wattages. High-intensity discharge lamps must also be re-ignited each half cycle; the peak instantaneous voltage required to accomplish this is typically 40% more than the instantaneous arc voltage during the majority of the half cycle. A properly designed ballast will accomplish lamp re-ignition at the start of each half cycle, without contribution to the voltage from the ballast's igniter circuitry.

The ignition voltage required is a strongly increasing function of gas pressure. Therefore, all HID lamps use a low-pressure rare gas during starting (typically 20–200 torr) and provide the required high operating pressure

by mercury vapor. Discharges in argon are easier to start than xenon, and hence this gas is used in mercury and M-H lamps. Xenon is preferred in HPS lamps, since it provides higher efficacy during normal operation, due to its lower thermal conductivity. At ignition, the mercury is cold, with a vapor pressure of a few millitorr and plays no role in lamp ignition. In operation, the arc tube wall is hot, and sufficient mercury is in the gas phase to raise the pressure to the level necessary to provide the design value of operating voltage. Full light output in HID lamps is only reached a few minutes after starting. If the lamp is switched off and then re-started while warm, mercury lamps require 10–20 min to return to full light output, Metal halide lamps are somewhat faster (5–12 min.) while HPS lamps can achieve this in about 1 min. Typical HID magnetic ballasts have a power transfer efficiency of 85–90%. Characteristics which limit ballast efficiency include wire/winding losses, core-magnetizing current losses, core eddy current losses (gauss-related), and fringing field/core gap losses. Understanding the interaction between lamp electrical requirements and ballast efficiency requires a thorough understanding of the lamp starting, glow-to-arc transition, and steady-state lamp-operating modes.

High-frequency electronic ballasts are being introduced to reduce weight, volume, and audio noise and improve maintenance, although these ballasts are substantially more expensive than magnetic ballasts. High-intensity discharge lamps operating above 1 kHz can have unstable arc/operating modes due to the excitation of acoustic resonances within the arc tube. Problematic acoustic resonances occur when high-frequency components of the lamp power couple with the natural sound wave resonances of the arc tube cavity to produce standing pressure waves, which can rapidly destabilize the arc. At sufficiently high frequency ($>$350 kHz) these instabilities are damped and stable operation of the lamp is possible. Another approach is to switch the frequency before sufficient energy is imparted to the acoustic resonance (frequency hopping). The difficulty with this method is that the frequencies at which acoustic resonances occur are sensitive to the operating conditions of the lamps, and it is difficult in practice to find a sufficiently broad band that will operate all lamps intended for the ballast. A common solution is to apply a low frequency ($<$400 Hz) square wave voltage and current to the arc tube, such that the total power dissipated in the lamp is approximately constant with time, minimizing the potential to excite destabilizing acoustic resonances.

D. Material Limitations Influencing Performance

Unlike incandescent lamps, there is no single mechanism in HID lamps which determines lamp life, nor is life inversely proportional to efficacy. Moreover, lives of HID lamps are prodigiously long; most mercury and HPS types have rated lives of 24,000 hr, while M-H lamps have rated lives from 6,000 to 20,000 hr. Since typical operating hours in dusk-to-dawn outdoor service, or two-shift indoor commercial service, are 4000 hr per year, these rates represent operating lifetimes of up to 6 years. Nevertheless, design choices that lead to higher efficacy are constrained by material limitations that result in shorter life. Mercury and HPS lamp electrodes are coated with an electron-emitting activator material, which evaporates in time, resulting in failure to ignite or re-ignite each half cycle. All HID lamp types are subject to excessive arc tube blackening by material evaporated or sputtered from the electrodes. In HPS lamps, wall blackening near the electrodes can lead to an increase in the amalgam temperature and the subsequent increase in gas pressure causes a voltage rise. This can result in "cycling" near the end of lamp life, in which the lamp voltage rises to a value which can no longer be sustained by the ballast and the lamp extinguishes. The lamp then cools to a temperature at which the igniter pulse is sufficient to restart the lamp and the process is repeated, resulting in a continuous switching cycle that repeats every few minutes. This problem can be solved by reducing the amalgam dose to the extent that all the mercury and sodium is in the vapor phase during normal lamp operation. This principle has led to the development of so-called "unsaturated vapor" HPS lamps. For these lamps, minimization of sodium loss due to chemical reactions with arc tube components is critical, since there is no pool of sodium available to replace it.

Sodium loss can occur as a result of an electrolytic process at the arc tube wall in M-H lamps. In operation, there are always a few parts per million of sodium ions dissolved in quartz in contact with sodium iodide, as a result of reaching thermochemical equilibrium in the reversible reaction between the quartz and the iodide. The quantity is not harmful to the quartz, nor does it represent a significant depletion of sodium from the initial dose dispensed. However, sodium ions are mobile in quartz, and negative charging of the outer surface of the quartz arc tube by photoelectrons emitted from various parts of the outer jacket will attract the sodium ions to the outer surface to be neutralized and evaporate. The depletion of the ionic concentration in the inner surface then permits the forward reaction to proceed to provide more sodium ions, which are electrolyzed in turn, until eventually a very substantial fraction of the original sodium dose has been lost. Outer jacket designs providing minimum photoelectric-emitting surfaces, or positively biased surfaces (such as a shroud), are used to mitigate this problem.

Chemical reactions between quartz and metals and metal iodides may form stable metal oxides at the wall and liberate silicon metal. The loss of metal by this process reduces the partial pressure of metal vapor and ultimately

changes the radiation output of the lamp. The silicon metal reacts with iodine to form volatile silicon tetraiodide, which decomposes at the electrode temperature, and deposits molten silicon on the electrode, drastically distorting its shape and degrading its performance. This becomes one of the life-limiting processes in M-H lamps, by adversely affecting the re-ignition process every half cycle, to the point that the ballast can no longer re-ignite the lamp. Efficacy of HID lamps increases as input power per unit length is increased, first because of a reduction in the fraction of power lost to heat conduction, and secondly due to increased arc tube temperature, which results in higher vapor pressures of the radiating species. However, increased arc tube temperatures result in reduced lamp life due to increased chemical reactions with the arc tube material, and in the case of quartz, a change in the crystal structure (de-vitrification). These constraints are generally embodied in the form of rule-of-thumb design rules regarding acceptable wall-loading (arc power input per unit area of inner arc tube wall surface). Acceptable lives in mercury lamps are obtained at wall-loadings of 10–12 W/cm^2 and quartz M-H lamps for general lighting are commonly found at 13–22 W/cm^2. The robustness of PCA at higher temperatures allows ceramic M-H lamps to run at ~40 W/cm^2, and HPS can use designs at 15–20 W/cm^2. For some applications, such as automobile headlamps and projection lamps, shorter lifetimes are acceptable and higher wall loading may be used.

Molybdenum foil seals in both mercury and M-H lamps have adequate life at much higher temperatures than in T-H lamps because they are protected from oxidation by a vacuum or an inert atmosphere in the outer jacket. The electrical lead-in in HPS lamps involves a niobium (also known as columbium) metal member, chosen for expansion match in PCA and sealed with a polycrystalline oxide mixture, fusible without melting either niobium or PCA. The niobium is protected from oxidation by a vacuum in the outer jacket. In HPS lamps, the maximum seal temperature is determined by the resistance of sealing compounds to attack by sodium, thereby limiting the cold spot temperature and hence the maximum pressure of sodium vapor in the arc tube.

E. Practical Considerations

1. Fabrication Technology

Fabrication of mercury and quartz M-H lamps involves aspects similar to a combination of T-H and incandescent lamp fabrication, since two hermetically sealed vessels are required: one of fused silica (quartz), incorporating molybdenum foil seals, and an outer jacket of glass using a stem seal similar to an incandescent lamp.

Mercury arc tubes require accurate amounts of mercury and argon, which determine the operating and starting voltages of the lamps, respectively. The outer jacket is filled with nitrogen gas to prevent oxidation and arcing, and to maintain the arc tube temperature at its required value. The inside of the outer jacket is coated with a phosphor, which converts UV radiation from the discharge to (red) visible radiation at 620 nm. Wide differences exist among manufacturers and types in the level of automation of manufacture of mercury lamps. Some popular lamp types are manufactured on fully integrated production lines that automatically seal and process the arc tube, mount it on the outer jacket stem press, and seal and process the outer jacket at speeds of 1000 per hr. For others, one or more stages of the process involve jigs-and-fixtures manual operations. Metal halide lamp fabrication is much more complex than for mercury lamps, due to the introduction of hydroscopic iodide salts. The salts themselves are dehydrated in advance to only a few parts per million of water content, and they must be handled in atmosphere glove boxes free of moisture and oxygen to a few parts per million. The precision of the cold spot temperature is even more critical in M-H lamps, and hence manufacturing tolerances are much tighter. Despite these difficulties, fabrication processes are now highly automated. This has been achieved by introducing more robust designs and reducing the number of lamp parts and the number of operations performed. Arc tube processing takes place on automatic exhaust machines. Final assembly and outer jacket processing usually proceeds through the mercury lamp line, sharing equipment and processes. Some lower wattage M-H lamps have vacuum outer jackets. A getter is also included to remove impurities from the nitrogen gas or maintain the vacuum.

High-pressure sodium arc tubes are composed of ceramic material and require a different sealing process than quartz. Polycrystalline alumina tubes are manufactured by either extruding or pressing very pure alumina powder of precisely specified surface area, to which controlled quantities of sintering aid compounds have been added. The tubes are then sintered at high temperature so that the resultant translucent tubing has a total diffuse transmittance of 95% or more. Arc tube sealing, exhausting, and filling may be carried out by two processes that differ in the stage at which the sodium amalgam is introduced into the arc tube. In one process, both niobium feedthroughs are cemented at once with a fusible oxide mixture (frit) into the tube ends in a vacuum or atmosphere furnace. One of these feedthroughs is equipped with an exhaust tube. The arc tube is evacuated and mercury-sodium amalgam is dispensed through the exhaust tube which is then pinched shut in a cold weld. In the other process, each end of the arc tube is sealed separately. The second frit seal is

made in an atmosphere furnace containing the desired fill pressure of starting gas, with mercury-sodium amalgam already dispensed into the tube and residing at the cool end which was sealed first. There is no exhaust tube, and no separate exhaust and filling process. The final assembly into an outer jacket and outer jacket processing are similar to those for mercury lamps, and they may share equipment. The sealing of PCA M-H arc tubes is similar to the second process described above. The M-H fill is introduced into the tube after the first seal is made and a second frit seal is then completed. The feedthroughs for the electrodes are introduced into the arc tube through a long capillary extension, with the frit seal at the far end. This allows the frit and niobium to remain relatively cold during lamp operation, so that reaction with the halide is minimized.

2. Influence of Impurities

In all three classes of HID lamps, gaseous impurities that interfere with ignition or re-ignition must be minimized. Hydrogen is a particularly troublesome species, since it is a principal component of water vapor, a common impurity, and difficult to remove. In addition to surface contamination by water, dissolved hydroxyl in quartz is a major problem. This is dissociated photolytically by UV rays from the discharge, and the liberated hydrogen diffuses rapidly into the arc tube interior. Manufacturing processes for quartz tubes are tailored to reduce the hydroxyl content to a few parts per million; optical absorption in an IR absorption band of OH is used as a quality control tool. Because of rapid diffusion of hydrogen through hot quartz, moisture in the outer jacket must also be minimized since it may also be dissociated by UV light, and the liberated hydrogen may diffuse through the arc tube wall into the interior. Hydrogen is especially harmful in M-H lamps because of reaction with iodide salts, forming HI and liberating metal. The metal reacts with quartz and is lost to the halogen cycle; the hydrogen iodide is gaseous and interferes with electron avalanche growth in ignition by capturing free electrons. Oxygen is also deleterious, especially in M-H and HPS lamps, because it promotes reaction of metal atoms with arc tube wall material. Hydrocarbon vapors in the outer jacket are also decomposed thermally and photolytically, liberating hydrogen as well as depositing amorphous carbon on the arc tube wall surface, reducing optical transmission.

F. Commercially Available Products and Performance Characteristics

High-intensity discharge lamps are primarily used in outdoor floodlighting, street and highway lighting, and indoor industrial, commercial, and retail applications, in which their exceptionally long service lives and high efficacies outweigh their high initial cost and the added expense of the required ballasts. In such service, relatively high mounting heights have been traditional so that large areas could be illuminated from relatively few fixtures. This has dictated in the past that the more popular types were from 8000 to 100,000 lm in the United States and from 5,000 to 50,000 lm in Europe. However, the increasing application of HPS and M-H lamps in lower ceiling height indoor commercial service, requiring smaller lumen packages ~4,000–10,000 lm—has resulted in many new lower wattage types being introduced within the last few years. Metal halide lamps are the fastest growing segment in HID lighting.

There has been a major shift in outdoor lighting to what is being termed "white light," for which the M-H lamp technology is the product of choice. The white light trend is most noticeable in downtown urban areas and parking lot illumination. The higher system and operating cost of M-H versus HPS is not influencing this shift to white light. The fastest growing segment of indoor lighting is retail. New M-H chemistries and arc tube materials are allowing color performance similar to T-H. The improved M-H color performance and longer life make the M-H lamp an economical solution. Most new retail constructions are incorporating higher ceiling height in their design, and the high output M-H lamp is a perfect choice to deliver increased lighting levels in these applications.

The poor CRI and efficacy of mercury lamps for general lighting appear to mark them for obsolescence; in the United States, sales of mercury lamps have been declining every year since 1983. Mercury lamps have an efficacy ~50 lm/W, about half that of HPS and M-H lamps, with a CRI of 16 with a clear bulb and up to 50 with a phosphor-coated bulb—well below that of M-H lamps. However, short arc, super high pressure (150–300 atm.) mercury lamps generate an intense white-blue light, due to strong broadening of mercury spectral lines. The high luminance and small dimensions of these arcs provide very efficient lighting suitable for various projection applications (e.g., video and data projection).

The most common type of HPS lamp is designed to operate with sodium pressure that gives maximum efficacy ~140 lm/W, with a CRI ~22. Sodium pressures are optimized to about 100 torr, to provide the maximum radiation in the sodium D line.

HPS lamps are also produced with improved CRI, at the expense of reduced efficacy and lamp life, by increasing the sodium pressure. One family of lamps has a CRI of 60, with a 15–20% reduction in efficacy, while a second family of 100 W lamps produces "white" light, with CRI of 80 but

an efficacy ~50 lm/W. At high sodium pressures, there is excessive resonance broadening of the sodium resonance line; this has the effect of shifting radiation away from the yellow region of the spectrum and into orange-red, where the eye is less sensitive. The higher sodium pressure also causes more rapid rates of adverse reactions with tube wall and sealing cement; life ratings are typically reduced by 50% or more. These lamps, nevertheless, have much higher efficacy and longer life than incandescent lamps and are therefore suitable for indoor lighting. Pulsed excitation has also been used to achieve white light in an HPS lamp.

High-pressure sodium lamps using reduced mercury doses are becoming increasingly popular in the United States, since they are not subject to failure due to cycling and because the lower mercury content ensures that they meet federal standards for mercury leeching. In addition, new lamps are being introduced that are entirely without mercury. Advances in PCA technology allow operation of arc tubes at higher temperatures, so that sufficient pressure is maintained to provide the necessary arc voltage for the required arc gap length in these lamps. These new lamps also require a higher fill gas pressure to maintain the emissive coating on the electrodes, and starting aids (in the form of metal wires that extend the length of the tube or tungsten cermets that are directly bounded to the arc tube) to lower breakdown voltage.

Several different metal-iodide combinations are available in M-H lamps. The preferred combination in the United States and in Japan uses a blend of sodium and scandium iodides. Scandium provides a multiline spectrum throughout the visible spectral range, similar to the rare-earth elements, and results in a CRI ~65, suitable for most commercial applications. Sodium increases the efficacy and permits color temperatures of 3000 to 5000 K, which is the preferred range in the United States. Sometimes lithium is added for an improved CRI of 75.

In Europe, a mixture of indium, thallium, and sodium iodides have been used: indium for blue emission, thallium for green, and sodium for yellow and orange. A gamut of color temperatures is possible, depending on mixture ratios. Also common in Europe is a dysprosium-holmium-thallium-iodide mixture, with or without sodium. The rare-earth elements have multiline emission spectra giving good color rendition. Thallium adds green, improving efficacy. In the absence of sodium, the color temperature of this mixture is ~6000 K; with sodium added, it can be lowered to the 3500–4000 K range. Metal halide lamps have made a large impact in display lighting applications in Europe, with an emphasis on compact lamp shapes and high CRI (>80), with reduced life (8000–12000 hr).

The introduction of ceramic technology has resulted in M-H lamps with improved efficacy and CRI, resulting from the higher cold spot temperatures achievable with PCA. Wall reactions by halide mixtures and failures due to sealing are life-limiting factors. These lamps use rare-earth combinations of dysprosium, holmium, and thulium, together with thallium, sodium, and mercury. Some use Kr^{85} as a starting aid. These lamps show reduced color variation and improved color stability over life and CRI of 80–90. Pulsed start M-H lamps are becoming popular in the United States, due to their improved lumen maintenance. Lamps rated at 150 W and below have used pulsed start since their inception. This technology is now being extended to 175 W and higher wattages by increasing the argon fill pressure to 100–200 torr, and removing the probe starting circuit. Applications with compact jackets and reduced wattages save energy with respect to probe start lamps. Another advantage of pulsed start lamps is faster warmup and hot re-strike.

Table V gives representative data for various examples of commercially available products in the mercury, HPS, and M-H families. Figure 14 shows some of the variety of types and sizes available.

G. Research and Development

Developments in HPS lamps are focused on improved production and manufacturing, as well as environmentally friendly lamps. HPS lamps with unsaturated mercury are being extended to other wattages and metallic (cermet) strips are being developed for arc tubes to improve reliability.

As we have seen, M-H lamps are extremely complex systems, involving many tens of chemical species and many thousands of spectral lines. Uncertainties in the data for all relevant processes limit our ability to model these systems in detail. However, the great advances in both computer technology and diagnostic techniques in the last decade provide the opportunity to understand the influence of many important parameters (e.g., cold spot temperature, arc tube design, chemical constituents) on lamp performance. Models are necessary to interpret the results of experiments, which use such techniques as laser and X-ray spectroscopy. The aim of such an exercise is to provide a diagnostic technique which is able to predict the performance of a lamp after many thousands of hours, thus reducing the amount of costly and time-consuming life tests. Models are also helpful in minimizing the number of DoE experiments in optimization of lamp design and establishing manufacturing tolerances.

The major emphasis in M-H technology today is in improved color and color consistency. New materials, processes, additive systems, and seals are being developed to realize this, and low cost, mass production methods of manufacture of PCA tubes with complex shapes are

TABLE V Representative High-Intensity Discharge Lamps

Power (W)	ANSI lamp code	Bulb[a] (type)	MOL[b] (in.)	Rated life (hr)	Nominal lumens (lm)	Efficacy (lm/W)	CRI	CCT (K)
Mercury[c]								
175	H39	BT28	8.3	24,000	8,500	49	45	4,000
400	H33	BT37	11.5	24,000	23,000	58	45	4,000
1000	H36	BT56	15.4	24,000	55,000	55	45	4,000
Metal Halide (M-H)								
39	M130	T6[d]	3.94	9,000	3,300	85	81	3,000
39	M130	PAR20[d]	3.75	9,000	2,300	59	81	3,000
100	M90	ED17[d]	5.44	10,000	9,300	93	85	3,000
100	M90	ED17	5.44	10,000	7,800	78	75	3,000
250	M58	E28	8.31	10,000	23,000	92	65	4,000
400	M59	BT37	11.5	20,000	36,000	90	65	3,800
1000	M47	BT56	15.38	13,600	110,000	110	65	3,900
1500	M48	BT56	15.38	3,000	155,000	103	65	3,700
High-Pressure Sodium (HPS)								
35	S76	ED17	5.5	24,000	2,250	63	22	1,900
50	S68	E23 $\frac{1}{2}$	7.75	24,000	4,000	78	22	2,000
150	S56	E28	8.3	24,000	15,000	100	22	2,000
250	S50	ET18	9.75	24,000	27,500	110	22	2,000
400	S51	ET18	9.75	24,000	50,000	125	22	2,000
1000	S52	E25	15.06	24,000	14,000	140	22	2,000

[a] Letters designate bulb shape; digits are diameter in $\frac{1}{8}$ in.
[b] MOL = maximum overall length.
[c] Phosphor-coated mercury lamps only.
[d] Ceramic (PCA) M-H lamps.

required. Applications are changing from the traditional single-ended bulbs to reflector lamps of different designs, such as the PAR38 and PAR20. These new lamps, together with the increasing application of small tubular bulbs, are spawning the development of new compact decorative fixtures, for which traditional magnetic ballasts are too heavy, noisy, and large. Compact electronic ballasts, with their improved lumen maintenance, are being developed for these applications. As cost of electronics is reduced, increasing effort will be directed toward development of high wattage electronic ballasts.

In quartz M-H lamp manufacture, efforts are continuing to reduce tolerances for color consistency and lumen maintenance over life. Pulsed start lamps offer an opportunity to use ballasts that improve lamp performance. Both CW and magnetic ballasts are being developed, including constant wattage designs as distinct from constant current. To improve environmental aspects of these lamps, lead is being removed from glass and solder. New manufacturing processes are being developed and implemented, including the replacement of solder with welding technology. Efforts are also continuing to reduce or eliminate the mercury content in lamps.

Limitations to lamp life and lumen maintenance due to transport of electrode material to the wall of the arc tube could be eliminated if the discharge were excited at radio or microwave frequency, referred to as electrodeless lamps. Although a number of electrodeless *fluorescent* lamps have appeared on the market (See Section VIII. F.2), technological problems associated with the development of electrodeless HID lamps have proved more challenging, and to date only one product is available commercially for general lighting, the Solar 100 lamp developed by Fusion Lighting. This lamp is operated by transmitting microwave power at 2.45 GHz from a magnetron through a waveguide to the lamp, which is contained in a resonant cavity. The lamp contains argon and a small amount of sulfur, providing a very "white" light, mainly from the sulfur molecules. The concept was originally used in a 5900-W lamp, the wall power loading of 250 Wcm^{-2} being a factor of 10 higher than conventional HID lamps, but the Solar 1000 operates at 1425 W, with greatly reduced wall loading (30 Wcm^{-2}). The bulb diameter is less than 3 cm, and the lamp life is long, but the system is limited by the life of the electronics, currently 15,000 hr.



FIGURE 14 Sample of HID lamp types.

VIII. FLUORESCENT LAMPS

A. Introduction

Fluorescent lamps are low pressure discharges, filled with a noble gas (usually neon, argon, or krypton or some combination thereof) at a pressure of a few torr, with a minority of mercury (typically a few millitorr). They operate at gas temperatures of 300–700 K, with electron temperatures ~11,000 K (~1 eV). Between 60 to 70% of electrical power in these discharges is converted to UV radiation by mercury atoms. A phosphor is then used to convert the UV to visible light, resulting in a total conversion efficiency of electrical power to visible light of about 25%.

Figure 15 is a schematic diagram illustrating the major components of a conventional linear fluorescent lamp. These lamps are typically long and thin—a standard T12 40-W fluorescent lamp is 4 ft (120 cm) long and 1.5 in. (3.6 cm) in diameter. (The notation Tn refers to a tubular lamp of diameter $n/8$ in.). Very short fluorescent lamps are inefficient because the electrode energy loss is significant compared to energy converted to light. Compact fluorescent lamps (CFL, see Section VIII.F.1) are made by folding the discharge tube, often with electronics which enable them to fit into a regular incandescent lamp socket. Inductively coupled ("electrodeless") fluorescent lamps (see Section VIII.F.2) have been developed in the last decade with considerable increase in lamp life due to the absence of electrodes.

Fluorescent lamps operate at wall loadings of 0.04–0.20 W/cm^2—two orders of magnitude less than those of HID lamps. However, since fluorescent lamps require no outer jacket, overall they yield 2000–4000 lm/L of lamp volume, only a factor 10 lower than HID lamps. Since there are a wide variety of fluorescent phosphors emitting nearly every color of the rainbow, and these may be used in blends or multilayered coatings, a very large range of colors and CRI is available from this family of lamps, without significant change in the discharge medium itself.

B. Radiation Mechanism

1. Discharge Plasma

The principal source of radiation in fluorescent lamps is in the UV at 254 nm, corresponding to the resonance transition from the $6\,^3P_1$ excited state of mercury to the

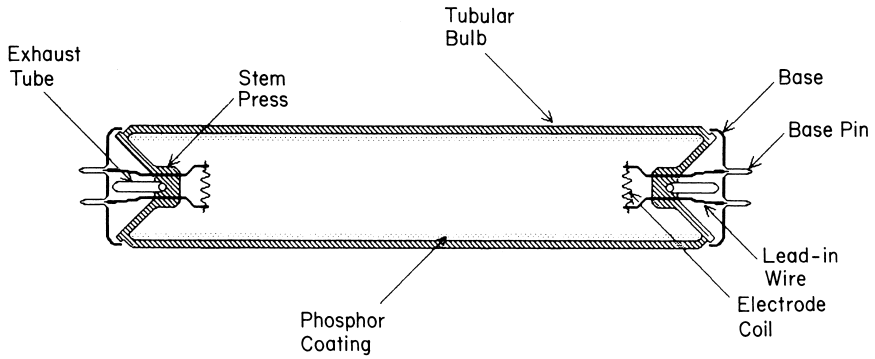


FIGURE 15 Schematic diagram of the principal components of a typical fluorescent lamp, shown in section.

$6\,^1S_0$ ground state. The discharge is very far from LTE (see Section III.A.2) and in many cases the electron energy distribution function (EEDF) cannot be described by a Maxwell distribution. The "effective" electron temperature is greater than the radiation temperature (which is itself different for the several energy levels of the mercury atom), and both are very much greater than the gas-kinetic temperature. The details of the energy balance that maintains these temperature differences are discussed in Section VIII.C.

Equation (4) in Section III describes radiation transport. The discharge has an optical depth of 50–100 at the center of the 254 nm resonance line, about unity for some of the stronger visible and near-UV lines, and much less than unity for other emission lines. There is generally little or no excitation of the noble gas atoms. The large optical depth of the resonance line results from the fact that the lower level of this transition is the ground state of the mercury atom, occupied by ∼99% of the mercury atoms present. Radiation from the plasma therefore escapes by a process of sequential absorptions and re-emissions, referred to as "imprisonment of resonance radiation." This has the effect of prolonging the effective radiative lifetime of the excited state, thus increasing the chance for dissipation of excitation energy by some nonradiative transition. The efficiency of generating resonance radiation therefore depends on the mercury atom number density (and hence vapor pressure). As mercury pressure increases, the number of atoms available for excitation also increases, increasing the total emitted radiation at each point within the discharge. However, increased mercury pressure also results in an increase in imprisonment time of radiation, and an increasing fraction of the energy used in creating excited mercury atoms is dissipated in nonradiative transitions Optimum mercury pressure results from balancing these two competing mechanisms. In conventional fluorescent lamps, the maximum efficiency of generation for 254-nm radiation occurs when the saturated mercury vapor pressure is ∼6 mtorr, corresponding to a "cold spot" temperature ∼40°C.

The mercury atom is an efficient radiator in a gas discharge because the electron energy required to excite the $6\,^3P_1$ resonance state (4.89 eV) is a little less than half the energy required to ionize a mercury atom (10.43 eV). Mercury atoms excited to the $6\,^3P_1$ state may radiate back to the ground state, or be ionized by an electron with an energy of 5.53 eV or greater, providing more electrons to maintain the discharge. This process is enhanced by the mixing between the resonance state and two neighboring metastable states—$6\,^3P_0$ (4.67 eV) and $6\,^3P_2$ (5.46 eV). The form of the EEDF plays an important role in the balance between these processes.

For these non-LTE discharges, integration of Eq. (4) of Section III to calculate the radiation flux density for all spectral lines is extremely complex, since it requires knowledge of $n(u, r)$, the radial density distribution of mercury atoms in the $6\,^3P_1$ upper level of the transition pair. Instead of being constrained to an LTE value, this number density must be determined by solving coupled rate equations involving electron-collision transitions to and from the state, as well as radiation emission and absorption transitions. These rate equations include the local electron density and the densities of all atomic species participating in electronic transitions, together with the diffusion of metastable atoms and electron-ion pairs to the wall. Thus, the total problem involves the solution of coupled integro-differential equations.

The problem of treating radiation transport in numerical models of fluorescent lamps is greatly simplified by the fact that the spatial distribution of radiating atoms in most situations of practical interest is very nearly the "fundamental mode solution" of the radiation transport equation. This occurs because discharge excitation has a broad spatial distribution, unlike laser or electron beam excitation, which can be highly localized. This fact enables modelers to replace the vacuum decay rate of the resonance level with a new, position independent constant which is referred to as the "trapped decay rate" (actually the fundamental mode decay rate) and depends on the average density of absorbing atoms, the dimensions of the discharge and the densities of atomic species responsible for broadening of the spectral lines. The trapped decay rate may be determined experimentally, or by solving the radiation transport equation *once* for each system under consideration (using an analytic method, numerical method or Monte Carlo techniques for example). In some cases, the trapped decay rate can be expressed as a simple formula in terms of the discharge parameters. The use of a single trapped decay rate for the system means that the solution of the rate equations may be decoupled from the radiation transport, greatly simplifying the numerical algorithms and reducing computation time.

2. Phosphors

Ultraviolet radiation reaching the walls of the discharge tube is absorbed and converted to visible light by coatings of luminescent phosphor powders applied to them. Such phosphors are composed of inorganic crystalline materials, optically transparent in the pure state but synthesized to incorporate specific luminescent centers to absorb the desired wavelengths of UV light and emit visible rays. Such luminescent centers are typically strongly coupled to ligand fields of the solid; therefore, excitation by a UV

photon creates excited state complexes in high vibrational energy states. These interact with lattice phonons to relax states of the excited manifold to lower vibrational energy before radiating. As a result of the loss of energy by phonon relaxation, the energy of the photon emitted on re-radiation is substantially less than that of the absorbed one. Phosphor development for application to fluorescent lamps has been an active area of research since 1934, when the first experimental fluorescence lamps were made. The discovery in 1942 of calcium halophosphate phosphors, with emission over a broad band of visible wavelengths, was a major breakthrough in the industry. These phosphors are still used in fluorescent lamps, since they are relatively inexpensive and provide CRI in the range 51–76. The discovery of rare-earth phosphors in the mid-70s, with narrow emission bands at 610 nm in the red, 545 nm in the green, and 450 nm in the blue, led to the development of three band phosphors with CRI values between 80 and 85. More recently, multiband phosphors have been developed with CRI values of 90 and above, but with lower efficacies compared to three-band phosphors.

Note that although phosphors may have near-unity quantum efficiency (i.e., one visible photon is emitted for each UV photon absorbed), energy efficiency is only ~50%, since a 254-nm UV photon carries 4.89 eV of energy, whereas a visible photon at the maximum of the eye sensitivity curve only carries 2.23 eV. Radiation output from the phosphor surface far exceeds blackbody radiation at the phosphor temperature (~40°C), in apparent violation of the fundamental conclusion of Section III. The luminescent centers, however, are radiatively coupled by UV photons to excited mercury atoms, which are in turn collisionally coupled to the electron gas in the plasma; the latter have an effective temperature of ~11,000 K. The blackbody radiance at that temperature could set the maximum thermodynamic limit to radiation output; the radiance of the phosphor surface is much less than that level.

Figure 16 illustrates spectral power distributions of (1) a triphosphor fluorescent lamp and (2) the conventional "cool white" fluorescent color that represented over two-thirds of fluorescent lamps in the United States prior to the early 1990s.

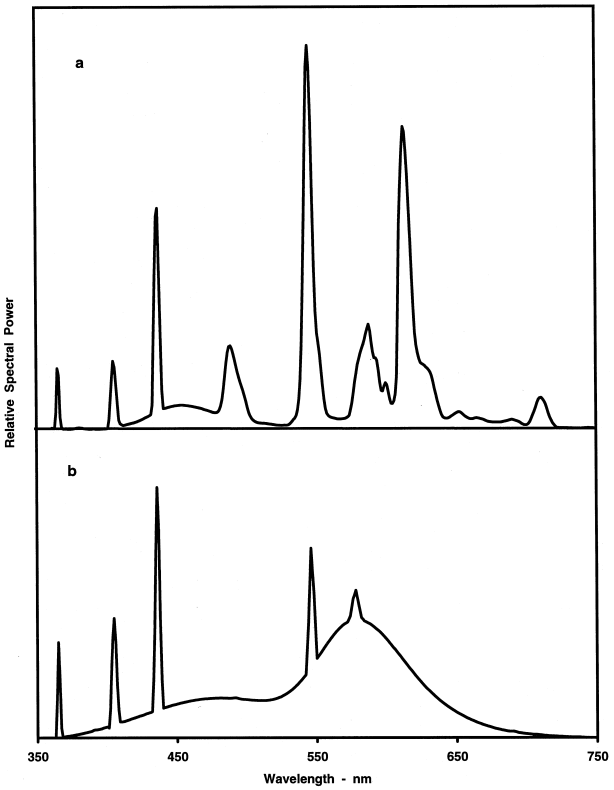


FIGURE 16 Spectral distribution of radiated power (W/nm) of two fluorescent lamp types at 40-W input power. (a) A triphosphor system including premium rare-earth phosphors: CCT = 3500 K, CRI = 85. (b) Standard cool white: CCT = 4200 K, CRI = 62. Visible mercury lines at 405, 436, 546, and 577 nm are evident in these spectral power distributions.

C. Energy Balance

1. Discharge Plasma

In a fluorescent lamp, electrons emerging from the cathode are accelerated through the cathode fall into a region of relatively weak electric field, the negative glow. There is an overproduction of ions in the negative glow, which is compensated by a dark region of low ionization, the Faraday dark space. Following the Faraday dark space is a region of constant electric field, the positive column and a bright anode glow is separated from the positive column by an anode dark space.

The positive column is the major source of UV radiation in the fluorescent lamp, and all other regions represent a loss of energy to the system, contributing to inefficient light production. As discharge length increases, the positive column increases, while the length of the other regions remains constant, leading to an increased fraction of electrical power dissipated in the positive column, and hence increased discharge efficiency. A detailed discussion of the role of electrodes and the near-electrode regions is beyond the scope of this text, and the remainder of this section concerns only the physics of the positive column.

Electrons gain energy in the axial electric field of the positive column, and dissipate their energy by collisions with atoms and ions in the gas, establishing a steady-state EEDF with an effective electron temperature

of ~11,000 K (~1 eV). Electrons with sufficiently high energy (more than 4.89 eV) may make inelastic collisions with ground-state mercury atoms, exciting them to higher states. Some of these excited states may relax to states of lower energy (including the ground state), emitting a photon with energy equal to the energy difference between the two states. In a fraction of the cases, the mercury atom (either in the ground state or an excited state) may receive enough energy to detach a valence electron, resulting in a new electron/positive-ion pair. "Two-step" ionization, in which an electron excites an atom and a second electron provides the energy to ionize it, becomes increasingly important as the power in the discharge increases. An electron of any energy can have a so-called collision of the second kind (or "super-elastic" collision) with an excited mercury atom, in which the excited atom relaxes to a lower electronic state, and the colliding electron gains the kinetic energy corresponding to this transition. Collisions of this kind substitute nonradiative transitions for radiative transitions of excited atoms, reducing the amount of radiation emitted.

All electrons make elastic collisions with gas atoms that transfer a small fraction of electron kinetic energy to the atom. Although the fraction of energy lost in any one collision is small, elastic collisions are very frequent; therefore, cumulative loss by this process represents a significant power dissipation. Electrons also redistribute their energy by collisions with other electrons. This process does not play a direct role in the energy balance, but makes a significant contribution to the electron distribution function, because electron-electron collisions tend to "Maxwellize" the EEDF. Electron-electron collisions are particularly significant for electrons with energies below the energy threshold of the first excited state of mercury, and for typical operating conditions of fluorescent lamps the EEDF may be approximated by a Maxwellian distribution (cf. Eq. (5) in Section III) for these energies. However, for higher energies (referred to as the "high energy tail"), the EEDF is depleted by loss of energy due to inelastic collisions. As power in the discharge increases, electron density increases, and for sufficiently high power, electron-electron collisions may be sufficient for the EEDF to be Maxwellian even for higher electron energies.

Because gas pressures are a small fraction of an atmosphere, the principal loss process for ions and electrons is diffusion to the walls, where they recombine to form neutral atoms. To maintain a steady state in the discharge, ions and electrons must flow to the walls at equal rates, but because electrons have a much higher mobility than ions, a radial electric field (the "ambipolar field") is established to accelerate ions and retard electrons. Ions thus diffuse away from the center of the discharge at a faster rate than neutral atoms can diffuse from the walls, leading to a depletion of mercury near the discharge axis. This process is known as cataphoresis and is more pronounced as discharge current is increased.

The noble gas serves a number of functions in a fluorescent lamp: first, it controls the ambipolar diffusion rate, which in turn establishes the required balance between ionization of mercury atoms to maintain the discharge and excitation of mercury to provide UV radiation; secondly, the noble gas acts as a buffer to reduce the ion bombardment on electrodes and the phosphor coating, prolonging lamp life; and thirdly, it helps in lamp-starting by reducing the voltage required to initiate the discharge. For a fixed discharge current (or electron concentration) in each fluorescent lamp, there is a unique value of the axial electric field in the positive column, and hence a unique potential between the electrodes (discharge voltage) at which the rate of ionization exactly equals the charged particle diffusion losses in steady-state operation. The ballast circuit must therefore include internal impedance that decreases discharge current when current through the lamp increases, so that the discharge can operate stably. The balance of diffusion losses and ionization may be influenced by the choice of noble gas, gas pressure, and discharge radius. Diffusion losses increase with decreasing molecular weight of the noble gas, reducing gas pressure, and smaller tube radius. Thus for steady-state operation, the discharge voltage tends to be lower for krypton than argon or neon, for higher pressure and larger tube radius. The situation is further complicated by the fact that there is an optimum gas pressure for maximum efficiency of UV production. In designing a fluorescent lamp, therefore, a compromise must be reached between the efficiency of operation and the electrical characteristics required by the ballast.

The efficacy of the discharge is principally determined by the fraction of the total power dissipated by electrons in the discharge which reaches the phosphor in the form of 254-nm radiation. For a standard 40-W fluorescent lamp, 10–15% of the input power is dissipated in electrode losses, 23% as elastic collision loss, 2% as visible and near-UV radiation, 8–9% as 185-nm resonance radiation, and 55% as 254-nm UV radiation. The astonishing efficiency of conversion of power in the plasma into 254-nm resonance radiation is the reason for the technological importance of this type of light source. Even allowing for a reduction of 50% efficiency in the conversion of UV to visible light at near unity quantum efficiency, fluorescent lamps still have high luminous efficacy (80 lm/W in commodity types, and up to 100 lm/W in premium types).

From the above, it may be concluded that the efficiency of conversion of electrical power to UV radiation is dependent on the fraction of high-energy electrons in the EEDF. On the other hand, the electrical characteristics are

largely determined by the electrical conductivity, which is principally determined by the low-energy electrons. The EEDF thus plays a key role in determining the parameters of the discharge. In principle, this could be calculated from the Boltzmann equation, including all the collisional processes described above, but in practice numerical models use analytic approximations to describe the EEDF.

The development of reasonably accurate discharge models incorporating rate constants for ~25 different collision processes as well as radiation transport was a major step forward in the understanding of the physics of this unique device. However, it did not result in significant improvement in efficacy of standard fluorescent lamps, which had already been optimized well by patient, empirical Edisonian techniques. However, a variety of new fluorescent lamps are being developed for a number of applications, which have operating parameters far removed from those of the standard fluorescent lamp, and this offers the chance for fluorescent lamp models to play an active role in lamp design.

2. Lamp-Electrical-Circuit System

Fluorescent lamps operate on alternating current, and the necessary current-limiting impedance provided by the ballast can be reactive rather than resistive. (Fluorescent lamps cannot be operated on continuous direct current, since mercury ions accelerate in the axial electric field as they are formed, depleting the mercury in the discharge, which accumulates at the cathode.) As in all gas discharge lamps, there are power losses in the ballast that must be added to lamp power to determine system power and efficacy. The power loss in a commodity-type (magnetic) ballast is typically 6 W per lamp. System efficiency can be improved by the use of electronic ballasts, using semiconductor components and operating the lamps at ~25 kHz operating frequency. Impedance losses can be reduced at these frequencies, and the lamp itself is more efficient because the discharge is maintained throughout the mains cycle, reducing electrode losses. For commercial lighting, the considerably higher cost of electronic ballasts is compensated by the energy savings represented by improved lamp efficacy. The electrical current in the lamp is maintained by thermionic electron emission from the cathodes. The cathodes are multiply-coiled helices of tungsten, the interstices of which are impregnated with alkaline-earth oxides for enhanced electron emission. During normal operation, they are heated by the passage of current through the tungsten wire of the helices and by bombardment of ions which gain energy in the cathode fall. In common with all discharge lamps, fluorescent lamps require a higher applied potential for ignition than for steady-state operation. There are three principal methods of accomplishing this. The simplest method of ignition is the switch-start system, preferred in most of the world except in the United States, of which the glow starter switch is the most common. A small glass bulb containing two bimetal contacts is filled with a low-pressure gas and mounted, together with a capacitor, in a container with two contact pins. This glow starter can be fitted into a socket, is inexpensive, and can be easily replaced. Before switching on, the bimetal contacts are separated by a small gap. On ignition, a glow discharge forms and the contacts heat and bend toward each other until they touch, extinguishing the glow discharge and providing relatively large current through the ballast and electrodes, which warms the cathode coils. Since the circuit is inductive, a high voltage pulse (600–1500 V) is produced across the ends of the lamp when the contacts open, which rapidly ionizes the gas and current flows between the lamp electrodes. If the pulse fails to ignite the lamp, the process is repeated, leading to a repeated flicker. During normal operation the lamp voltage is too low to cause a discharge in the starter switch.

In the United States, either rapid start or instant start systems are used. In rapid start ballast systems, individual filament transformers wound on the ballast core continually supply 3.75 V to the cathodes, and the resulting current flow heats them to the temperature required to provide sufficient electrons to initiate avalanches in the breakdown process. This represents an additional system loss of 2.5 W in a T12 fluorescent lamp. Instant start ballasts apply a high voltage across the lamp terminals when the lamp is switched on and the electrodes are cold. The high voltage results in increased "sputtering" of material from the electrodes, with a consequent reduction in lamp life, but gain in efficiency due to the absence of a heater current.

Electronic ballasts, particularly in their instant start, multilamp configuration, now represent 60% of all ballasts sold in the United States for 4- and 8-ft lamps. Pending United States Government regulations for those lamp lengths will require the use of electronic ballasts in new fixtures by the year 2005, and will eliminate the rapid start magnetic ballasts for replacement purposes by the year 2020. Magnetic ballasts will be allowed for other fluorescent lamps and HID lamps.

As fluorescent lamps approach the end of their service life, the paste on the electrode coils become depleted. This generally occurs at one electrode first and the power increases dramatically in an attempt to maintain the discharge current. This in turn generates excessive heat at the worn coil, which may break and result in melted sockets or cracked lamp glass, particularly in small diameter lamps which are powered by high frequency electronic ballasts. In 1998, NEMA (National Electrical Lamp Manufacturers Association) agreed: "When lamps of 5/8" diameter

or less are operated with high-frequency ballasts, the ballast should be capable of sensing end-of-life conditions and must not operate the lamps in a manner that can result in overheating and/or deformation of the lamp base or socket." Company proprietary technologies have also been developed for lamps themselves to sense end-of-life and self-extinguish.

D. Material Limitations Influencing Performance

Fluorescent lamps are extremely long-lived devices, with rated lives up to 20,000 hr at 3 burning hours per start (3 hr on/20 min off) on rapid start circuits. There is no fundamental relationship between efficiency and life, since lamp life is primarily determined by the erosion of electron emission material from the cathodes. The process of ignition results in momentarily enhanced erosion, and lamp life is adversely affected by increasing numbers of ignitions per thousand burning hours. A lamp with 20,000 hr life rating at 3 hr per start will last close to 35,000 hr (more than 4 years) if simply ignited once and never switched off. A lamp which is started once per minute of operation, will last 2000–5000 starts—only a few hundred hours. Lamps that are intended for service requiring many ignitions per hour are operated on ballasts supplying continuous filament heat, even when the lamp is not operating, to minimize damage to the electrodes at starting. Electron emission materials that are more durable, but have equivalent electron emission as the alkaline earth oxides, would increase the flexibility of operation of fluorescent lamps in intermittent service, but no such materials are currently available. A major material limitation in fluorescent lamp operation is the durability of fluorescent phosphors. Phosphors deteriorate in operation as a result of photolytic decomposition and color center formation, ion bombardment. and chemical reactions with mercury, glass, and impurity gases. The rate of deterioration of phosphors under these influences increases with the wall loading (i.e., the flux density of UV rays and ions) and with wall temperature, and the degree of this deterioration depends on the particular type of phosphor. Sputtering of electrode material from the cathodes during starting also discolors the phosphor. The effective life of a fluorescent lamp will be determined by the time over which it produces sufficient light for the application required. Standard cool white halophosphate fluorescent lamps lose ~20% of their output over 8000 hr of operation and this can be reduced to ~10% by use of three-band phosphors.

Phosphor deterioration is a major obstacle to application of fluorescent lamps as substitutes for incandescent lamps, since smaller lamps require higher wall loading for comparable light output. For lamps with halophosphate phosphors, doubling the power input results in more than twice as much lumen depreciation. Three-band phosphors are much more durable under high-loading conditions and are hardly affected by operation at double and triple the customary wall loading. Thus, most compact fluorescent lamps intended for incandescent replacement service use these phosphors, the cost of which is many times that of conventional phosphors.

E. Practical Considerations

1. Fabrication Technology

Fabrication technology of fluorescent lamps has much in common with high-speed incandescent lamp manufacturing. It is highly automated with automatic transfer of parts in process from one station to another and little or no manual handling. There are two principal differences: (1), the necessity to apply a uniform coating of powdered fluorescent materials to the inside wall of the tubular bulb; and (2), the constraints imposed by processing of electrode materials and the evacuation of very large internal lamp volumes through exhaust tubes of small diameter (limited by the capability to flame-seal rapidly and reliably).

The phosphor itself is manufactured by blending raw materials of controlled particle size and firing at elevated temperatures to synthesize the proper compounds. The necessity for control of particle size in firing results from the fact that these materials are very susceptible to mechanical damage in the usual types of milling processes common to formulating paints. The phosphor is applied to the bulb, essentially in the form of a paint with the phosphor as a pigment, flow-coating over the inside surface of the bulb to a thickness controlled by viscosity and drying. The remaining binder is removed by pyrolysis, passing the coated bulbs through furnaces at nearly the melting temperature of the glass, with axial air flow to react with the carbon and hydrogen constituents. After coating and binder burnout, the stems with lead-in wires, electrodes, and exhaust tubes, are flame-sealed to the ends of the bulbs, and the assembly is transferred automatically to the exhaust machine. Because alkaline earth oxides used for the electron emission mix are extremely reactive to moisture, these materials are applied to the electrode coils as alkaline earth carbonates, and the carbonates are decomposed to oxides by heating as part of the evacuation process. Due to the large volume of gas to be exhausted in standard lamps, and the small size of the exhaust tube, conventional vacuum evacuation is not fast enough. Therefore, the air and carbon dioxide are removed from the bulb by flushing inert gas in one exhaust tube and out the other; the entire process takes place

in the viscous flow regime instead of in free-molecule flow. In this way, several liter-atmospheres of air can be removed and replaced by the noble gas filling in a fraction of a minute. Basing, aging, testing, and packing are automatic at production speeds as high as 3000–6000 lamps per hour.

Like all other light sources, fluorescent lamps are sensitive to certain impurities, which must be effectively removed even at the extremely high processing speeds dictated by necessity for low-cost manufacture. Water vapor, the bête noire of all lamps, is troublesome in fluorescent lamps as well. It dissociates in the discharge, liberating hydrogen that attacks the phosphor and liberates oxygen. Oxygen reacts with the electron-emission coating, which "poisons" the cathode, and with mercury vapor, which forms solid deposits of dark mercury oxides on the surface of the phosphor. Residual hydrocarbons from the phosphor coating application result in carbon deposition on the phosphor as well as the liberation of hydrogen that will further damage the phosphor. If the bakeout temperature is increased too much, alkali from the glass will react with the phosphor, damaging it initially. Thus the baking process must be critically controlled. Hydrogen and water vapor will also act to increase the ignition potential of the lamp in early life; this is usually a transient effect, however, because of cleanup during operation by the reactions noted. Such gas contamination may, in extreme cases, prevent the lamp from igniting the first time that it is placed in service, thus causing premature lamp failure. Testing for this effect, as well as cleanup of small amounts of impurities, is the purpose of the aging step in the manufacturing process. Compact fluorescent lamps (CFLS) are fabricated using very narrow (T4 or T5) glass tubes, which must be bent and/or fused and joined to form a compact unit which can replace an incandescent bulb. This process can now be automated and production rates have been achieved which are comparable to those for linear fluorescent lamps.

F. Commercially Available Products and Performance Characteristics

1. Electroded Fluorescent Lamps

Fluorescent lamps are used for commercial, industrial, institutional, and retail applications. There has been a major shift in recent years from the halophosphate cool white phosphors with CRI ~60 to rare-earth phosphors with CRI ~70, 80 or even 90. This change has been accelerated by the increasing replacement of the standard T12 lamps by the more efficient, smaller diameter linear T8 lamps and compact fluorescent lamps, which utilize less of the more expensive rare-earth phosphors per lamp. Fluorescent lamps are available in five standard white colors with CCT of 3000, 3500, 4100, 5000, and 6500 K.

Table VI presents characteristics of fluorescent lamps intended for commercial and merchandising illumination and some of these are illustrated in Fig. 17. The most common lamps range in length from 2 to 8 ft (60 to 240 cm) and in diameter from 5/8″ to 1.5″ (1.6 to 3.65 cm). In the United States, T12 (1.5″) lamps are generally filled with a mixture of argon and krypton. These lamps are more efficient than lamps containing argon, because the presence of krypton reduces electrode losses. In Europe, T12

TABLE VI Representative Linear Fluorescent Lamps (120 V, United States Only)

Power (W)	Bulb	Type	Rated life (hr)	Light output (lm)	Efficacy (lm/W)	CRI	CCT (K)
40	48T12	CW, RS, Hp	20,000	3,050	76	62	4,200
34	48T12	CW, RS, ES, Hp	20,000	2,700	79	62	4,200
34	48T12	CWX, RS, Hp	20,000	2,200	65	87	4,100
32	48T8	RS/IS, Tp	20,000	2,800	88	75	3,000
32	48T8	RS/IS, Tp	20,000	2,750	86	75	6,500
32	48T8	RS/IS, Tp	24,000	3,000	94	85	4,100
31	22.5T8	U, RS/IS, Tp	20,000	2,600	84	75	5,000
95	96T12	CW, HO, RS, ES, Hp	12,000	8,000	84	62	4,200
110	96T12	CW, HO, RS, LT, Hp	12,000	8,600	78	62	4,200
195	96T12	CW, VHO, RS, ES, Hp	10,000	13,000	67	62	4,200
35	60T5	IS, Tp	16,000	3,650	104	82	3,500
80	60T5	HO, IS, Tp	16,000	7,000	88	82	3,500

Bulb: Maximum overall length (inches); tubular diameter, Tn = (n/8 inches); CW, cool white; CWX, deluxe cool white; RS, rapid start; IS, instant start; ES, energy saving; HO, high output; VHO, very high output; U, U-shaped lamp; LT, for cold temperature applications; Hp, halophosphor; and Tp, triphosphor.



FIGURE 17 Typical types of fluorescent lamps used in industrial and commercial applications.

lamps often contain argon and the narrower T8 (1″) lamps include 25% krypton in the argon gas fill. The increase in discharge voltage due to the increased ion-electron diffusion losses in the narrower tube may be compensated by the presence of krypton, so that for a given rated power the T8 lamp can run with the same voltage and current as the equivalent T12 lamp and the same ballast may be used. In the United States, argon-filled T8 lamps with instant start electronic ballasts have a system efficacy above 100 lm/W. Future developments may lead to increased application of even narrower T5 (5/8″) lamps, which are more efficient than T8 lamps with much higher surface brightness.

T8 and T12 lamps may be bent into a U-shape, with a two-pin base at each tube end connecting the electrodes to the ballast. These lamps can replace linear lamps which are twice as long, in specially designed luminaires.

Compact fluorescent lamps have been designed such that the volume of the light source is sufficiently compact for them to replace incandescent lamps in commonly used fittings. Table VII gives the characteristics of a representative sample of CFLs, some of which are illustrated in Figure 18. Compact fluorescent lamps are mainly used for domestic and decorative commercial applications, where they typically consume only a quarter of the energy required by an incandescent lamp to provide the same light output. There are three basic versions of CFL, with a multitude of shapes and sizes within each. Integral-type (or self-ballasted) CFLs, as the name implies, contain the lamp and all the electronics required to run it within a single unit, complete with a base that can be fitted directly in incandescent lamp sockets. The disadvantage of this design is that when the lamp fails, both lamp and electronics must be replaced. In the second type of CFL, the tube is separate from the electronics, which are contained in purpose-designed fixtures, such as table lamps. These lamps have special caps to connect to the electronics and when the lamp fails it may simply be replaced. The third type of CFL also has a separate tube, but an adapter containing the electronics may also be purchased separately, again allowing the user to replace the lamp upon failure. Well-designed adapters can last many lamp lives and may also contain an electronic ballast.

Tubes may be convoluted in one plane, circular, coiled, or have two or more parallel-limbed tubes. The tubes may be placed in a reflector or covered by a diffuser to give them the appearance of a globe, bullet, or regular household

TABLE VII Representative Compact and Electrodeless Fluorescent Lamps

Power (W)	Shape	Length (in.)	Diameter (in.)	Rated life (hr)	Light output (lm)	Efficacy (lm/W)	CRI	CCT (K)
Separate ballast								
5	Single tube	4.2	T-4	10,000	230		82	2,700
18	Double/quad tube	6.0	T-4	10,000	1,250		82	3,000
18	Triple tube	4.3	T-4	10,000	1,200		82	3,000
18	Triple tube	9.0	T-5	12,000	1,250		82	4,200
Integral ballast								
15	Triple tube	5.6	T-4	10,000	925		82	3,000
20	Triple tube	6.3	T-4	10,000	1,280		82	3,000
23	Triple tube	6.8	T-4	10,000	2,750		82	3,000
20	Circle	3.5	T-9	10,000	1,200		82	3,000
13	Globe	5.1		10,000	600		82	3,000
Electrodeless								
100	ICETRON	12.3	T-16.5	100,000	8,000		80	3,000
85	QL	19.2 cm	11.0 cm	100,000	6,000			2,860
65	Genura			15,000	1,100		82	2,700



FIGURE 18 Sample of compact fluorescent lamps.

(A-line) incandescent bulb. They have typical lives of 10,000 hr, CRI of 82, efficacy in the range of 55–60 lm/W and CCT between 2700 and 3000 K, for self-ballasted lamps and 2700 to 5000 K for CFL in dedicated fixtures. Some CFLs contain an amalgam in place of liquid mercury, which greatly extends the temperature range over which they can be operated at high efficacy. Most nonamalgam CFLs are designed for maximum efficiency when operated base up, such that the excess mercury collects at the end opposite the base. When these lamps are operated base down, the mercury collects near the electrodes, where the temperature is higher and the mercury vapor pressure is consequently higher than the optimum value. Efficacy of amalgam-containing CFLs are independent of orientation.

2. Electrodeless Fluorescent Lamps

The prospect of developing electrodeless lamps has challenged the lighting industry for more than a century, since Tesla demonstrated in 1891 that light could be produced in the presence of a high-frequency electromagnetic field. The introduction of lamps based on the electrodeless principle for general purpose lighting has been hampered by the lack of readily available, compact, cheap, and reliable driving electronics. However, in the last decade, progress in semiconductor electronics and power switching technology has made this approach commercially feasible. A number of companies have recently introduced electrodeless fluorescent lamps, in which the power is introduced in the discharge by inductive coupling of radio frequency power. The plasma in an inductive discharge forms a single turn secondary to an exciter coil, which may be placed in or around the discharge. Provided sufficient power is applied to maintain the discharge, high coupling efficiencies may be achieved at low frequencies, with the benefits of reduced electromagnetic interference (EMI) and cheaper electronics. The principal features of currently available electrodeless fluorescent lamps are schematically illustrated in Figure 19 and the most important properties (system power, life, efficacy, and CRI) of these lamps are included in Table VII.

QL (Philips) was introduced in 1991 and operates on the re-entrant cavity concept patented by Bethenod *et al.* in 1936, illustrated schematically in Fig. 19a. The inductive coil is placed within the re-entrant cavity and consists of a number of turns on a ferrite core. Its appearance resembles a regular incandescent bulb, although somewhat larger. The main feature of the *QL* lamp is its exceptional life (100,000 hr) making it ideal for situations where maintenance is difficult. Separating lamp and electronics via

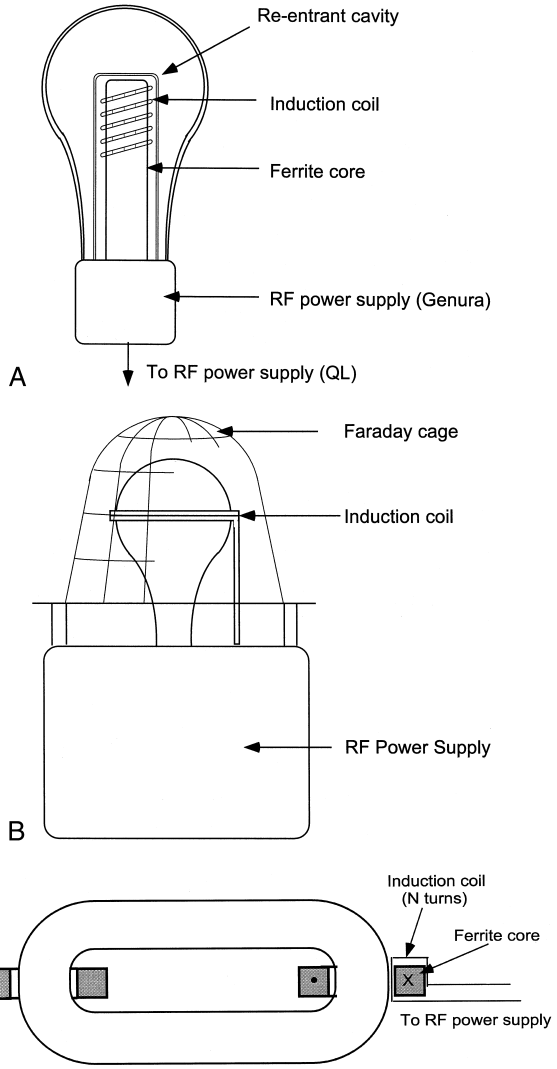


FIGURE 19 commercially available electrodeless fluarescent lamps. (a) Re-entrant cavity lamps (QL, Genura). (b) Lamp with outer coil (Everlight). (c) Toroidal lamp (ICETRON/ENDURA.)

a shielded coaxial cable reduces heating problems and a heat pipe is used to take excess heat from the ferrite. The lamp operates at a gas pressure of a few hundred millitorr of argon and the operating frequency is 2.65 MHz, a band allowed for lighting.

Genura (GE Lighting) was introduced in 1994 and also follows the re-entrant cavity concept (cf. Fig. 19a). It is the first compact electrodeless fluorescent lamp with integrated electronics, intended as a high-efficacy direct replacement for incandescent reflector lamps. As in QL, Genura operates at 2.65 MHz, but with a gas pressure of a few hundred millitorr of krypton. Efficacy is a factor 4 greater and life a factor 7.5 longer than the lamp it replaces.

Everlight (Matsushita) was introduced in 1991 in the Japanese market only. The coil is wound on the outside of the lamp (cf. Fig. 19b) requiring increased electromagnetic screening compared to other inductively coupled lamps, which is provided by a mesh screen construction outside the lamp. The lamp operates at the industrial frequency of 13.56 MHz using neon as a buffer gas to provide visible light during starting, and is intended as an incandescent replacement.

ICETRON (OSRAM SYLVANIA)/ENDURA (OSRAM Gmbh) has the lowest applied frequency (250 kHz) of all available electrodeless lamps and has a design similar to that of a toroidal device used for fusion (cf. Fig. 19c). The lamp forms a closed tubular loop, which passes through two toroidal ferrite cores, each wound with an induction coil. The voltage in the lamp is thus induced by a closed magnetic path. ICETRON/ENDURA is a high-power, high-brightness lamp with emphasis again on life (100,000 hr) and use in inaccessible areas. The lamp operates at a few hundred millitorr pressure of krypton or argon.

G. Research and Development

Research and development in fluorescent lamp technology is focused on four areas: development of more efficient fluorescent lamps for commercial and merchandising service; development of improved phosphors; reduction of mercury in fluorescent lamps to comply with increasing environmental legislation; and development of compact fluorescent lamps for incandescent replacement, with the eventual hope that they can supplant incandescent lighting in many residential applications.

The first of these areas involves a range of activities from changes of diameter, fill-gas mixture and pressure, and phosphor improvements, to fundamental studies aimed at increasing the primary efficiency of generation of resonance radiation.

Phosphor development is an activity that has been ongoing since the introduction of the fluorescent lamp 60 years ago. In the past, phosphor development has focused on extending the range of colors available and on evolutionary improvements in quantum efficiency and lumen maintenance of existing commercial phosphors. The primary focus at the present time is on discovering less expensive substitutes for the rare-earth phosphors that, nevertheless, retain the high efficacy, good color, and extreme durability under high loading of those materials.

Mercury is absorbed in phosphor and in the electrode materials of fluorescent lamps and important research activities are continuing to minimize this process and reduce the mercury content in these lamps. Considerable progress has been made in this area, leading to a series of "ecological" products, distinguished by the presence of a green label or cap on the fluorescent tube.

Life of conventional linear and compact fluorescent lamps is largely determined by the electrodes and the development of more robust electrodes or more durable electron-emitting materials could make a significant contribution to the efficiency of these lamps, since it would allow them to operate at lower gas pressures, where efficacy is higher. Electrodeless lamps operate at lower pressures, with lamp life limited only by the electronics and the phosphor, but much development is still required if these lamps are to become true incandescent lamp replacements.

IX. LOW-PRESSURE, NON-MERCURY LAMP

A. Low-Pressure Sodium Lamps

The low-pressure sodium (LPS) lamp was introduced in the 1930s and is still the most efficient light source available. Low-pressure sodium lamps emit radiation principally in two resonance lines, 589.0 and 589.6 nm (*D* lines), which are near the peak of the eye sensitivity curve, corresponding to about 520 lm/W. In fluorescent lamps, the conversion efficiency of 254 nm radiation to visible light using current phosphors is about 160 lm/W, so an LPS lamp with the same energy efficiency in converting electrical power to resonance radiation as a fluorescent lamp would have more than three times the efficacy, i.e., about 300 lm/W. In the 1930s, LPS lamps achieved 50 lm/W, but significant improvements since that time have led to efficacies as high as 200 lm/W. However, the radiation produced is principally yellow, which gives very poor color rendition, and the use of these lamps is mainly restricted to outdoor applications, such as highways, where brightness is more important than color definition. Even here, these lamps represent a very small fraction of such lighting.

Low-pressure sodium lamps contain a minority of sodium vapor in a buffer gas which is typically neon.

The optimum vapor pressure of sodium in these lamps is 3 mtorr, which is attained at a temperature of 260°C, compared to 40°C for the optimum mercury pressure in a fluorescent lamp. In order to maintain the lamp at the required temperature, additional means of heating the lamp are necessary. In early LPS lamps, this was achieved by including a neon buffer gas at a pressure of 20 torr, to provide gas heating by elastic scattering from electrons, and enclosing the arc tube in a vacuum-sealed outer jacket to reduce losses due to convection. In the current generation of LPS lamps, a coating on the inside of the outer jacket, which reflects IR radiation back to the arc tube, reduces the heat losses still further, enabling the lamps to operate with neon pressures below 10 torr. The main difficulty in attaining the theoretical efficacy limit of the LPS is to achieve both the optimum current density and wall temperature simultaneously. The discharge power may be increased for constant current by raising the discharge voltage or lengthening the lamp. The latter solution has the advantage, in common with the fluorescent lamp, of reducing the fraction of input power which is lost due to the electrodes. The discharge in LPS lamps is lengthened by bending the tube into a U-shape; in a 90 W LPS lamp, 30% of input power is output as visible radiation, 5% as IR radiation, 22% of losses are due to electrodes, and the remaining 43% is due to other losses, such as gas heating.

Conventional LPS lamps operate on a 50/60 Hz ballast. The effect of sodium depletion by cataphoresis (see Section VIII.C.1) can be reduced by use of a square wave ballast, to reduce the peak value of the current, but this has limitations due to the control gear requirements. Increases in efficacy of 10–20% have been reported in discharges at operating frequencies of 100–400 kHz. This is well above the frequency associated with sodium ion diffusion (1 kHz) and of the order of the decay of sodium atoms (100 kHz) in a typical LPS discharge. In contrast to fluorescent lamps, operation much below 100 kHz can lead to a reduction in efficacy compared to 50/60 Hz operation, while a monotonic increase in efficacy is found as the operating frequency is increased from 100 to 400 MHz.

A further problem affecting the efficacy of LPS lamps is associated with radiation transport. Radiation trapping of the *D* lines is more effective than trapping of the 254-nm line in mercury discharges, and most of the radiation from the lamp is emitted near the surface. In a U-shaped lamp, some of the radiation emitted in one leg may be trapped in the other. Changes in the shape of the lamp, to replace the circular cross section with one with an increased ratio of surface area to volume (such as a cross or crescent shape) have proved successful in increasing the efficacy of these lamps.

During the starting phase, before the lamp has heated sufficiently to vaporize the sodium, the discharge is essentially a rare gas plasma. This phase lasts typically 10–15 min. A small amount (0.5–1%) of argon is added to the neon, forming a so-called Penning mixture, to reduce the starting voltage. Penning ionization occurs when ground-state argon atoms collide with neon metastable atoms to produce argon ions. During the long starting phase, however, argon ions diffuse to the wall and may become embedded in the glass arc tube, leading to the removal of argon from the gas mixture and a rise in starting voltage. The glass used in LPS lamps must therefore be both non-absorbing for argon, and resistive to chemical interaction with sodium.

B. Rare Gas Discharges

Rare gas discharges have been the subject of extensive research for more than a century. However, low-pressure discharges in rare gases have not been used for general lighting. The most readily excited radiation spectrum is in the VUV and the phosphor conversion efficiency to visible light is inefficient. Rare gas discharges have been used in special light sources such as street signs and photocopier lamps, as well as plasma displays. These discharges have the advantage that their performance is independent of the ambient temperature, unlike fluorescent and LPS lamps, and are therefore suited to outdoor applications. The efficacy of these discharges would be greatly improved if phosphors could be developed with quantum conversion efficiencies to visible light which are substantially greater than 1.

Heavy rare gases such as neon, argon, krypton, and xenon have an identical structure for the major energy levels. Therefore, the elementary processes in the discharge plasma of these gases are identical and may be described in a similar way. The major excited levels are the $1s$ (4 levels) and $2p$ (10 levels). Resonance radiation from two of the $1s$ levels is in the VUV, while visible radiation is from the $2p$ to the $1s$ levels. The principle visible lines in neon are red, argon in the near infrared, krypton yellow, and xenon blue.

Neon discharges are the most efficient of all rare gases in producing visible light, and the bright red neon line is finding application as the third stop light on some automobiles. The narrow bore (4-mm outer diameter) and uniform color make it ideal for this purpose. The lamp runs on a direct current electricity supply and starts virtually instantly, so that, in common with LEDs, the human response is 150 msec shorter than for a comparable incandescent lamp with a red filter. Efficacy is 13 lm/W, comparable to incandescent lamps. These narrow bore lamps run with a discharge current of 8 mA,

the current density being comparable (or higher) than in standard fluorescent lamps. In the 2–20 torr pressure range, neon discharges are also the best of the rare gases in providing visible light from VUV when converted by currently available phosphors. Use of a green-emitting phosphor, together with the visible radiation can provide an amber light suitable for turning signals on automobiles.

C. Dielectric Barrier Discharge lamps

Dielectric barrier discharges are finding increasing applications as UV sources. Recently, OSRAM introduced a new lamp (PLANON) based on this principle which, through pulsed excitation, has achieved an efficiency of 60% conversion from electrical power to VUV xenon excimer radiation. The lamp is filled with 100 torr xenon and is extremely thin ($\sim$10 mm) with large surface area ($\sim$800 cm^2).

The dielectric barrier discharge is a nonthermal plasma in which the EEDF is very far from LTE. A dielectric coating covers at least one electrode to limit the electron current. The discharge is filled with high-current density (100–1000 A cm^{-2}) micro-discharges of radii $\sim$0.1 mm. The duty cycle of the voltage pulse must be adjusted to minimize electron density (and hence the high-energy tail of the EEDF) while maximizing the production of metastable Xe* atoms. The metastable atoms then combine with ground-state Xe in a three-body interaction to produce the Xe_2^* excimer which has a broad band emission centered at 172 nm.

X. LUMINESCENT LIGHT SOURCES

A. Powder Electroluminescent Panels

Powder electroluminescent panels are made by embedding a thin layer of electroluminescent phosphor in a dielectric medium between two parallel plate electrodes, one of which is transparent. Light is emitted when a voltage (AC or DC) is applied between the electrodes. In the case of AC panels, a white layer of high dielectric constant material is usually placed between the opaque electrode and the phosphor layer. This concentrates the electric field across the phosphor grains, minimizes the electric breakdown in the cell, and reflects light generated in the cell back through the transparent electrode.

Only a limited number of phosphors are available for these panels. The most common phosphor used in AC panels is zinc-sulfide activated by copper, which provides a green light. The spectra can be augmented by application of organic fluorescent paints. The effective life of these panels is governed by the lumen maintenance—for ceramic panels operated at 50 Hz, the time to reach half of initial lumens is 6000 hr. At low frequencies, panels run at low-current density ($\sim 1.5\ \mu A mm^{-2}$) and remain close to room temperature, but current density increases with frequency and the panels become hot, and the effective life is reduced to a few hours at 5 kHz, with a compensating increase in luminance.

These panels are not bright enough to be applied for general lighting and the main current use of AC panels is for backligthing liquid crystal displays (LCD). Direct current (DC) panels have found little acceptance in the market place, although they are a subject of continuing research and development.

B. Thin Electroluminescent Films

In this application, a thin phosphor film is used in place of a powder layer. The film is placed between two insulating layers, which separate the film from the two electrodes. The transparent electrode is a layer of indium-tin-oxide, coated on a glass substrate. The most common phosphor used for this application is zinc sulfide activated by manganese, which provides a yellow light. The most common form of panel has a 0.5-μm phosphor layer, separated from the electrodes by 0.2 μm layers of yttrium oxide.

Because of their high luminance (300–3000 cdm^{-2}, depending on voltage and applied frequency) thin film panels are used in matrix-addressed graphic displays and alphanumeric indicators.

C. Light-Emitting Diodes (LED)

Light-emitting diodes have the same electrical characteristics as simple silicon or germanium diodes. Pure semiconductors have a very high electrical resistivity, and the electrical characteristics can be radically altered by doping the crystalline material with a small amount of impurity. The dopant may produce an excess of electrons in the conduction band, forming an n-type material, or remove electrons, leaving positively charged "holes" in the valence band, forming a p-type material. If a forward bias voltage is applied to a crystal in which the conductivity changes from p-type to n-type within a small transition region, a current flows and electrons and holes may recombine, releasing energy in the form of heat or light. In "direct gap" semiconductors, the energy released is approximately that of the band gap. By suitably controlling the introduction of particular impurities, "indirect gap" semiconductors have been developed which generate electronic states in the forbidden gap, such that during recombination photons are released with energies less than the band gap.

The most important materials for commercially available LEDs are gallium arsenide, with a band gap of 1.49 eV in the infrared, and gallium phosphide, with a band gap of 2.25 eV, near the peak of the eye sensitivity curve. By adjusting relative fractions of these materials in a gallium arsenide phosphide alloy, intermediate band gaps with emission in red, orange, and yellow may be realized. Manufacturing of LEDs has much in common with standard semiconductor processing of multilayer devices.

Light-emitting diodes have found widespread application as indicator lamps, because of their robustness, small size, and long life. They are also being used in some automobiles for tail and stop lights. Considerable research activity is being undertaken to increase the efficacy of these devices for application to general lighting. One approach is the use of organic materials, such as light-emitting organic polymers, with multiple regions of emitting material in different colors for white light. In common with other developments in the semiconductor industry, rapid technological advances are being made. However, the introduction of products to compete with the major light sources described in this article is still some years away.

XI. SUMMARY

A. Cost Comparison

A comparison of the cost of purchase and electricity consumption of typical examples from each of the main lamp families is summarized in Table VIII. Lamp cost can vary considerably within each lamp family, and common, lower cost lamp types have been used to represent the lamp families. The cost of energy is taken as $0.08 per kilowatthour, an average value for the United States in the late 1990s. Because the lumen output and the lives of the different lamp types vary, the comparison must be based on equal amounts of light, and both lamp and energy costs are prorated on a million lumen-hour basis.

TABLE VIII Cost Comparisons for the Major Lamp Families

Lamp family	Cost per million lumen-hours ($)			Energy cost as a percent of total cost
	Lamp	Electric energy	Total	
Incandescent	1.04	4.70	5.74	82
Tungsten halogen (T-H)	1.36	4.17	5.53	75
Fluorescent	0.10	1.02	1.12	91
Mercury	0.10	1.70	1.80	95
Metal halide (M-H)	0.11	1.28	1.39	92
High-pressure sodium (HPS)	0.07	0.71	0.78	92

The total cost to produce one million lumen-hours for each lamp family (column 4 of Table VIII) is obtained by combining the cost of lamps (column 2) with the cost of energy (column 3). Clearly, incandescent and T-H lamps are the most expensive way to produce light, while HPS lamps are the least expensive. Regardless of the lamp family, energy cost is the principal component in the cost of light (column 5). For this reason, the cost of producing light can generally be reduced by selecting a more efficient lamp within a lamp family, even if the cost of that lamp is greater.

Analysis based on lamp and energy costs is an indicator of the relative cost of light as produced by the various lamp families. However, when the application is known, life cycle cost analysis can determine absolute cost of alternative lighting systems. This uses the present value of the energy and all other costs associated with the light such as ballasts, luminaires, periodic maintenance, installation and electrical distribution, impact on heating/cooling systems, ownership costs, etc.

HPS lamps are commonly chosen for applications where their relatively poor CRI and undesirably low color temperature are not important issues, typically for out-of-doors lighting and in some industrial situations. Use of mercury lamps has been declining for many years, its use supplanted by HPS with much higher efficacy but poor color properties, or by M-H lamps with higher efficacy and superior color properties. For some applications, the consumer may be limited to particular source types. For example, of the basic lamp types, incandescent and T-H are best suited for narrow beam spotlights, whereas fluorescent lamps cannot be used for this purpose.

B. New Market Developments (2000)

Despite the advantages of more efficient light sources with respect to "total cost of light," illustrated in Table VIII, the penetration of the United States market by energy-efficient lamp products was initially slower than anticipated. In particular, the replacement of incandescent by compact fluorescent lamps in residential service was disappointing. The principal obstacle to this substitution was the high first cost of compact fluorescent lamps compared with incandescent bulbs. The residential purchaser strongly resisted paying a higher first cost, even with the anticipation of recouping the higher price several times over in reduced electricity bills later.

In the early 1990s a new factor entered into the decision-making process. There was a growing shortage of electric-generating capacity in some areas of the country, especially the Northeast. Here, the electric power distribution system had been overloaded on several occasions of high peak consumption, requiring a 5%

reduction in line voltage. In this situation of small or even negative reserve capacity, utilities recognized that substitution of light sources of reduced power consumption for equal light output offers a means of reducing peak power demand. For example, substituting an 18-W electronically ballasted compact fluorescent lamp for a 75-W incandescent lamp reduces peak demand by 57 W, thereby adding 57 W to reserve capacity to meet other peak load requirements. The enhancement of reserve capacity offered a powerful economic incentive for this substitution. The utility purchases such lamps wholesale and sells them to residential customers at a reduced cost, or sometimes leases them to the customers at a nominal monthly charge. The residential customer is freed from having to pay the high first cost of the energy-efficient product, but enjoys the savings on the electricity bill. From the utility point of view, improving reserve capacity by subsidizing energy-efficient products is cost effective. Equally important, this advantage is recognized by Public Utilities Commissions, which allow the investment to be incorporated into calculations for rate-making.

SEE ALSO THE FOLLOWING ARTICLES

COLOR SCIENCE • ENERGY EFFICIENCY COMPARISONS AMONG COUNTRIES • LIGHT EMITTING DIODES (LEDS) • LUMINESCENCE • POWER TRANSMISSION, HIGH VOLTAGE

BIBLIOGRAPHY

Coaton, J. R., and Marsden, A. M., eds. (1997). "Lights and Lighting," John Wiley and Sons, New York.

Elenbaas, W. (1972). "Light Sources," Crane Russak and Co, New York.

Proud, J. M., and Luessen, E. H., eds. (1986). "Radiation Processes in Discharge Plasmas," NATO ASI Series, Plenum, New York.

Rea, M. S., ed. (1993). "IES Lighting Handbook," Illuminating Engineering Society, New York.

Stoer, G. W. (1986). "History of Light and Lighting," Philips Lighting Division, Eindhoven, The Netherlands.

Waymouth, J. F. (1971). "Electric Discharge Lamps," MIT Press, Cambridge, MA.

Waymouth, I. F., and Levin, K. E., eds. (IYXO). "Designer's Handbook, Light Source Applications," GTE Marketing Services, West Seneca, New York.



Machine Design

Gerard Voland
Illinois Institute of Technology

Uffe Hindhede
Black Hawk College

Ralph Erisman
Carl Sandburg College

I. Fundamentals of Machine Design
II. Three Major Phases of Machine Design
III. Design Considerations

GLOSSARY

Analysis Separation of any whole into its parts for the purpose of examining these parts with regard to proportion, function, or interrelationship.
Dynamics Discipline that deals with the motion of a system of rigid bodies under the influence of external forces.
Hydro Greek stem meaning *water*, used in hydrostatics and hydrodynamics.
Hysteresis Internal molecular friction.
Involute Form of spiral that can be described mathematically.
Iteration Repetition.
Kinematics Study of motion in a system of interconnected rigid bodies without reference to the forces acting on the system.
Machine Combination of rigid or resistant bodies having definite motions and capable of doing useful work.
Material fatigue Phenomenon that leads progressively to fracture under cyclic loading.
Mechanics Study of physical systems under the influence of forces.
Mechanism Device for modifying motion, with one fixed and rigid member and one or more rigid and moveable members, connected in such a way that the motion of one compels the motion of the other.
Metallurgy Science and technology of metals and alloys.
Morphology Any scientific study of form and structure.
Statics Discipline that deals with equilibrium of rigid bodies acted on by external forces.
Synthesis Combination of parts to form a whole or entity.

MACHINE DESIGN is the process of using scientific disciplines to specify completely the construction of a mechanical device that will be able to perform some specific task. The mechanical device will have some power input source connected by suitable means to the desired functional output. The role of the designer is twofold: to describe the functional output by mechanical means and to specify the physical size of each component so that there is no malfunction when the device functions at its stated capacity. The first criterion is achieved by means of the designer's creativity, and the second is achieved by mathematical, graphic, and computational methods.

Design is also the first step toward the manufacture of the finished machine. Designers determine the specifications for all the component parts and supervise preparation of the completed engineering drawings required.

I. FUNDAMENTALS OF MACHINE DESIGN

A. Nature and Composition of Machines

A machine is a device for doing useful work such as manufacturing or transporting goods. To do useful work, a machine must have a source of power or be able to generate its own power. Machines that receive energy from natural sources such as moving air and water and the burning of fossil fuels are known as prime movers. They transform potential energy into mechanical energy and are themselves sources of power.

Secondary movers, such as electric, hydraulics, and pneumatic motors, receive their energy from prime movers. Both sources of power receive a different form of energy as input. The output, however, is nearly always a rotating shaft that can supply power to other machines.

Machines that receive energy from prime or secondary movers and do useful work are known as driven machines or operators. Besides a power source, the driven machine must have means for transmitting and modifying forces and motions to carry out its task. The devices commonly known as machines are therefore dual in their makeup in that they have a power source separate from those parts doing useful work. Figure 1 illustrates the components of such a power-work unit.

The dominant power sources today are electric motors and internal combustion engines. Motors are used primarily in stationary machines, while engines are the preferred power source in most self-propelled machines.

TABLE I Classification of Machine Elements[a]

Function	Name
Fastening and joining	Rivets, screws, nuts, bolts, pins, snap rings, keys, washers
Storage of energy	Springs (coil, torsion bar, leaf, disk)
Power transmission and speed change	Belts, sheaves, pulleys, chain, coupling sprockets, gears, friction wheels, power screws, ball and roller bearing screws, linkages
Carrying and transmitting rotary motion	Axels, shafts, sliding and rolling bearings, linkages
Transmission and control of rotary power	Pistons, connecting rods, flywheels, cams, brakes clutches
Sealing, enclosing, and controlling fluids	Seals, gaskets, valves, cocks, tubing, fittings, piston rings
Service elements	Handwheels, levers, steering wheels, cranks, linkages
Admitting, enclosing, restraining, or supporting integral machine part or machines	Machine frames, housings, and foundations

[a] Reprinted with permission from Hindhede *et al.* (1983). "Machine Design Fundamentals," Copyright 1983 John Wiley & Sons, Inc.

However simple, any machine is a combination of individual components generally referred to as machine elements or parts. Thus if a machine is completely dismantled, a collection of simple parts remains such as nuts, bolts, springs, gears, cams, and shaft—the building blocks of machinery. A machine element is, therefore, a single unit designed to perform a specific function and capable of combining with other elements (Table I). Sometimes

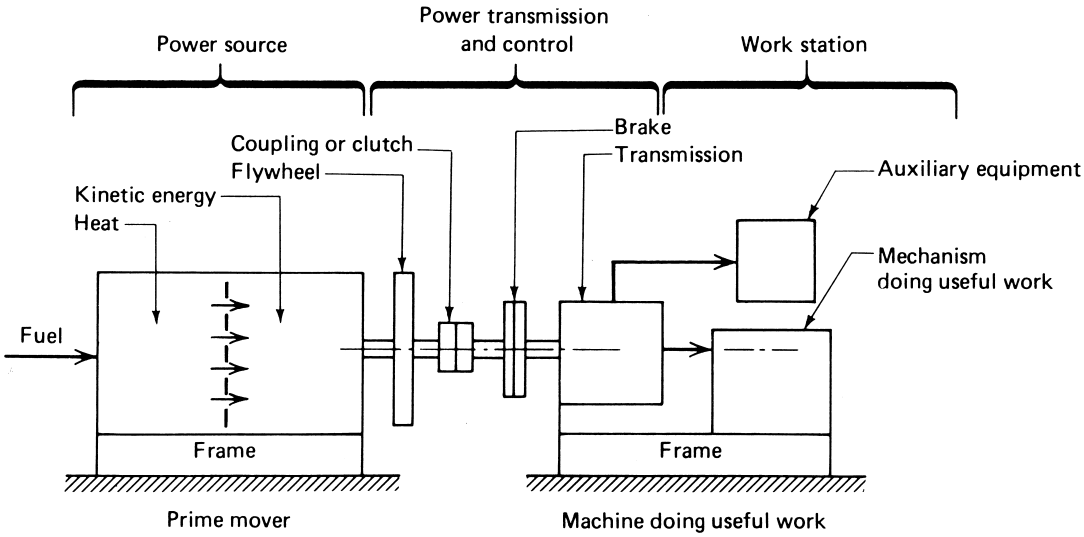


FIGURE 1 Basic components of a power–work unit. Power is transmitted from a source, through transmission and control, to a work station. [Reprinted with permission from Hindhede, U. *et al.* (1993). "Machine Design Fundamentals." Copyright 1983 John Wiley & Sons, Inc.]

certain elements are associated in pairs, such as nuts and bolts, or keys and shafts. In other instances, elements are combined to form specialized subassemblies or constructional elements such as bearings and couplings.

Because new inventions are constantly being made, the number of such construction elements is steadily growing. When they reach the stage of being commonly used and the patent rights have expired, they are often referred to as machine elements. Designers endeavor to increase the store of such components. Computer-aided design [CAD] systems with their vast capacity for storage are therefore of great practical importance in design.

When rigid machine members are combined to form a device with one fixed and one or more mechanically actuated members, we have a mechanism. The slider-crank, composed of a crank, a connecting rod, and a piston, is one of the most useful mechanisms in existence because of its ability to modify motion: rectilinear to rotational, rotational to linear, and so on. In engines, explosive power is applied to the slider crank while the crank rotates to provide an easily adaptable power source. With a power source at the crank, the slider can function as a pump or a compressor performing useful work as a driven machine. Thus when a mechanism is power actuated, it becomes a machine.

The other important function of machines is force amplification. Six simple machines, essentially force-amplifiers, have been known for centuries: lever, wheel and axle, pulley, inclined plane, wedge, and screw (Fig. 2). Originally conceived to relieve man of backbreaking chores, they proved equally useful when muscle power was replaced by engine power. Modern technology has added two other simple machines: gearing and the hydraulic press. Mechanisms and simple machines are the building blocks of complex machines. With eight simple machines and about two dozen basic mechanisms available, the realization of a vast array of machines is possible.

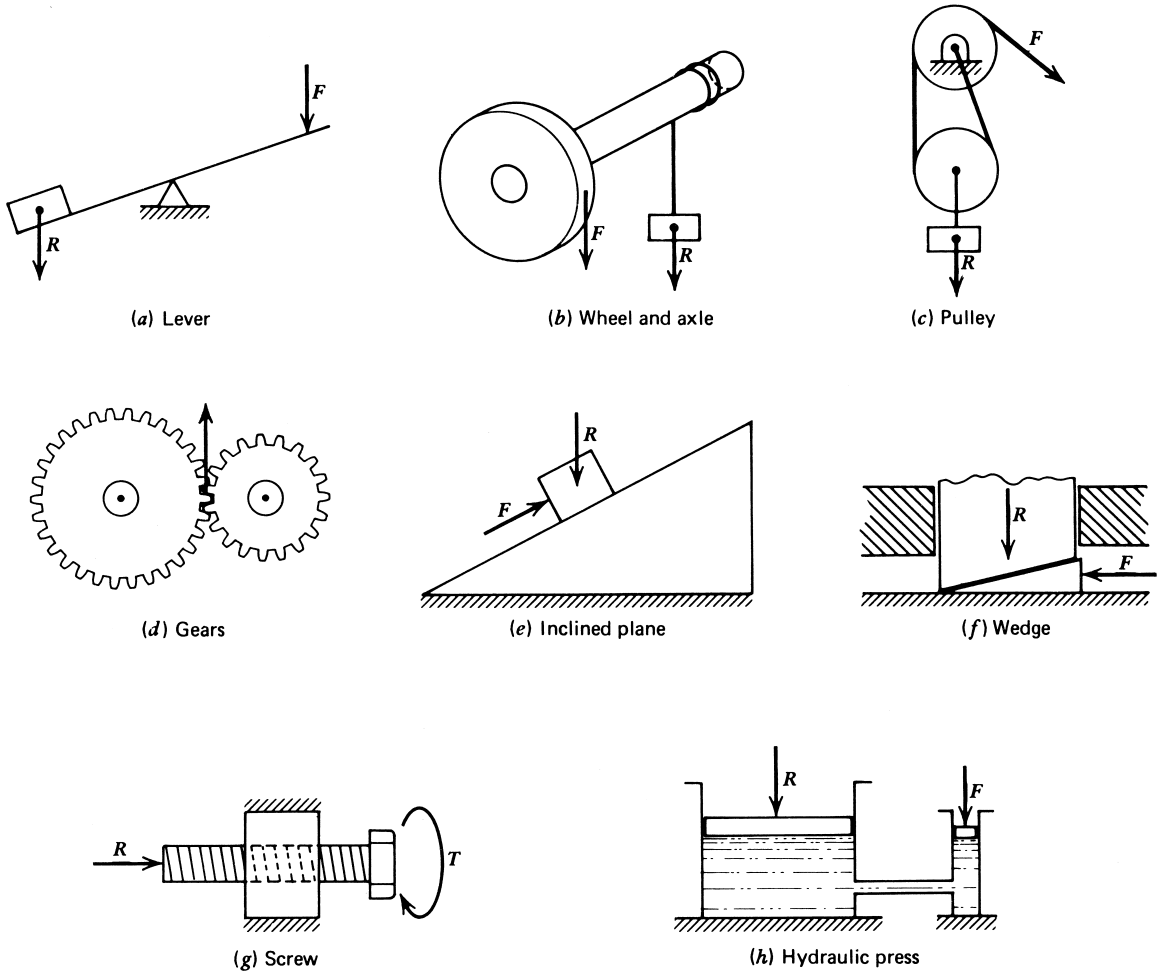


FIGURE 2 Simple machines. [Reprinted with permission from Hindhede, U. *et al.* (1983). "Machine Design Fundamentals." Copyright 1983 John Wiley & Sons, Inc.]

B. The Morphology of Design

1. The Basis of Design

There may well be only a few acceptable arrangements or configurations of parts in a complex technological device such as a motor vehicle. The arrangement, however, is neither self-evident nor scientifically predictable. The early designers of motor vehicles could not ask science to tell them where to put the engine, clutch, brake, or fuel tank. In time, of course, the wrong choices would be revealed, but not by scientific analysis.

Although many machine members in use have been influenced by science, their form and function, dimensions, and appearance may have been determined by designers, inventors, and engineers using nonscientific modes of thought. Engineers sometimes have downgraded this intellectual element of design because its origins lie in the arts and not in science. Industry, however, clearly recognizes the competitive advantage of creative thinking that results in patents or innovative designs.

The inherent difficulty of machine design once was that three-dimensional components were seen pictorially but confined to two-dimensional plan drawings on paper. Practical considerations compelled designers usually to use three such two-dimensional views in mutually perpendicular planes to depict machine parts and to show their dimensions properly. This system, known as orthographic projection, continues to be extremely useful and versatile, provided that designers have the artistic ability of visualization to make the transition from drawing to parts and from parts to drawing. This mental transition or nonverbal form of thought has been greatly facilitated by three-dimensional computer graphics and solid modeling.

2. The Significance of Design

Machine design is the single most important activity in the mechanical industries. Success or failure of a company has it roots in product design, whether it is done in-house or contracted out. It is here that manufacturing costs and profits are determined. Even the best of manufacturing facilities are of little use if the design of a product is inherently faulty.

Designers of machinery are individuals who use their talents to solve user-product problems on an on-going basis. They typically spend most of their time and effort on questions of applications and functions, usually within diminishing space requirements. That is, the designer must constantly seek to satisfy the customer's desires with respect to adaptability to the environment and will always seek to reduce the size limitations. Other concerns do not necessarily have lower priorities but are secondary in the sense that form and function are usually fixed in the designer's mind first. This means that designers must first cope with geometry. Appropriate layouts must confirm that the proposed device will not occupy space already allocated to other objects, that it will be able to reach or attach to other parts of the whole, and that it can be assembled.

Individuals who have an engineering degree, an associate degree, or who have completed an in-house apprenticeship in drafting and design usually design machines. Machine design is a team effort requiring a variety of talents. Some designers are quite inventive and many hold patents; others, usually engineers, have more analytical minds. A third category of designer is artistically inclined and skilled in adding form, shape, style, and esthetics to machinery. Machine design is an art as much as a science.

Creative designers, who bring machine elements together in new combinations, are able to assemble in their minds new mechanisms that as yet do not exist. They design and redesign their machines as they walk or drive and see them in height, breadth, and depth as real objects, and they can also see them and even hear them in action. Such nonverbal thinking, which is a central concept in machine design, involves perceptions, the stock-in-trade of the artist. Thus, design requires a challenging blend of exploratory artistry coupled directly with the application of scientific and engineering principles.

Machine design in general is concerned with development of power sources and functional mechanisms. However, designers invariably specialize; they may concentrate on such areas as design of engines or turbines or the functional portions of automobiles, machine tools, or automation equipment.

Since design is the first step toward manufacturing, it is important that potential designers have some experience in manufacturing and industrial engineering. The activities in these departments are closely related to product design. An understanding of manufacturing processes results in simpler parts, form, lower cost, and fewer in-house disagreements. Industrial engineering is a key to human engineering and design for assembly. Even time spent in sales can be valuable for designers, because designers ultimately must satisfy customers' needs.

Design is a strenuous activity because of the discrepancy that always exists between what industry would like to produce and what is currently feasible. The most successful designers are those who have the necessary talents and qualities to perform equally well in any of the related disciplines.

3. The Limitations on Design

Machine design is limited geometrically by the present capabilities of machine tools. Machine tools occupy a unique

position among machines in that they not only reproduce themselves but also produce the parts from which other machines are made. For this reason they must be more accurate than the machines they generate. Because of this demand for accuracy, they are limited to simple motions such as linear translation, plane rotation, and a combination of these. Precision-planed surfaces, parallel or perpendicular to each other, are easily produced on lathes, milling machines, and planners. Helices, the basis for screw threads and helical springs, are produced by synchronized rotation and translation. Although the design of machine parts is generally limited to such simple forms and shapes and kinematics, combining the output of simple machines can generate more complex machines.

4. Theoretical Foundations of Machine Design

Mechanics, the theoretical foundation of machine design, deals with the relations between masses, motions, and forces. The laws of mechanics are fundamental laws of nature, and designers soon learn that they are constantly enforced. They are Newton's three laws of motion, Coulomb's law of solid friction, Hooke's law of stresses and strain, the principle of energy conservation, and the laws of hydrostatics and hydrodynamics.

Mechanics has two branches: solid mechanics and fluid mechanics. Solid mechanics has three branches: statics, kinematics, and dynamics. In machine design, statics deals with the distribution of forces in rigid machine members in a state of rest or in a state of constant velocity. Kinematics deals with motion of machine members without regard to forces, and kinematic design insures that the various machine parts have proper relative motion. Dynamics, based on Newton's second law, considers motion of machine members as they are influenced by forces and acceleration. Fluid mechanics, the second principal branch of mechanics, deals with pressure, flow, and mechanical reactions of fluid media in machines.

The tools of machine designers are the disciplines of statics, dynamics, kinematics, fluid mechanics, metallurgy, and strength of materials. To these can be added failure analysis, which gives the insight needed to learn about design and use limitations in specific applications. The application and absorption of power will generate forces, reactions, and torque, either singly or in combination, in every component. The ability of the components to withstand these impositions of force is a function of the physical size and the inherent strength of the material used. Low material strength requires large sizes; high-strength material components tolerate smaller sizes. This relationship is tempered by another phenomenon: there is no appreciable difference in deflection rates between high- and low-strength steel components of the same size. From the standpoint of strength, high-strength steels can offer the advantage of smaller size; but if deflection is the design criterion, the high strength material offers no advantage.

Machine design is largely concerned with component design, that is, with proper sizing of machine elements to perform a given function at some stated life criterion.

The key component design is the ability to interpret the mode of failure and find suitable means to circumvent such failure. Failure does not necessarily mean that the part breaks but that it no longer functions as intended. There are four basic models of failure (Fig. 3):

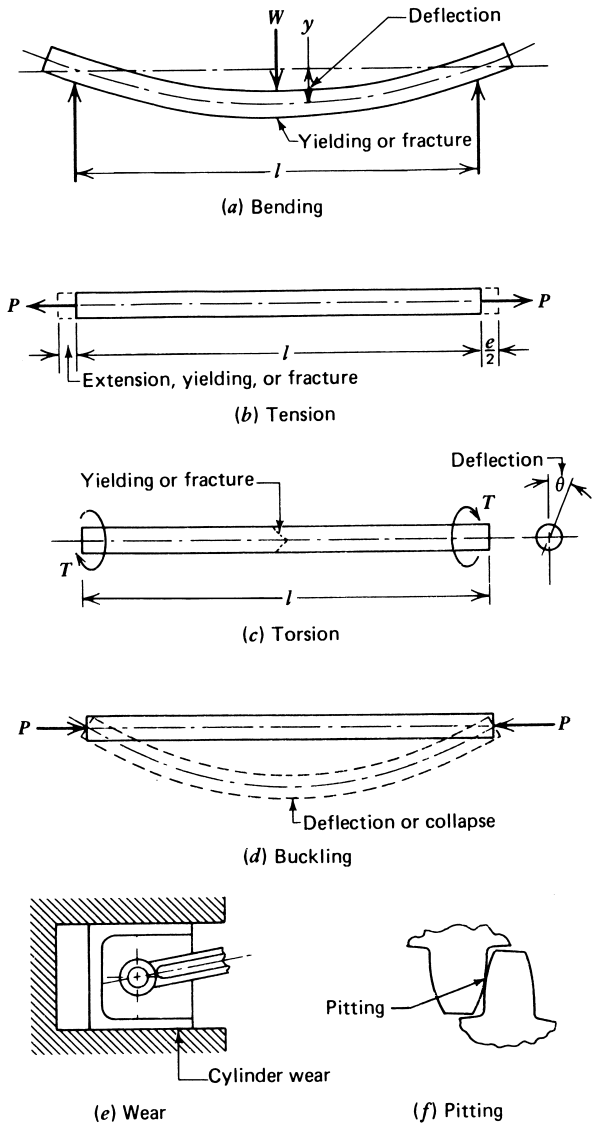


FIGURE 3 Common modes of failure of machine members. [Reprinted with permission from Hindhede, U. *et al.* (1983). "Machine Design Fundamentals." Copyright 1983 John Wiley & Sons, Inc.]

1. Lack of strength (material rupture)
2. Lack of rigidity (excess deflection)
3. Lack of stability (buckling)
4. Wear (removal of vital surface material)

The corresponding design criteria involve a comparison of working conditions with allowable conditions.

Strength:	Maximum induced stress $\leq$ allowable or working stress
Rigidity:	Maximum operating deflection $\leq$ allowable deflection
Stability:	Maximum load $\leq$ safe load
Wear:	Maximum wear $\leq$ permissible wear

Additional considerations include design against rust or corrosion and against surface crushing or indentation, which is controlled by heat treatment, material selection, and special surfacing techniques. In the United States alone it has been estimated that corrosion costs over $10 billion annually and that at least $1 billion of this is from damage done to machinery.

Heat expansion, generally a predictable phenomenon, is an important design criterion. Because machine parts are manufactured at room temperature but often operate at much higher or lower temperatures, the design must allow for expansion or contraction; otherwise, large stresses may develop. For the same reason, special considerations must be given to material selection for machines that operate in extreme environments. Many common engineering materials become brittle at moderately low temperatures.

II. THREE MAJOR PHASES OF MACHINE DESIGN

The design process begins with an analysis of simple parts and ends with a synthesis of several parts acting in unison. Because the division between analysis and synthesis is not always clear, it is difficult to treat them separately.

The machine design process can be divided into a number of interdependent steps or activities. We choose to describe the process in terms of three major phases: needs assessment, creative design, and implementation. (Although some authors subdivide the process into more numerous and specific stages of development, the process itself remains unchanged: it requires the same set of activities to be performed, however, these actions may be subdivided or described.) The first phase—needs assessment—in a machine design project is devoted to establishing that a need exists for a certain type of machine to serve a specific purpose. Often the possibility of such needs is transmitted through the sales department. The second stage—creative design—begins with a feasibility study to determine if the machine can be built at a reasonable price. If so, where and when? It is at this stage that there must be a meeting of the minds of the design, manufacturing, and sales departments. Alternative solutions then are generated, analyzed, and synthesized into a final design concept. The last phase—implementation—focuses upon creating a detailed description of this final concept, constructing and testing a prototype of the design, and finally moving it into the production mode.

Although this theoretical model of the design process is sequential and linear in its form, the reality of design is that one must work iteratively and concurrently among all the various stages of the process. Such simultaneous consideration of all aspects of the development, fabrication, operation, and disposal of a design is known as *concurrent engineering.* A design's fabrication, assembly, distribution, maintenance, repair, disassembly, and recycling (or ultimate disposal) must be considered as it is developed, not after the final conceptual design has been completed. The results of later stages of the process may have an impact on earlier design decisions, thereby necessitating that the designers revisit and modify those decisions. Concurrent engineering has reduced the need for such corrective actions, resulting in significant savings of time, effort, and money.

Of the three stages, the second (creative design) is by far the most lengthy, costly, and decisive. It is here that the project must prove itself or be rejected. This step is at first purely visionary in that the physical principle of operation for achieving the desired result is thought out, whereas no consideration is given to the final shape, rated output, or material selection. The physical principle to be employed in the solution of the problem at hand is stated and decided upon. If the principle is new, the design becomes an invention. Putting these principles into effect requires calculations and material specifications. It is a process in which the initial analysis is synthesized into layouts and specifications.

The design process takes place on a purely intellectual plane; the calculations and layouts are mere visible expressions of this mental activity. As noted, it is an iterative problem solving process in which the design progresses from the abstract to the absolute. It involves extensive processing of information that results in a move from uncertainty about the success or failure of a design toward certainty.

As designs develop, their descriptions can evolve from a series of hand-made sketches to simple line drawings, then to accurate layouts, and finally to exact geometrical drawings and computer-generated solid models. Such depictions and models of a design can be highly condensed, and often serve as a road map for all other activities in a company, such as sales, service, testing, cost accounting, advertising, and subcontracting. Moreover, they may become legal documents when work is subcontracted.

Computer-aided design (CAD) software allows the designer to create and modify concepts quickly and easily. Creativity has been greatly enhanced by CAD techniques, which permit easy access to, and combinations of, the designs of available machine elements at the touch of a button. The complete geometric form of a part can be specified in a three-dimensional solid model database. Unlike earlier computerized wireframe representations of a design in which only the edges could be prescribed, solid models also allow one to calculate the mass, center of gravity, and other important properties of the part under development. Moreover, *finite element analysis* (FEA) software can be applied to solid models in order to evaluate stress patterns, heat transfer behavior, and other characteristics. In addition, solid modeling data can be used by computer-controlled tools to fabricate a given part according to specifications, thereby reducing the time and effort needed to produce a design.

The computer is such a powerful tool that some designers have erroneously predicted the demise of practical engineering. Although computer simulation field conditions can greatly reduce actual testing, discrete practical testing is still the best assurance for guaranteeing product reliability. The imperfections of a new design can be traced to incomplete knowledge of such factors as friction, heat, or vibration. The enormous complexity of modern machines and the numerous interacting factors preclude the probability of covering all potential failures from an analytical standpoint.

III. DESIGN CONSIDERATIONS

Solutions to most design problems do not arise solely from analytical relationships but are compromises that satisfy design requirements and practical limitations such as materials, available tooling, and service ease. Design is an evolutionary process. Most designs are revised "*n*th-generation" versions of earlier solutions. Totally new machines are rare.

The following steps are necessary considerations in the analysis of a design problem.

1. A kinematic design establishes the necessary motion requirements.
2. A force analysis establishes the magnitude of the acting forces and moments.
3. A strength and rigidity analysis establishes the basic dimensions of machine members.
4. A human factors analysis insures that controls and operator interactions are safe and efficient.
5. An environmental analysis establishes any adverse effect the design may have on the quality of air, water, and soil, as well as the effect that air (oxygen) and humidity may have on the design itself (rust, corrosion, etc.).

The design process starts at the point at which useful work is done. In engine design, the top of the piston is the focal point. In a punch press it is the tool that custom shapes, or forms a part.

To buy or to build is a question raised about every major part. Many components are standardized and available at low price; hence only a limited number of machine parts are made in-house. Design, like manufacturing, boils down to dollars and cents. Design for low-cost manufacturing simply means that the parts can follow the normal path through the machine shop. Anything that is out of the ordinary entails extra cost; in fact, costs increase as soon as a part requires special attention from the manufacturing engineers.

Experience indicates that the following guidelines should be followed in designing machines.

Equal importance of parts—There is no such thing as an unimportant or less important part in any machine. When a part fails to function properly, the entire machine malfunctions.

Importance of separate functions—To economize, designers are often tempted to combine two or more separate functions into one single component. Experience indicates that nearly always the best answer is to separate the functions and design accordingly.

Uniform strength—The ideal design would have all sections in a machine part offer the same resistance to failure. This would provide maximum efficient use of material with consequent cost benefit.

One problem—one solution—A gear may be fastened to a shaft by different means, such as key and key way, set screw, or press fit. Although any single method, when properly calculated and fit at assembly, should be adequate, inexperienced designers tend to be overly cautious and use both a key and press fit. Good engineering design practice requires only one solution for each engineering problem.

Constant wall thickness—To simplify manufacturing of hollow-cast or welded parts it is common to use the same wall thickness when possible. This is clearly a case in which manufacturing considerations may take precedence over strength and rigidity so long as the result is a stronger or more rigid structure than is required (Fig. 4).

A. Force Analysis

A working machine exerts forces and has forces exerted on it due to applied loads, weight, power transmission, and accelerations of members. Movable machine members

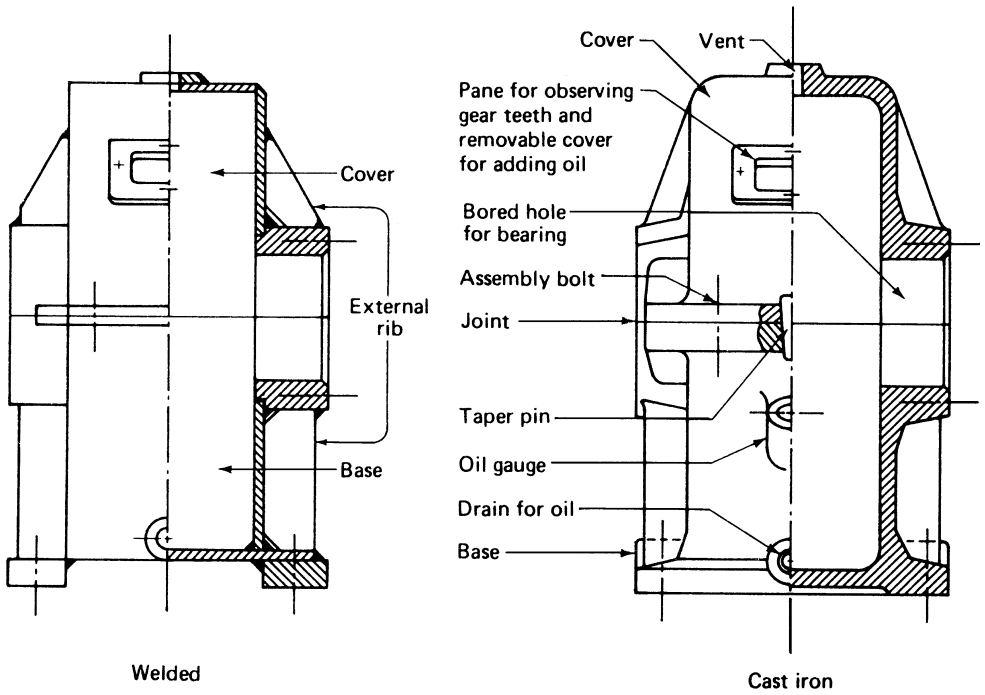


FIGURE 4 Frames with constant wall thickness. [Reprinted with permission from Hindhede, U. *et al.* (1983). "Machine Design Fundamentals." Copyright 1983 John Wiley & Sons, Inc.]

experience translation, rotation, and a combination of translation and rotation that results in changes in position or displacement, velocity, and acceleration.

Forces are dual in nature. They can do useful work, as in centrifugal pumps, or generate harmful stresses, as in flywheels. Forces are purposefully generated to do useful work, but in the course of action, there is the inevitable loss of energy due to negative work by friction and hysteresis in deflecting machine members. Forces are especially useful when they generate rotary motion and rotary power, the very foundation of machine dynamics.

In machine design, the term *load* has several meanings. It can be the force or moment exerted on a single machine member or, in a broader sense, the burden placed on a machine measured in units such as horsepower or kilowatts. An overload is a load greater than that which a machine is designed to withstand continuously.

Determination of machine member loads is often the single most difficult phase of machine design and the one most neglected in education. The fact that little can be done to determine size and shape of machine members until the loads are known only accentuates the initial difficulties.

Once the working principle (the "what" of design) has been settled, the next question is how can it be best accomplished. This amounts to the detailed analysis of every component from the input of power or motion to the final output. In this aspect, designers must recognize that energy seeks the path of least resistance. A good designer strives to achieve the ideal of uniform resistance to failure for every component.

In determining component size, the designer must thoroughly understand all the forces that are applied to or absorbed by individual component parts and the entire assembly. A knowledge of strength-of-materials principles is essential for designers at this point. They should be fully aware of the differences between simple and compound stresses. Designs using simple stresses (tension, compression, and shear) are preferred to designs having compound stresses (bending, buckling, and torsion). Designers recognize that it may be impossible to eliminate all bending action but that there are more efficient ways of designing bending members. They are aware, however, that torsion members are more efficient with symmetrical cross sections. Finally, they recognize the importance of column ends to increase the resistance to buckling.

Most service loads are at best statistical averages. Consider, for example, a motor-driven machine tool. Power input varies because of variations in voltage and amperage. The belt drive may slip and wear. The tools wear. The raw material varies in size, shape, hardness, and toughness. And, of course, no two machine operators work in exactly the same manner. In self-propelled machines, the variations are even greater due to impact loads that at times may exceed the force of gravity. The designer must consider all of these factors.

B. Choice of Power Actuation

Perhaps the foremost design consideration is the choice of power actuation. The designer has several options, each with advantages and disadvantages for any particular product. The final choice may be the result of personal preference. The usual choices are (1) human muscle, (2) electrical, (3) pneumatics, (4) hydraulic, and (5) internal combustion.

After deciding on the power source, the designer carefully traces the path of the power from the point of input to the final point of output. Each component must be analyzed to determine the proper size to insure adequate strength for the specified loads.

The point of dissipating the power can be near or far from the point of power input, and power can be transmitted through few or many components. Just as it is true that a chain is no stronger than its weakest link, no machine is stronger than its weakest component. (Of course, unlike the links in a chain, some components may play more critical roles in a machine than do others.) In machine design, *every* component is scrutinized individually to ascertain whether it might fail.

C. Selection of Materials

Determining the sizes of components depends on geometrical shape and on the characteristics of the materials from which they are made. Steel is stronger than brass. Heat-treated steel is stronger than annealed steel. Cast iron is better for compression than tension members. Some materials are stiffer (that is, they have a higher modulus of elasticity) than others. Ability to dampen vibrations is not constant in all materials.

Selection of materials is facilitated somewhat by the information in engineering handbooks on the properties (strength, ductility, hardness, etc.) of various materials. In special cases, however, designers must work with metallurgical or manufacturing engineers to develop machine members for special service at minimum cost consistent with quality design.

Design formulas for evaluating sizes and shapes of machine members must take into account material properties. Of primary interest are properties indicative of material behavior under the loadings most frequently encountered in machinery. These properties are ultimate yield and fatigue strength; elastic and shear moduli; and range of ability to be hardened.

These properties give designers a basis for predicting material behavior. They are, however, data obtained under ideal laboratory conditions and therefore do not predict the effects of such factors as variations in size, corrosive atmosphere, high temperatures, and lack of homogeneity. Consequently, designers commonly face the problem of applying such data to situations that are considerably different from those in the laboratory.

To allow for lack of complete design information, design texts, handbooks, and engineering codes recommend use of design factors, variously known as safety, derating, or service factors. Design factors are the bases for transforming theoretical designs into practical designs. But designers are cautioned that when new machines are developed and failure may result in human injury, there is no substitute for prototype testing.

In machinery, metallic materials dominate, because under most circumstances:

1. They are rigid below, but plastic above, a certain stress (yield stress) and can therefore be shaped by rolling and forging operations.
2. They become liquid above their melting points and can thus be shaped by casting or joined by welding.
3. They are superior to nonmetals in ease with which they can be worked into a large variety of forms and shapes.
4. They are more resistant to wear than nonmetals.

Among metallic materials, steel dominates because of its versatility, great strength, and low cost. Low carbon steel dominates the steel group.

A major drawback to the use of metallic materials in machinery is their tendency to fatigue and corrode. Fatigue is the phenomenon that progressively leads to fracture under cyclic loadings. Corrosion is the alteration of materials, originating at the surface, caused by unintentional chemical or electrochemical attack.

The use of nonmetallic materials in machinery is growing. Their use is generally limited to lightly stressed parts. The real importance of plastics lies in their ability to be shaped by noncutting processes. This means that complex machine components can be shaped economically.

D. Friction, Wear, and Lubrication

The soaring cost of natural energy has forced designers, in the early stages of design, to pay closer attention to friction and wear. Wear represents one of the most significant operating costs of machinery. Every year thousands of otherwise functional machines are taken out of service because minute amounts of material have been worn from vital contact surfaces.

Wear is the gradual and unintentional abrading of surfaces in contact as a result of relative motion between them. Irregularities and asperities found on all surfaces roughen and abrade machine members as they move relative to each other under pressure. The process is accelerated as material is dislodged and serves as a grinding

agent. As the action continues, excessive clearances are generated, and precision parts are damaged as they ingest the tramp metal circulating through the system. The end result is a loss of efficiency and a buildup of heat, followed by breakage of parts or disabling of the entire machine.

Designers seek to control wear by specifying maximum limits of surface roughness in the manufactured parts. Materials should be compatible, which usually means that they will be dissimilar—cast iron against steel, for instance. Surfaces are specified to be hardened either by heat-treating the entire part or by some method of hard facing. Finally, there must be an adequate lubrication system. Wear is greatly reduced by keeping sliding metals apart through the use of a lubricant. Mineral oil has the desired characteristics of maintaining a strong separating film while offering minimum resistance to shearing action.

The lubricating system is an important aspect of the design of all machines because they have moving parts. To be effective the lubricating system must have several essential features: The lubricant should be positively directed to the rubbing surfaces, there is a means of filtering out the contaminants, and there is sufficient volume of oil to prevent its rapid deterioration. Suitable gaskets and seals are provided to prevent leakage. A cooler may also be required to stabilize temperatures.

E. The Importance of Weight Reduction

If the weight of a highway truck is reduced by one pound, it can carry an additional $100 in freight annually. Similarly, designers have good reason to lessen the weight of materials in machines, particularly those that are self-propelled. However, any attempt to save material nearly always results in a more complicated shape for the component concerned, with subsequent greater cost of labor. Material and labor costs are, as a rule, inversely proportional to each other. The optimum point for any design occurs where the sum of the two is at a minimum (Fig. 5). Of course this is only one of several criteria for designing machinery.

F. Methods of Assembling and Joining

An assembly is a combination of two or more parts, joined by any of several different methods, that facilitates the production of a larger assembly. All assembly operations follow a similar pattern: detailed parts are built into subassemblies; the subassemblies are assembled into major assemblies, and these into the finished product.

Theoretically, machines could be made from a few complex parts. However, because of the intricate shapes such designs would require, many parts cannot technically or economically be made in one piece. Figure 6 shows two

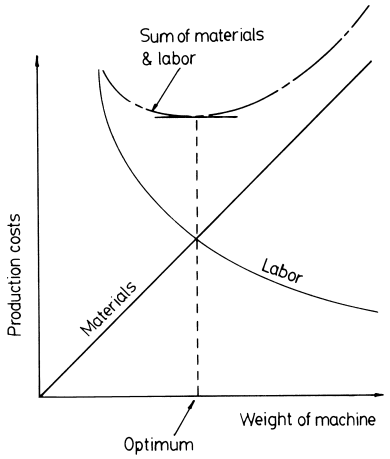


FIGURE 5 Determining the optimum weight of a machine from the cost of materials and labor. [Reprinted with permission from Leyer, A. (1973). "Machine Design." Copyright 1973 Birkhäuser Verlag AG.]

distinct parts from a tractor: a rear axle housing and a differential housing joined by means of cap screws. Figure 7 shows the tractor rear axle housing partially removed from the differential housing, exposing internal parts, thus facilitating inspection, repair, or replacement.

Assemblies may be made one way in one plant and another way in another plant. Established plant practices, available equipment and facilities, tooling, fixture design, production volume, and labor cost account for the

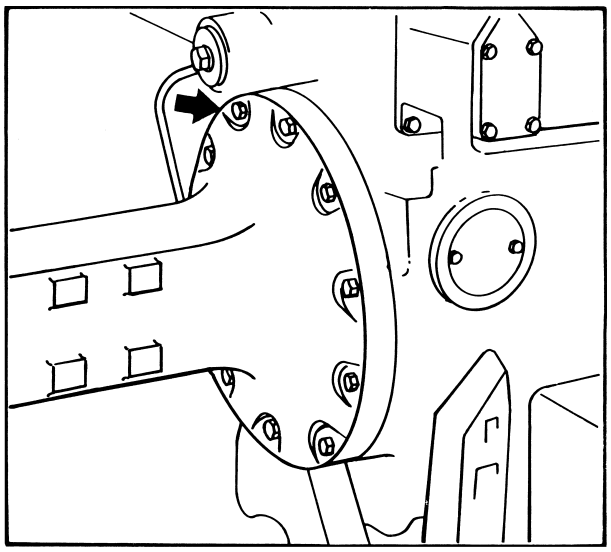


FIGURE 6 Tractor axle housing fastened to the differential housing with capscrews. Abrupt changes in form from a boxlike housing to neck-like housing prohibit one-piece construction. [Reprinted with permission from John Deere Service Publications. Copyright 1974.]

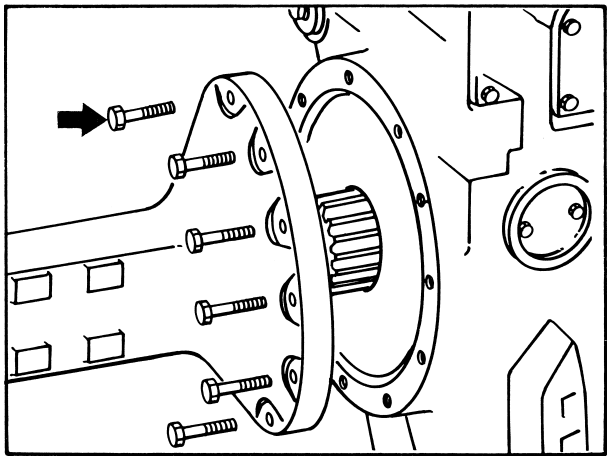


FIGURE 7 Detachable fasteners facilitate inspection, repair, or replacement of machine members. [Reprinted with permission from John Deere Service Publications. Copyright 1974.]

differences. Because there is no general solution to an assembly problem, designers must consider several aspects, including production volume, quality, and ease of service.

Design for assembly (DFA) is a design activity that can significantly reduce manufacturing costs and labor requirements. The most effective principles of DFA in achieving such reductions include the following:

1. Minimize the total number of parts required in a design assembly
2. Minimize the number of part variations (e.g., the number of different fasteners that are used in a design)
3. Design parts to be multifunctional (e.g., a single screw might be used to both fasten and align a set of components)
4. Avoid separate fasteners (e.g., molded snap fits might replace screws in an assembly, or self-tapping screws could be used to eliminate the need for separate washers and nuts)
5. Minimize the number of assembly operations
6. Maximize tolerances between parts for easy assembly
7. Provide sufficient and unobstructed access for easy assembly

Depending upon the nature of the product, reductions up to 40% in manufacturing costs are possible, as are increases of over 100% in productivity. Impressive savings such as these are possible because more than 50% of total production time can be spent on assembly or fastening.

1. Rigid Connections

Individual parts in machinery are held together by devices known as fastenings,most of which are detachable. They vary in kind and use from bolts and machine screws to keys, splines, and rivets for machines varying in size from airplanes to watches. Designers use detachable fasteners for a number of reasons, such as to simplify manufacturing and repair and to insure the safe and accurate assembly of critical parts. Threaded fasteners predominate because of their excellent cost efficiency.

Equally important to assembly are the methods of joining parts permanently by fusion or alloying. Actually, machine members, or parts from which machine members are made, can be joined in four distinct ways: by molecular process, interlocking, friction, and interference. Occasionally a combination of methods is used.

2. Molecular Processes

These are methods of fusion or alloying that occur at the molecular level. They include welding, soldering, brazing, and adhesive bonding. In each case the bonding agent becomes an integral part of the structure and does not function separately (Table II).

Parts joined by the molecular processes are very similar in appearance, and the geometric shapes will also be similar. Parts assembled by brazing, soldering, and adhesive bonding usually require that surfaces be overlapped. Joining also takes place at much lower temperatures. Each process owes its existence to cost-effective advantages over other methods in specific applications. In general, welding is the most economical method of joining the low-carbon steels that predominate in machine design.

The popularity of arc welding lies in its ability to create, inexpensively, a butt joint that has a strength equal

TABLE II Comparison of the Main Characteristics of Joining

Method of joining	Materials to be joined	Temperature required
Arc welding	Identical or similar base metals	Melting temperature of the base metal
Pressure welding	Identical or similar base metals	Depends on method
Brazing and soldering	Identical, similar, or dissimilar base metals	Melting temperature of the filler metal (base metal must remain solid)
Adhesive bonding	Any solid metal to any other solid metal; or solid metals to nonmetallic materials (plastics, high-strength fibers)	Depends on bonding agent; may require pressure

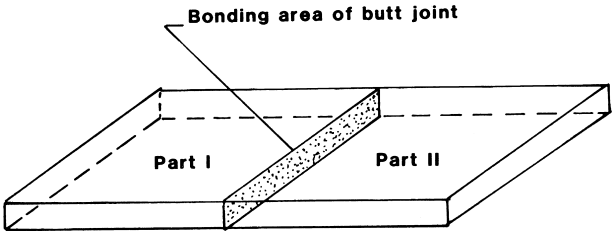


FIGURE 8 Butt joints for flat parts are made at the edges.

to that of the adjacent base materials. As Fig. 8 shows, the butt joint becomes an integral part of two plates and provides a relatively smooth surface with considerable savings in weight, compared with the lap joint shown in Fig. 9. Figure 10 shows a comparison of welded and riveted structures.

Many phases of welding are important to machine design, but the initial problem is the determination of the size of welds necessary for a given part. This requires stress analysis of parts of the design that are subject to either static or fluctuating loads. Especially critical is the reaction of the assembly to dynamic loads, which can be particularly rough on metals whose strength, hardness, and effective thickness may decline because of the complex reactions (thermal and metallurgical) involved in welding.

3. Interlocking

Joining by interlocking of projections has two subdivisions:

1. Interlocking of rigid members such as pins, keys, retaining rings, and splines that engage with suitable matching recesses (Fig. 11)
2. Plastic deformation followed by interlocking, as in riveting (Fig. 12); rivets are primarily shear devices, and very cost efficient

4. Friction

This method involves the application of large normal forces to matching surfaces that then generate frictional forces tangential at the mating surfaces. Typical joints based on friction include setscrews, tapered shafts, and concentric, tapered rings. Bolted joints loaded transversely depend on the friction that is developed by clamping to prevent relative motion (Fig. 13).

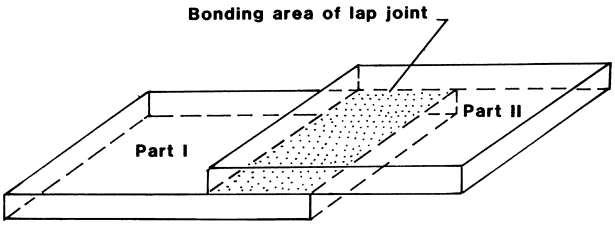


FIGURE 9 Lap joints of flat parts are made at the sides. This technique provides a larger bonding area.

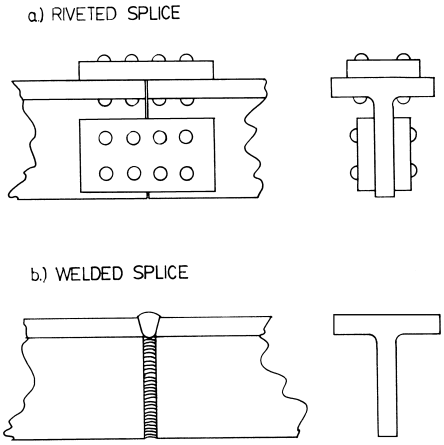


FIGURE 10 Comparison of welded and riveted structural joints. Joining two parts by riveting requires that appropriately spaced holes be made in the parts to be joined; as shown in (a), these holes reduce the cross-sectional areas of the members by up to 10%. The joint may also require the use of one or more gusset plates, which adds to the weight and cost; as (b) illustrates, welding can eliminate the added weight and cost.

5. Interference Fits

Fits between cylindrical parts govern the proper assembly and performance of countless rotating machine parts. For parts meant to be permanently assembled, interference fits create the necessary tightness. No fastening elements are used; instead, the inherent elasticity of the materials is used to secure a rigid connection (Fig. 14).

G. Methods and Means of Power Transmission

Because machines are designed to transform, transmit, and control power, machine designers devote much of their effort to selecting and designing the components directly responsible for transmission and control of power. Indeed, the components are the largest single group of machine elements, a group that includes couplings; belt, chain, and gear drives; shafts; bearings; clutches; and brakes.

In machinery, power can be transmitted in four fundamentally different ways; mechanically, hydraulically, pneumatically, and electrically. Figure 15 shows principal means of power transmission. Power can be transmitted mechanically in three basic modes: (1) by a flexible wrapping connector, such as a belt, rope cable, or chain; (2) by direct contact as by cams, gears, or friction wheels; and

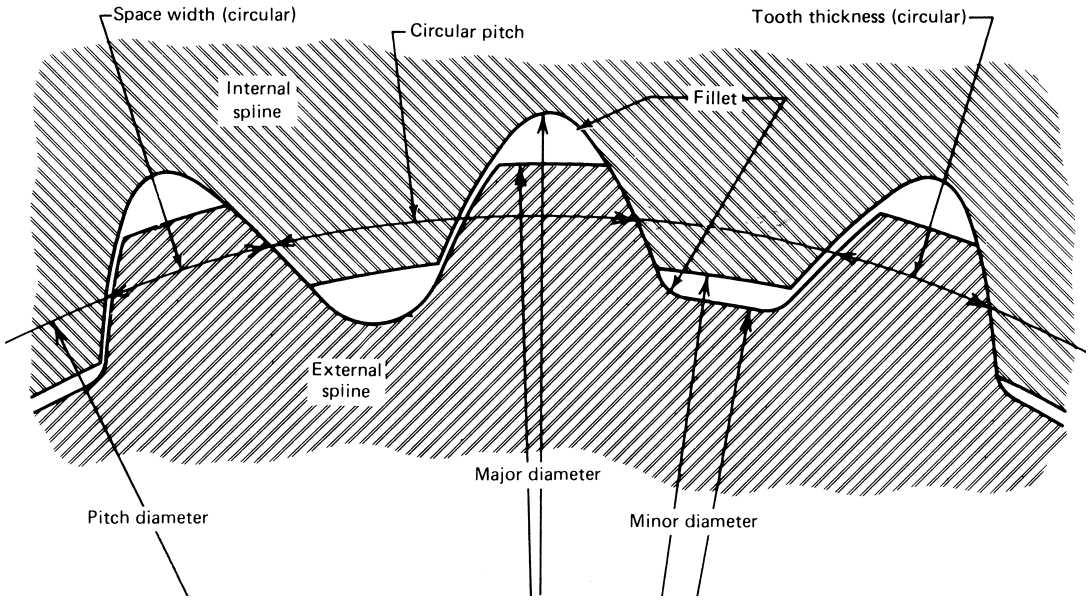


FIGURE 11 Mechanical interlocking.

(3) by a rigid connecting link or coupler (Fig. 16). In every case the mechanism has a driver and a follower, with power input at the driver and output at the follower.

Before a designer can select a drive, however, the following basic data should be ascertained:

1. Type of power source (motor or engine)
2. Amount of power required (input and output)
3. Type of operation (continuous or intermittent)
4. Operating conditions (start and load duration)
5. Magnitude of speed (input and output)
6. Speed modification (output to input)
7. Environmental conditions (ambient temperature, corrosive elements, etc.)

Transmitting motion and power is accomplished by the generation and absorption of forces. Thus the magnitude and manner of application of forces are basic to machine design. By controlling these variables, designers control speed and work as well as power. Time is a third variable with which designers contend. Unfortunately, they are unable to control this variable; they can only use it as a basic reference.

Mechanical means of power transmission are generally associated with shaft-mounted rotating members such as belts, chains, gears, and couplings. The choice of machine member depends first on the relative shaft position; but speed, torque, and shaft distance are also important (Fig. 17). Figure 15 showed a schematic outline of the choices available to designers, and Fig. 18 gives the approximate breakdown based on industrial usage. Note the predominance of couplings and gears.

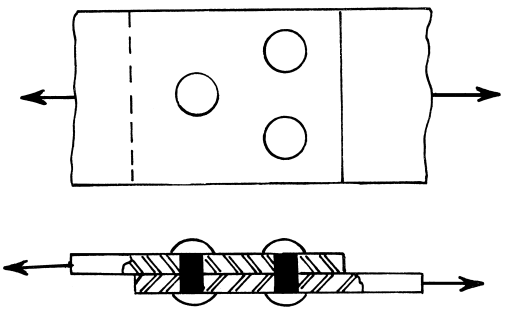


FIGURE 12 The typical riveted joint has three or more rivets loaded in shear. The rivets are clinched so that no relative motion takes place as long as the allowable shear stress is not exceeded.

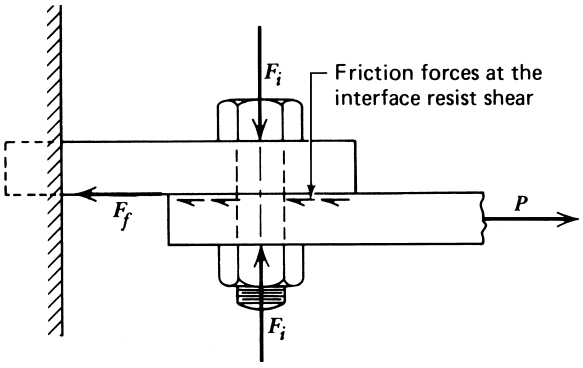


FIGURE 13 Transverse loading. [Reprinted with permission from Hindhede, U. *et al.* (1983). "Machine Design Fundamentals." Copyright 1983 John Wiley & Sons, Inc.]

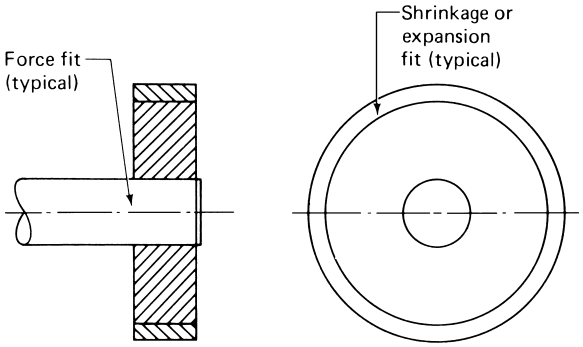


FIGURE 14 Interference fits. [Reprinted with permission from Hindhede, U. *et al.* (1983). "Machine Design Fundamentals." Copyright 1983 John Wiley & Sons, Inc.]

When motor speed is the same as shaft speed, a simple flexible coupling is often used. For speed reduction and smooth flow of power, however, gears, chains, or belts are needed. Gears predominate because of their versatility, compactness, and high efficiency. Unlike most belts and chains, gears (beveled and worm) permit change in the plane of rotation, a characteristic that greatly increases their usefulness. Gears provide power transmission over a wide range of speeds but are generally used for short center distances, such as those found in vehicle transmissions.

When shaft distances are large enough to render gears impractical, chains or belts are used. Drive chains, like those on motorcycles, reduce overall weight; but they wear and elongate. Wear can be minimized with adequate lubrication and the use of protective housings to keep out abrasive material. A high-torque capacity at low speeds is a major advantage of chains.

For higher speeds and lower torques, belts are preferred. Belts do not need lubrication but require a means of adjustment to prevent excessive slippage. The lower power capacity and higher speeds of belts, compared with chains, are due to their reliance on frictional forces to transmit power. Both belts and chains are often used in a speed reduction system. A product division exists for speed reduction systems in that belts are used for the first step, characterized by high speed and low torque; the second step uses a chain because of reduced speed and increased torque.

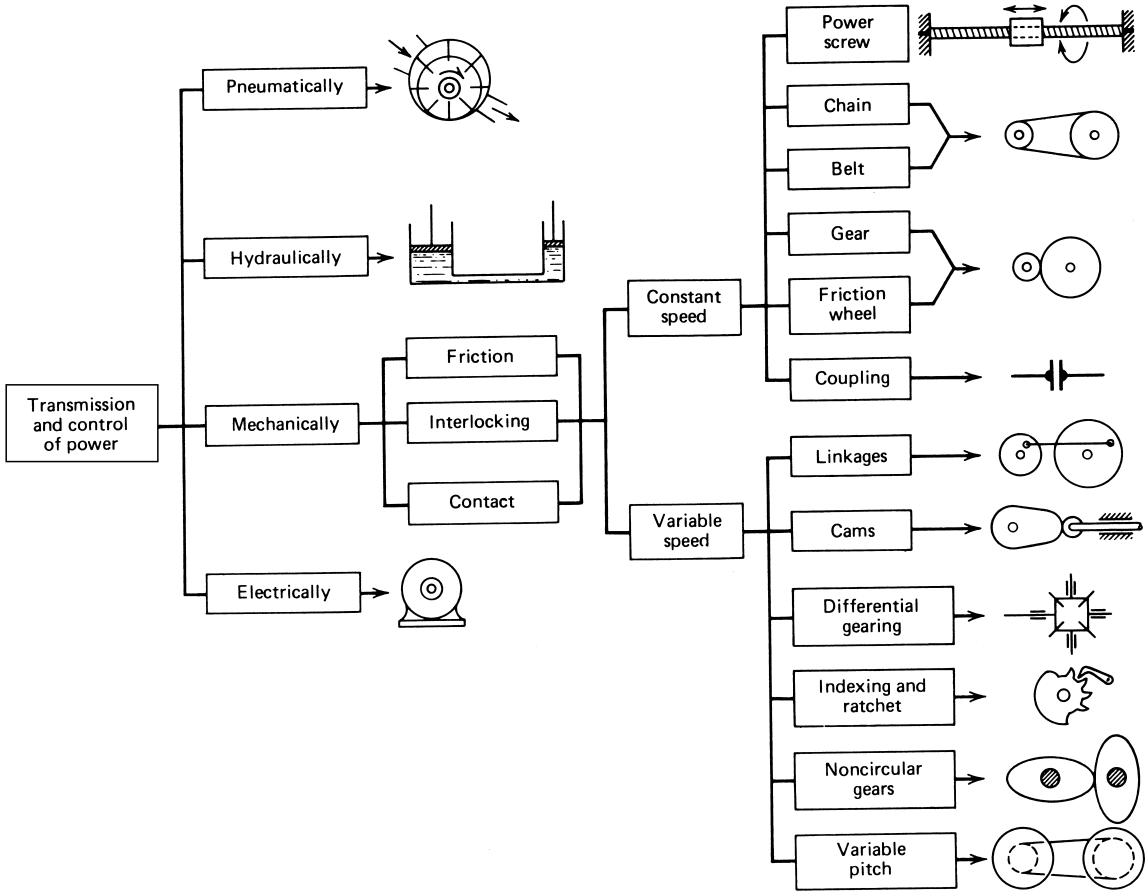


FIGURE 15 Principal means of power transmission. [Reprinted with permission from Hindhede, U. *et al.* (1983). "Machine Design Fundamentals." Copyright 1983 John Wiley & Sons, Inc.]

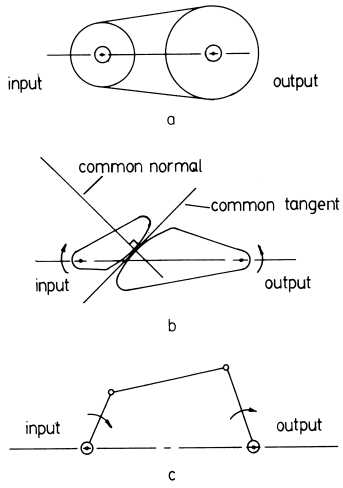


FIGURE 16 Methods of transmitting motion and power: (a) flexible-wrapping connectors, (b) direct-contact bodies, and (c) linkage.

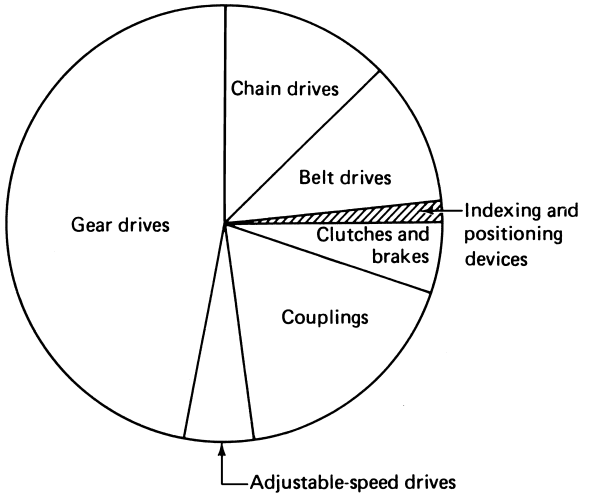


FIGURE 18 Use of mechanical components in industry. The chart shows the approximate breakdown based on sales. [Reprinted with permission from Hindhede, U. *et al.* (1983). "Machine Design Fundamentals." Copyright 1983 John Wiley & Sons, Inc.]

Designers of machinery use a fluid drive in preference to mechanical systems when (1) power is to be transmitted between points too far apart for chains and belts; (2) high torque at low speed is required; (3) a very compact unit is needed; (4) a smooth transmission, free of vibrations, is required; (5) easy control of speed and direction is necessary; or (6) output speed must be varied without stepping.

Fluid power is the use of a fluid, which can be air or an oil, under controlled conditions to perform useful work.

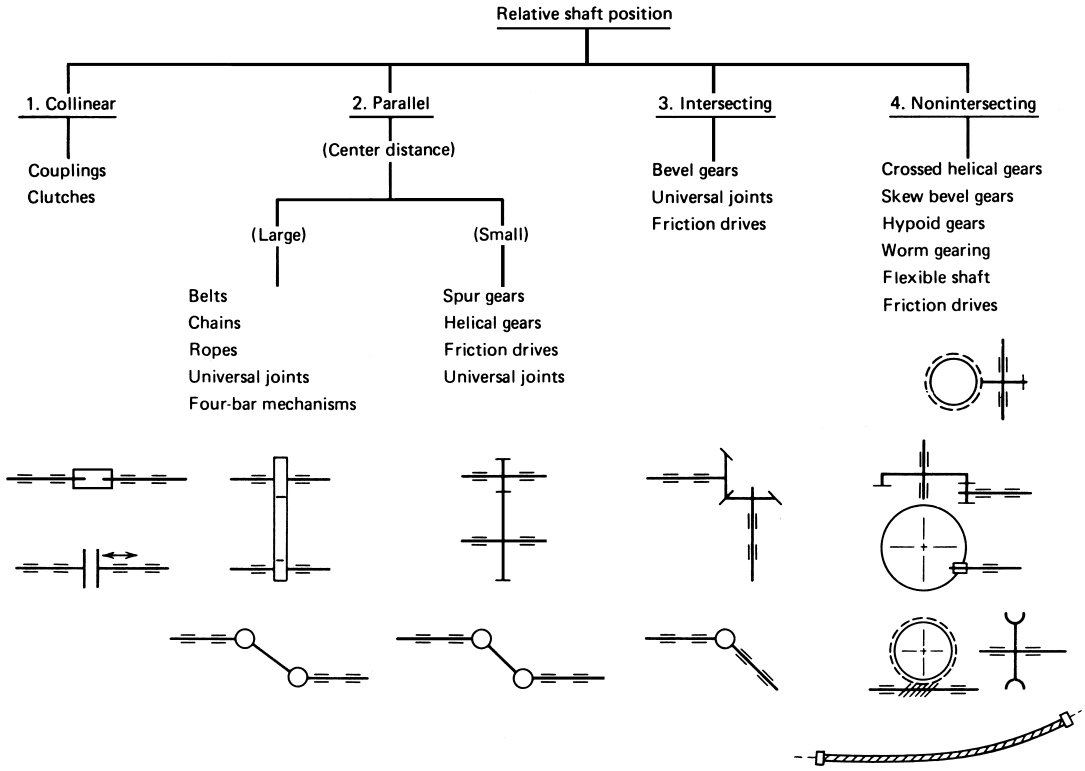


FIGURE 17 Common mechanical means of power transmission between machine shafts. [Reprinted with permission from Hindhede, U. *et al.* (1983). "Machine Design Fundamentals." Copyright 1983 John Wiley & Sons, Inc.]

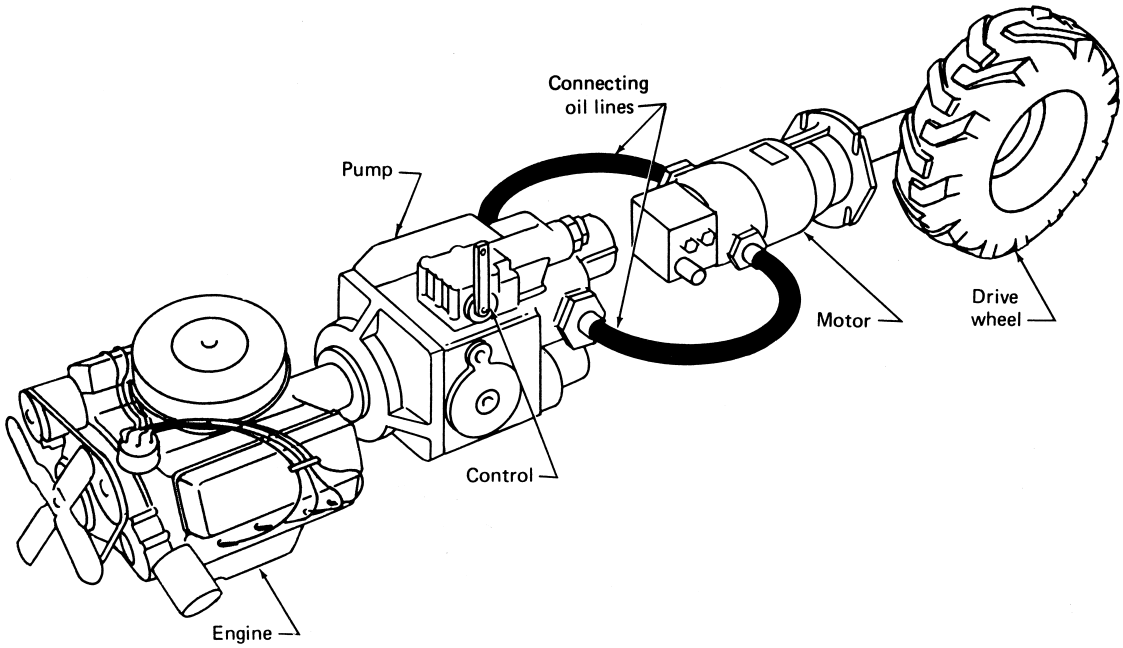


FIGURE 19 Hydrostatic drive in a complete power train. [Reprinted with permission from John Deere Service Publications. Copyright 1972.]

In machinery, oil or hydraulic fluids predominate because their noncompressibility enhances the efficiency of the system. In fluid power systems, fluids are moved under pressure by pumping to transfer the maximum amount of energy with a minimum amount of fluid. Power can be transmitted hydrostatically and hydrodynamically. Such systems are called hydraulic systems.

A hydrostatic transmission using oil fluid contains a pump, motor, and controls in one unit. It provides stepless variation of speed, torque, and power in addition to smooth and controllable acceleration, instantaneous reversal, ability to be stalled without damage, and easy control (Fig. 19). If a linear actuator (power cylinder) is used instead of a motor, the output becomes a powerful push-pull device.

Designers of machinery do not usually design hydraulic components; instead they synthesize a system with purchased components. To do so they must have knowledge of hydraulics.

Hydrodynamic systems use the kinetic energy in fast-flowing oil fluids to transmit power. A fluid body accelerates by absorbing kinetic energy from some external source of energy. When the same body decelerates, it does so by giving off energy to some external energy-absorption unit. This is the basis of fluid couplings.

A fluid coupling is really a clutch-coupling (Fig. 20). At low speed and during acceleration, it is a coupling that transmits torque unchanged but with some loss of speed. In this respect, it resembles a centrifugal clutch, but it operates more smoothly and has the ability to transmit torque while blocking transfer of harmful shocks and vibrations. These unmatched operational qualities are the reason that fluid couplings are used extensively in machinery.

H. Frames and Housing

In nearly all machines there is one member that occupies a fixed position or carries the entire unit along with it during motion. This member is the frame. The frame is the backbone of the machine and keeps the component parts securely in proper alignment. It can be an effective vibration-dampening device. Also, the frame often serves as the means of containing the lubricant and excluding the external contaminants (Fig. 4).

In self-propelled vehicles, for example, the engine, transmission, and drive axle assemblies are usually attached to the frame, as well as to the front axle with its suspension system and steering mechanism. Frames in vehicles are subjected to bending, torsion, direct loads, and a combination of these loads. Internal forces generated by the acceleration and deceleration of the vehicle must be absorbed by the frame.

Design of frames has undergone a revolutionary change due to the finite element method, a way to determine the size and shape of components using the computer. Initially developed for the design of architectural structures, FEA can be applied to almost any structure and has become a significant design tool. If the frame is to be of

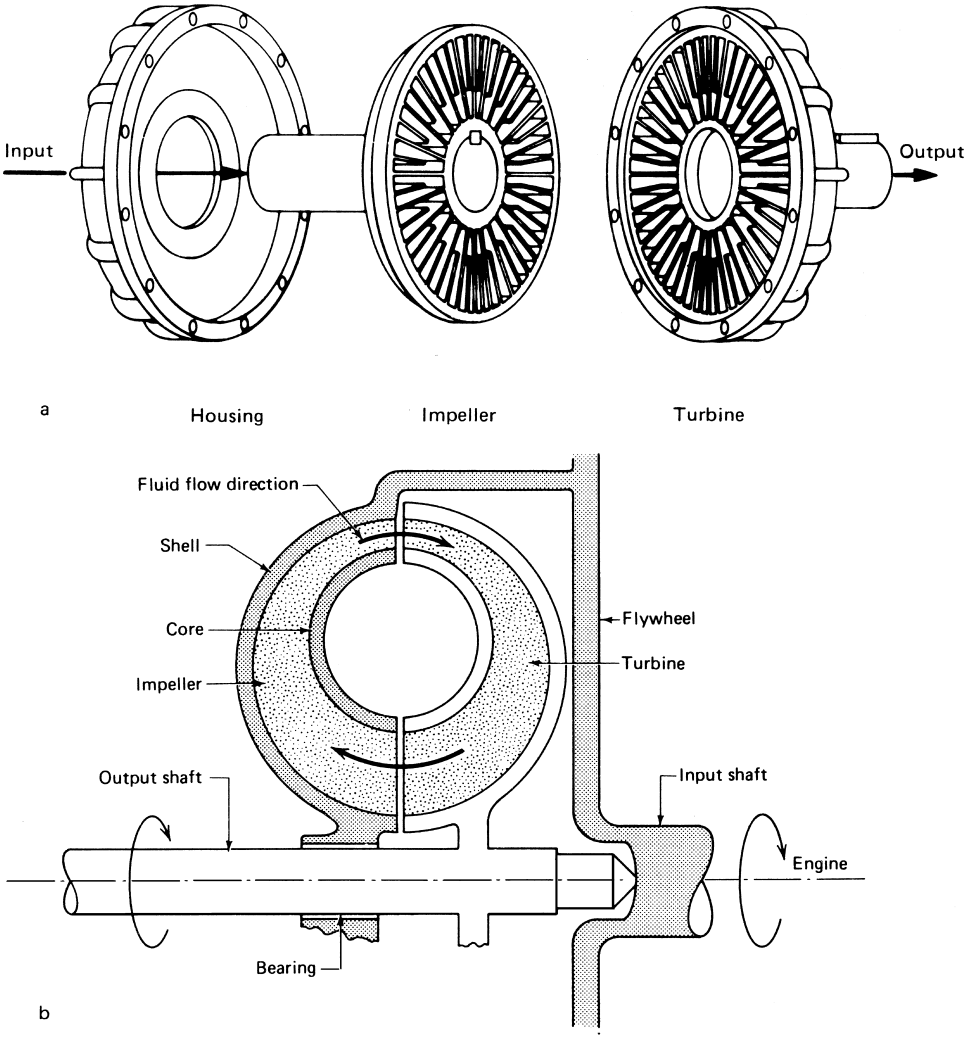


FIGURE 20 Fluid coupling. This kind of hydrodynamic drive transmits power without changing torque: (a) the three main parts of hydraulic coupling and (b) schematic outline of a fluid coupling. [Reprinted with permission from Hindhede, U. *et al.* (1983). "Machine Design Fundamentals." Copyright 1983 John Wiley & Sons, Inc.]

cast construction, a designer, aided by FEA software, can discretely place the material exactly where it is needed. This is an important capability. In the absence of such sophisticated design aids, the designer can be well served by full-scale testing with stress-coat and strain-gauge techniques.

I. Design for Human Factors

Consideration of human factors is especially important in the design of frames and housings. In the automotive industry, for example, frames must be designed to accommodate people as operators, passengers, and maintenance personnel.

The goal of human factors engineering (or "ergonomics") is to insure that people can interact with machines safely, comfortably, and efficiently. This means that designers must place controls where they will be accessible to the operator and comfortable to use. Hand tools should be light, free from excessive vibration, and safe to use. Concern for ergonomics is clearly evident in the design of space vehicles, which are literally built around the astronauts.

SEE ALSO THE FOLLOWING ARTICLES

AUTOMOBILE TECHNOLOGY • ELECTRIC MACHINES, DESIGN • FRACTURE AND FATIGUE • MANUFACTURING PROCESSES TECHNOLOGY • MECHANICS, CLASSICAL • SIMULATION, REALISTIC (ENGINEERING) • TRIBOLOGY

BIBLIOGRAPHY

Anderson, D. M. (1990). "Design for Manufacturability," CIM Press, Lafayette, CA.

Boothroyd, G., and Dewhurst, P. (1984). 'Design for assembly,' *In* "Machine Design," Pention/ICP, Inc., Cleveland, OH.

Ferguson, E. S. (1977). "The mind's eye: nonverbal thought in technology," *Science*, August 26.

Hindhede, U. Zimmerman, J. R., Hopkins, R. B., Erisman, R. J., Hull, W. C., and Lang, J. D. (1983). "Machine Design Fundamentals, A Practical Approach," John Wiley and Sons, New York.

Leyer, A. (1973). "Machine Design," Blackie & Sons, Glasgow, Scotland.

Norton, R. L. (2000). "Machine Design, An Integrated Approach," Second Edition, Prentice-Hall, Upper Saddle River, NJ.

Voland, G. (1999). "Engineering by Design," Addison-Wesley Longman, Reading, MA.



Manufacturing Processes Technology

H. C. Kazanas
Lowell P. Lerwick
University of Illinois

I. Manufacturing: Design, Materials, and Processes
II. Materials and Measurement Technology
III. Primary and Casting Processes
IV. Materials Removal Technology
V. Forming and Fabrication Processes Technology
VI. Miscellaneous Processes Technology
VII. New and Emerging Processes Technology

GLOSSARY

Automation Designing of machine tools or processes so that, once started, they will repeat or cycle their basic functions accurately, continuously, and automatically and do so without human assistance.

Computerize Equip with or operate, control, process, or produce by means of machines capable of operating on their own internal system of logic.

Cybernetics Analytical science dealing with the overall problem of communication and control within system; in manufacturing processes, usually thought of in terms of artificial intelligence regulating automation.

Fabrication Putting together of materials to make up parts, subassemblies, assemblies, or complete products, especially during the latter stages of manufacturing; usually involves the coordination of several processes such as handling, holding, and joining.

High technology An emerging area of applied theory and practice characterized by an emphasis on systematic interdisciplinary research and development as a substitute for a historical database; also called science-intensive technology.

Machine tool Power-driven tool specifically designed to execute predetermined manufacturing materials processing operations.

Manufacturing process Operation performed as part of the overall series of changes that alters, either partially or completely, a material from an ambiguous form or state to a predesigned configuration or condition.

Materials Physical element, elements, or substances of which machines, products, parts, goods, or stock items are composed or from which they are made.

Science of materials Multidisciplinary field devoted to the study of the nature, behavior, and use of materials for specific technological applications or purposes; sometimes called materials science, or materials science and engineering.

Technology Systematic application of scientific and engineering theory and knowledge to some practical purpose or activity.

MANUFACTURING processes technology is a generic reference to the technological methods by which raw materials are converted to useful or desirable products. The term denotes the industrial processes required to make manufactured products. Manufacturing is the overall process of coordinating human resources, machines, tools, and capital to convert preselected materials into preconceived products. As a productive enterprise, manufacturing represents one of the largest single segments of world industry and accounts for the largest percentage of the gross national product (GNP) of any single industry. There are thousands of different manufacturers producing many thousands of different products, including ships, airplanes, automobiles, machinery, appliances, clothes, and food. The number of industrial processes required to produce these products is astronomical. Defining and describing these processes would merit a multivolume series of texts. As a means of restricting the scope of the topic, this article will address only the processes that are basic to most modern and most central or important manufacturing industries.

I. MANUFACTURING: DESIGN, MATERIALS, AND PROCESSES

Originally, the word *manufacture* meant to make something by hand. The earliest manufacturing processes were done with simple hand tools for many centuries (see Fig. 1). The gradual accumulation of knowledge, derived from experience with these tools and other simple implements, slowly evolved into the process technology that made possible the transition from hand tools to machine tools. Just when this happened is not certain. Crude lathes were reportedly used to make Etruscan wooden bowls as early as 700 B.C. The date of the modern machine tool age is fixed as 1775, the year English-man John Wilkinson invented the first consistently accurate horizontal boring machine. It was the development of machine tools capable of producing themselves (and other machines) that ushered in the era of modern manufacturing and set the foundations of modern technology.

In early times, manufactured products were produced largely on the basis of an individual's perceived needs or wants. The suitability of the end product for its intended purpose was highly dependent on the materials found in nature, plus the creativity, ingenuity, and skills of the artisan attempting the task. Quality depended mostly on the worker's skill in fashioning and using relatively crude processing tools. This type of manufacturing posed serious limitations in terms of production output, diversity

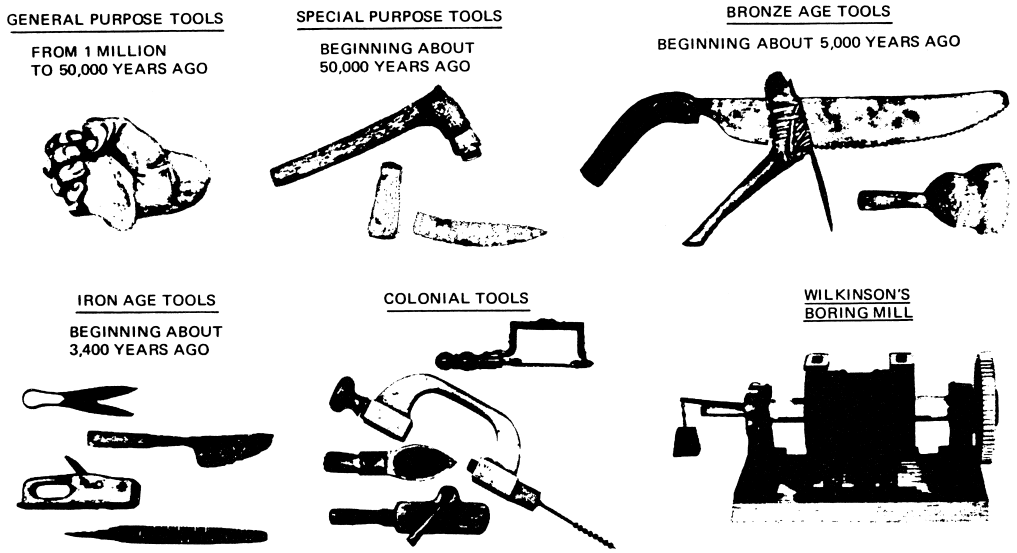


FIGURE 1 Evolution of hand tools resulting in the development of machine tools that led to modern manufacturing.

of product design, product cost, consistency of quality, and interchangeability of components.

Modern manufacturing methods and machines build part of the workers' skills into the production machines and processes themselves. This makes possible the employment of a greater number of people with relatively limited skills, while increasing output, reducing production cost, improving product design capabilities, and providing more reliable product quality. But modern manufacturing is also dependent on availability of materials. Product diversification and high production output invariably call for diversity in the materials used. Fortunately, modern materials science has done much to meet that challenge. But the advent of new and improved materials, most of them "man-made," has in turn required new and improved manufacturing processes. When special alloys were developed to meet aerospace needs, it was soon learned that existing processes could not be efficiently used on these alloys. Thus, new and radically different processes had to be developed. There is a direct relationship between the types of materials used and the processes required to produce a product for a specific function. This relationship is so close that it has become quite common for engineering schools and colleges to regard materials and processes as a single subject, mutually related in both theory and application.

Modern manufacturing is an industrial activity that requires many resources. For efficient, economical, and competitive production, all available resources must be carefully organized, coordinated, and controlled. The objective of this management activity is to create an integrated organization capable of successful production in terms of quality, profit, and share of the market. It is important, therefore, to understand that manufacturing processes, though specialized in nature, are an integral and important part of a larger system called the organization or the company. The nature of the organization and coordination required by manufacturing companies varies with the size of the company and the type of product or products being produced. But even though companies may differ organizationally, certain elements and functions are found in almost any manufacturing organization. Among these are ownership, general administration, sales and marketing, purchasing finance, personnel, product development, manufacturing, and quality control. Figure 2 provides a conceptual overview of the interrelationships between processes and other major manufacturing functions within a typical company.

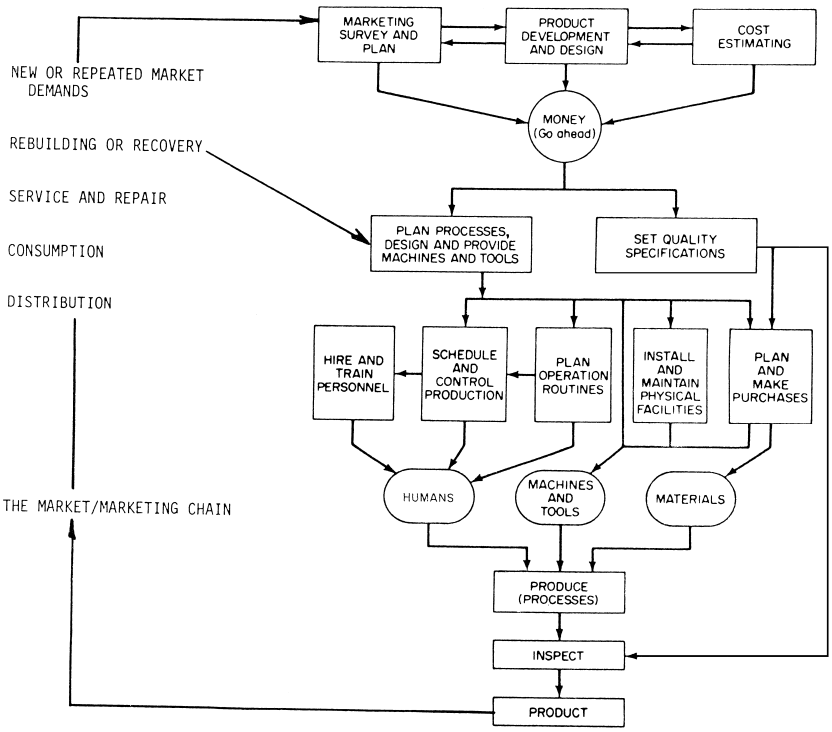


FIGURE 2 Functions that lead to efficient manufacturing. [From Doyle, L. E., Keyser, C. A., Leach, J. L., Schrader, G. F., and Singer, M. B. (1985). "Manufacturing Processes and Materials for Engineers," Prentice-Hall, Englewood Cliffs, New Jersey.]

Basic manufacturing involves four basic stages or steps, which can be defined as follows:

1. Product research and development or product design. This is where the concept of the product is formulated. It is also the stage where the basic "go" and "no go" decisions are made by management.
2. Production planning and tooling. Most of the technical planning to produce a product that will be competitive in the market is done at this stage. Cost estimating and process design are elements of the planning stage of manufacturing.
3. Manufacturing or production. Once the product has been designed and its production planned, the next step is to produce it.
4. Marketing. After the product has been manufactured, it must be marketed to make it available to the customer. The marketing stage will not be discussed in this article, which is concerned with the processing of the product.

These four stages are organized in different ways in industry, depending on many factors. Nevertheless, they represent the four logical areas of manufacturing a product, from its conception through its delivery to the consumer. As previously stated, the market/marketing chain is usually considered to be beyond the scope of manufacturing and processing per se. Yet it is important to note that this chain presents the company with a host of variables over which it has no direct control. Although the chain leads back to the company's marketing survey and product development teams, most of the factors are subject to the whims of the consumer, the strength of the competition, and the winds of fortune in the macroeconomy. To compensate for the uncontrolled variables of the outside market, internal manufacturing functions must be carefully planned. Every process must be evaluated for cost efficiency; every feature of the product must reflect deliberate design. Everything that can be controlled between engineered parameters must be held to tolerance. Ideally, nothing is left to chance. The challenge is to satisfy the market at minimum cost and maximum profit. Most of this planning and design is done by engineers, and almost all engineers are concerned with some aspect of the planning/design challenge.

II. MATERIALS AND MEASUREMENT TECHNOLOGY

Technology can be defined as the application of scientific knowledge to practical problems. When scientific knowledge of materials is combined with modern processing techniques the result can be termed manufacturing (industrial) processes technology. When scientific knowledge is focused specifically on the practical applications of materials, it can be defined as industrial materials technology. Although the two terms are so closely related as to be interchangeable, this discussion of materials and measurement will focus on materials technology. Under subsequent subheadings the processes technology perspective will be reinstated.

Selecting a material from which to make a product was not always done as it is today. In primitive times, choices were limited and therefore simple. Finished products were more nearly expressions of art than of science. For example, one made from wood whatever could be "drawn" by native creativity from the natural medium. Such "art" (craft) could be extremely practical and utilitarian, but it was seldom supported by a scientific understanding of the materials being processed (although it must be conceded that sound engineering concepts have been in effect for some time). Until fairly recently, materials selection was hampered by broad gaps in knowledge. The development of materials is in itself an interesting field of study. For the purpose of this chapter, however, Table I provides an overview of materials development.

The approach to materials selection as practiced in industry now is quite scientific. This change has come about as a result of progress in two major fields. First, advances in understanding service requirements permit accurate prediction of the conditions to be encountered by the material. Second, a great deal has been learned in recent years regarding the internal elemental atomic structure of materials.

Because of modern advances, design engineers are able to choose with great accuracy the one material that will best suit a need. Moreover, as a result of the new knowledge of materials structures, it is often possible to produce a material when no ideal one is already available to fill a specific need. This ability to control the structures of materials to produce or build in desired properties has led to the "materials revolution." This revolution relies heavily on the collection and use of data that are measurably significant. Determinations of the theoretical relevance and importance of such data have fallen within the newly emerged field of reliability engineering. The practical application of reliability engineering principles has, in turn, been functionally designated as quality control. The prime responsibility of quality control is to ensure that the end product meets engineering expectations.

It should be noted that quality levels are a matter of engineering rather than arbitrary moral judgment. Measured data are "good" if they fall within the parameters predetermined as acceptable by the marketing and product engineers and their supporting reliability engineers and scientists. Measurements are "bad," and require

TABLE I A Brief History of Materials and Technology

Time	Materials evolution	Level of technology
Prehistoric	Materials taken directly from animals, birds, vegetable and marine life, or the earth and sea. Natural resins and oils.	Technology of the nomad/hunter. *Ad hoc* use of available materials as tools to process other tools. Fire used to process food.
New Stone Age (8000 B.C.)	Survival products. First human-designed products of clay, gold, silver, copper, shells, bone. Refashioning of materials begins: pottery, jewelry, armor, weapons, and tools. First manufactured materials: cloth, papyrus, paper, beer, and wine. Timber converted to wood products.	Flintstone technology—stones made into tools. Water, fire, and controlled heat used to make pottery. Melting discovered. Agriculture—domestication of plants and animals as a source of food, power, and products. Metal tools appear.
(3500 B.C.)	Meteoric iron discovered.	
Bronze Age (3000 B.C.)	Lead, tin, zinc, bronze, brass. Metallic alloys in common use. Value of gemstones recognized. Value of materials as fuels noted.	Smelting and alloying methods improved. The wheel and axle, lever, inclined plane, wedge, and screw used as power-making or power-saving mechanisms.
Iron Age (1600 B.C.)	Iron as a war material Iron replaces bronze as metal of convenience.	Hittite Empire mining, ironmaking, and metalsmithing.
(1000 B.C.)	Materials categorized by type and kind. Damascus steel.	
Meridian	Roman cement, limestone, silica.	Greek natural philosophy, science, mathematics, and geometry. Earth, water, fire, and air as the basic elements. Roman engineering.
Medieval (A.D. 476–1450)	Textiles, cut stone, glass, cast iron, gunpowder, linseed-oil ink, carbon, silicon, sulfur, etc.	Dark Ages paradox, alchemy and metallurgy. Renaissance of physics, science, and technology.
Enlightenment (A.D. 1700)	Elemental nature of matter recognized.	Chemistry becomes an applied science. Beginning of the industrial revolution. Machine tools appear in great numbers.
19th century	Forty of the elements being regularly used to make industrial materials.	Existing science and engineering applied as manufacturing technology.
1868	First synthetic plastics: cellulosics.	
1909	Phenolics.	
1926+	Alkyds, vinyls, styrenes, aminos, polyethylene, polyester, epoxies, polypropylene, etc. Moden plastics, textiles, crystals, ceramics, concrete, and composites.	Einstein and the beginning of the materials revolution. Computers and cybernetics.

corrective action, if they violate these established spheres of tolerance. Obviously, a product that falls below design standards for function or performance runs the risk of customer rejection and failure in the marketplace. It is not always as obvious that arbitrary attempts to exceed the expectations established for the product can result in serious manufacturing cost overruns, thereby damaging a company's economic visability.

The final choice of a material rests on a combination of properties, economics, and overall acceptability; but first, the exact requirements and conditions of service must be determined. Performance data from similar existing products often provide the base data. Stress analysis techniques and other physical experiments can be used to calculate the types and magnitudes of factors affecting a new design. Models, mock-ups, test runs and computer simulations can be used to analyze the range of service conditions to be encountered. There are certain requirements for the proper selection of a material for a specific application. Designers usually consider the following points:

1. Integrity of shape. If a material is to be used under load, its shape must remain within design tolerances.
2. Strength. The material must resist breaking under normally anticipated service loads.
3. Service life. The material must perform its function for some predictable period of time based on normal care and frequency of use.
4. Special functions. A material may have to perform a unique function or service in respect to electrical, thermal, or magnetic effects. Unique service requirements often limit the list of usable materials to those with special properties. Some examples are materials subjected to ultrasonic, microwave, or X-ray in-service environments.
5. Integrity of structure. The internal structure of the material used must remain stable over time independent of anticipated temperature or other environmental factors.
6. Processability and formability. Materials must be transformed into the shapes and assemblies shown on the designer's blueprints. A company must have, or be able to acquire, the technology necessary to convert the material

into a manufactured product. Thus, process capability is an important factor in materials selection.

7. Cost. Materials selection should consider the overall as well as the apparent cost of a specific material. Apparent cost considers only the price of the material itself. Overall cost is the sum of the materials cost and the cost of processing. The overall cost of a finished product can often be reduced substantially by combining the proper selection of materials with an appropriate processing technology.

8. Environmental effects. No responsible contemporary manufacturer can escape considering the effect that a product, and the processes used to produce it, will have on the environment. Hence, designers and engineers must consider the ecological and ergonomic consequences of their decisions. Such problems as waste disposal and recycling of used materials present new challenges in the initial selection of materials.

The list of service need considerations comprises initial points that imply the need for much more specific data and information regarding the properties of the material per se. Those properties are the responses of the material itself, independent of size and shape, to specific mechanical and environmental loads or influences. Although many properties can be measured, only those that pertain to industrial applications are discussed here. They can be broadly classified as mechanical and characteristic (or physical) properties.

Mechanical properties refer to the response of a material to applied forces and loads. Basic measurements are in terms of tension, compression, shear, and torsion. Critical parameters are determined with reference to elastic behaviors and ultimate yield strength. Additional critical considerations are resistance to cyclic load and fatigue; impact strength; flow or "creep" resistance; service over time at temperature, usually called stress–rupture life; ductility under stress and strain; and relative hardness. Materials selection should also recognize the demands placed on a material by limitations to its 12 degrees of kinematic freedom. Materials in service can often be expected to experience positive and negative events along the classic x, y, and z planes.

In addition, rotating positive and negative forces may be encountered as roll, pitch, or yaw constraints. In normal dynamic service, a material will usually have to cope with the resolution of simultaneous moments and technical events. Therefore, a material's response to dynamic and static forces has important implications concerning its suitability for a particular application.

Tests of characteristic properties indicate how a material is affected by nonmechanical variables such as electric fields, thermal energy, or chemicals. Major thermal considerations are coefficients of thermal expansion, resistance to thermal stress or shock, and thermal conductivity. Specific heat characteristics can be important. Electrical conductivity and resistivity, dielectric constancy and strength, and chemical or corrosion resistance are often important. Density and optical properties must sometimes be considered. Usually, the more abstract properties related to appearance, safety, or human sensory and biological reactions to materials must be taken into account.

Under the discipline of the science of materials (or, more simply, materials science), engineers and scientists are attempting to understand the elemental nature of materials in an effort to develop materials tailored to service needs. Through the study of atoms by applied physics, chemistry, and biology, humans have come close to discovering the nature of matter, which is the basis of all materials. Most of the experimental data for this effort have been generated by materials testing processes. The two main purposes of materials testing are to (1) study the behavior of sample materials under specific conditions and (2) identify defects in materials and products. In general, tests (and testing procedures) used in manufacturing are classified as either destructive or nondestructive, based on the condition of the sample specimen after testing. The main purpose of destructive testing is to determine the ultimate properties of the material under study. Nondestructive tests are used primarily to determine whether there are probably defects in the materials or finished product. Nondestructive testing relies heavily on the data collected by destructive testing as criteria for margins of engineering safety. In addition to the data generated by materials testing, materials are monitored during their processing by manufacturing quality control personnel. This quality control function is usually referred to as process inspection. At the risk of straying too far afield, one must note that *all* of the data used in manufacturing materials technology depend on the existence of an enormous array of accurate theories, formulas, instruments, and measuring tools. No body of data is more reliable than the theories and techniques used to gather it.

Under the heading "materials and measurement," the importance of materials selection and the need for measurably accurate materials data have been stressed. The focus has been on materials rather than process technology. However, there is a crucial interrelationship between materials and processes that must be remembered. In the production of modern-day sophisticated end products, the way a material is processed can be as significant as the material itself. By their nature, manufacturing processes subject materials to loads, forces, elemental contamination, and technical abuses that can affect the integrity of the end product. Thus, information regarding the in-process

history of a specific material can be of major significance. In certain "high-tech"industries, the need for scientifically controlled processing capabilities is now supercritical. The construction of sterile manufacturing satellites in deep space has become more of a necessity and near-future reality than a fantasy.

III. PRIMARY AND CASTING PROCESSES

Manufacturing relies on a number of techniques and engineering principles to shape or form materials into designed products. Most products require some degree of casting, molding, cutting, forming, and eventual joining to become a complete end product. Figure 3 illustrates how these processes might be categorized.

It has become customary to divide the concept of manufacturing into subcategories referred to as primary and secondary processes. Primary processing includes all the methods and techniques necessary to convert raw material into standard industrial classifications in terms of sizes, shapes, lengths, and other predesignated stock inventory nomenclature. Examples are iron ore converted into steel sheets, bars, angles, and other stock engineering shapes; timber into dimensional lumber and laminated plywood sheets; and crude oil "cracked" into various semiprocessed, but standardized, liquid or gaseous states. Secondary processes traditionally mean those manufacturing processes which alter or convert standard industrial materials and goods into a finished consumer product. Thus, the distinction between primary and secondary processes appears to be rather clear-cut. Primary processing produces a product intended to be used as stock unfinished industrial material; secondary processing is the manufacturing of an unfinished stock material into a finished consumer product. Figure 4 shows the primary processes typical of the steelmaking industry.

However, much of the logic behind the distinction between primary and secondary processes seems rooted in the notion that raw materials possess a natural immutability. That is, raw materials have a set of natural properties that are not easily changed and therefore dictate the useful range of industrial purposes to which they can be put. The raw nature of a material tended to determine what kinds of finished products could be made from it. For example, iron ore, which had gone through primary processing into stock steel shapes, was considered "pure" even though it was known to contain traces of other elements (sulfur, phosphorus, silicon, and manganese being the most common). The presence of these "tramp elements" was regarded simply as a natural handicap on the range of viable product design applications.

With the advent of the materials revolution, the attitude toward materials processing has been reversed. Increasingly, functional service requirements, and not natural availability, dictate what a material must be and how it

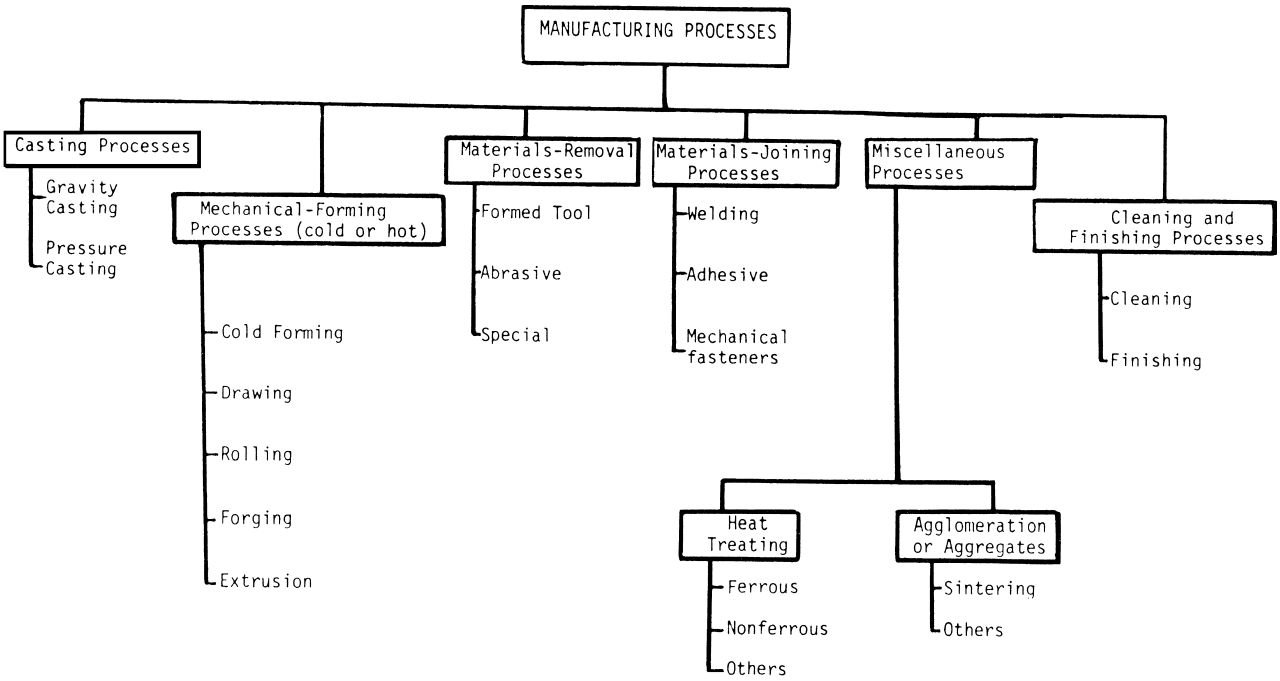


FIGURE 3 Classification of basic categories of manufacturing processes.

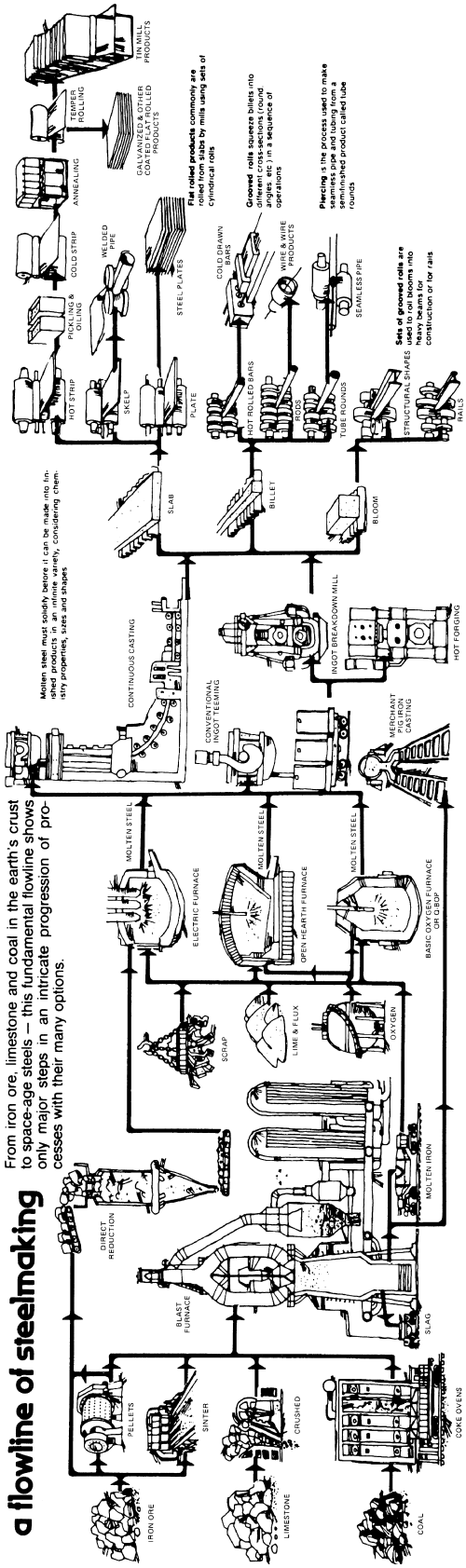


FIGURE 4 The primary processes of steelmaking. [Courtesy of the American Iron and Steel Institute.]

must be processed. The old standards for commercial purity have given way to new demands for scientific purity as the basis for materials engineering. Consequently, many of the so-called tramp elements must be processed out of the material during primary processing. Likewise, woods can be treated during primary processing in ways that meet specific service needs. Obviously, these new techniques are sophisticated and often expensive, but the point is that humans now have the technological ability to use them. As a result, the old clear distinction between primary and secondary processing has become unclear. The question becomes: Is an initial premanufacturing process, which is done to satisfy finished end product requirements, a primary or secondary process? In this case, the raw material is being processed toward a specific end function rather than a general materials inventory. It is in the area of modern synthetic materials that the issue truly becomes blurred. Petrochemical, chemical, ceramic, and plastic materials are prime examples of materials processed from a raw state with an end rather than an origin in mind.

To establish topical boundaries, the balance of this article will address the processes that have traditionally been defined as secondary manufacturing processes, that is, the conversion of stock industrial materials into marketable consumer products. However, the reader is reminded that these boundaries, from stock material to end product, are relative rather than absolute. In the past, manufacturing tended to be viewed as a linear series of operations that began with source materials and ended with the eventual disposal of the finished product as arbitrary scrap or waste materials. In the more contemporary "end consequences" view, manufacturing is increasingly regarded as a cyclical series of events in which waste materials are reprocessed into significant source materials. Thus, depending on the whole systems design criteria for a given product, it can sometimes be difficult to isolate primary from so-called secondary, product-related, processes.

Even under the traditional categories of primary and secondary processes, there is one classification of materials processing that fails to fall neatly into either one. This comprises the casting and molding processes. Cast or molded products can be supplied as stock industrial items requiring additional processing to meet end product requirements. That is, they must be converted from rough shapes to more precise configurations in terms of dimensions, angularity, geometric tolerances, design features, and so forth. But by using modern materials together with precision casting or molding techniques, it is also possible to produce items that accurately and completely conform to engineering service requirements. Consequently, cast and molded products are often regarded as being fully or semiprocessed items, irrespective of whether they were made in a primary or secondary setting.

Almost every office and home has numerous objects made by casting or molding. The typical automobile has a variety of castings of different materials made by different casting processes. Many different materials may be cast or molded, and the processes take a variety of forms, techniques, and variations. Examples range from the towering industrial machines cast from iron and steel to tiny toys made of molded plastics. Between these extremes are such products as ceramic vases, glass bottles, electronic components, and a multitude of metallic items, from gold rings and silverware to automobile parts, in an endless variety of metal alloys, ceramics, plastics, glass, concrete, and other materials.

Casting is one of the oldest processes known. Materials have been cast for over 5000 years. However, the ability to cast iron and steel has been limited, and the oldest known casting of an iron object is an ornamental lion cast in China in about A. D. 500. Casting is a process of giving shape to an object by pouring or forcing liquid material into a formed hole or cavity called the *mold* and allowing the liquid to solidify. As the material solidifies in the shaped cavity, it retains the desired shape. The mold is then removed, leaving the solid shaped object.

Molding is similar to casting. A shaped mold is used to impart the shape. However, the material is not a liquid but is only in a softened or "plastic" state. The material is forced into the mold under pressure and allowed to solidify. The mold is then removed to expose the shaped object.

To describe this complex array of processes, some system of organization is necessary. In Table II it can be seen that some processes can often be used to produce the same object. Also, the various processes can be categorized and classified in several ways: by the manner in which the materials are forced into the cavity (the two basic systems are gravity and pressure); by the nature of the mold material, which can be classified as either permanent or nonpermanent; and by reference to the materials being cast.

Processing companies that cast or mold metals are usually called foundries. Firms that process other materials such as plastics or ceramics are generally called either casting or molding shops. However, foundry, casting, or molding shops can also be categorized by their products. Firms that restrict their output to the mass production of one type or line of products are referred to as production shops. Other companies specialize in producing a variety of products in relatively limited quantities; these are commonly called job or jobbing shops.

Foundry and other casting shops are also described by how they market their products. A captive shop is one that is owned by a larger company and produces items principally for the company's own internal needs. An independent shop, by contrast, is usually separately owned

TABLE II Classification of Casting Processes[a]

Category	Process	Product material	Mold material	Mold class (unit or permanent)	Typical product
Gravity	Lost wax (investment)	Metals	Sand or plaster	Unit	Statues, turbine parts, transmission impeller
	Pit molding	Iron and steel	Green sand	Unit	Large industrial machines, turbines, and generators
	Floor molding	Iron and steel	Green sand	Unit	Medium to large industrial equipment, pumps, engines
	Bench molding	Iron and steel nonferrous metals	Green sand	Unit	Engine parts, plaques, hardware
	Shell molding	Ferrous and nonferrous metals	Cured sand	Unit	Semiprecision cast parts—gears, engine parts
	Ceramics molding	Ferrous and nonferrous metals	Cured sand	Unit	Precision parts, dies, rotor blades
	Rotational molding	Plastics, concrete, metals	Metal	Permanent	Pipe, trash containers, balls, toys
	Pouring	Concrete	Steel or wood	(1) Unit	(1) Building construction
				(2) Permanent	(2) Pipe, blocks, steps
	Slush and slip molding	(1) Nonferrous metals	(1) Iron and steel	Permanent	(1) Toys, small figures, and statuary
		(2) Ceramics	(2) Plaster		(2) Vases, trays, dishes, etc.
	Dip molding	Plastic, metal, rubber	Metal	Permanent	Insulated tool handies, toy balloons, housings
Pressure	(1) Centrifugal investment	Metal	Plaster	Unit	(1) Custom jewelry, intricate machine parts
	(2) Sand	Metal	Sand	Unit or permanent	(2) Intricate machine parts, pipe, housings
	(3) Centrifuge	Metal	Sand	Unit	(3) Small machine parts
	(1) Die casting	Nonferrous metals, glass	Iron and steel	Permanent	Sewing machine cases, door handles, carburetors
	(2) Injection molding	Plastics, rubber	Metal	Permanent	Toys, molded rubber handles, eating utensils, auto interior trim, "foamed" styrene products
	Blow molding	Plastics, glass	Iron and steel	Permanent	Glass containers and plastic containers
	Continuous	Metal (steel, copper, and aluminum alloys)	None or graphite "guides"	Permanent	Wire, rod, and intricately shaped rods
	Compression molding	Thermosetting plastics	Iron and steel	Permanent	Electrical construction hardware, gears, insulators
	Transfer molding	Thermosetting plastics	Iron and steel	Permanent	Small intricate parts, transistor sockets, etc.
	Lay-up molding	Fiberglass plastics	Iron and steel	Permanent	Boat hulls, machine covers, chairs
	Vacuum forming	Thermoplastics	Metal	Permanent	Packaging "blisters," small machine housings (electric drills, etc.), auto trim
	Straight-pressure molding	Thermoplastics	Metal	Permanent	Packaging "blisters," small machine housings (electric drills, etc.), auto trim
	Free blowing	Thermoplastics	None	None	Airplane canopies, skylight domes

Comparison of Casting Methods (Approximate)[b]

	Green sand casting	Permanent mold casting	Die casting	Sand–shell CO_2–core casting	Ceramic and investment casting
Relative cost in quantity	Low	Low	Lowest	Medium high	Highest
Relative cost for small number	Lowest	High	Highest	Medium high	Medium

continues

TABLE II (*continued*)

	Green sand casting	Permanent mold casting	Die casting	Sand–shell CO_2–core casting	Ceramic and investment casting
Permissible weight of casting	Unlimited	100 lb	30 lb	Shell oz–250 lb CO_2 $\frac{1}{2}$ lb to tons	Ounces to 100 lb
Thinnest section castable, in.	$\frac{1}{10}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{10}$	$\frac{1}{10}$
Typical dimensional tolerance, in.	0.012	0.03	0.01	0.010	0.01
Relative surface finish	Fair to good	Good	Best	Shell good CO_2 fair	Very good
Relative mechanical properties	Good	Good	Very good	Good	Fair
Relative ease of casting complex design	Fair to good	Fair	Good	Good	Best
Relative ease of changing design in production	Best	Poor	Poorest	Fair	Fair
Range of alloys that can be cast	Unlimited	Copper base and lower melting preferable	Aluminum base and lower melting preferable	Unlimited	Limited to unlimited

[a] Courtesy of American Colloid Company.
[b] Some metals cast easier than others.

and makes items by contractual agreement or to satisfy a known market.

The distinction between production and job shops, and between captive and independent producers, is not limited to the casting and molding industry. Most processing operations are organized around one or more of these business and marketing concepts.

Ideally, the biggest single advantage of a casting or molding process is that it quickly produces an item that requires a minimal amount of material removal. In economic terms, this advantage can translate into reduced production costs. Like the frozen popsicles made in home freezer molds, the finished products meet consumer expectations in terms of size and shape; again, ideally, in this case no further processing is required. Unfortunately, before this single process ideal can be achieved, certain conditions must be met. The major conditions are as follows:

1. A cavity that incorporates all the design features of the product must be provided. This mold cavity must accommodate the flow and shrink characteristics of the material being formed. Molds that meet this basic requirement can become a major cost factor in their own right.

2. The material being cast must have allotropic characteristics, that is, be able to exist in a solid, liquid, or gaseous state. Since allotropic transformation is a function of either heat or internal chemistry, or a combination of both, suitable means must be found to manipulate and control the prerequisite allotropic behavior. Basically, the material must be liquid enough to pour, yet eventually become a solid shape.

3. The material and mold, as a total process, must be designed to eliminate the negative effects of potential air or gas entrapment. Venting of mold cavities to allow escape of air and gases can become a serious design problem.

4. Shrinkage or expansion of the material's volumetric mass during its transformation from a liquid to a solid can result in product defects. Cores and other techniques must often be used to reduce bulk volume into thinner shells. Consequently, control of shell wall thicknesses and internal configurations can become a major consideration. External or female mold cavities, with all their inherent problems, are still relatively simple to design. The design of functional internal (male) core molds, by comparison, is truly complex.

5. It must be possible to remove the finished product from its mother mold or male core. This simple requirement tends to be taken for granted, but mold and core removal is often a designer's number one challenge, particularly since it is usually desirable to design molds that can be reused as permanent tools. Nonpermanent molds (i.e., molds that must be destroyed to redeem the part) are usually regarded as a comparatively expensive alternatives to permanent molding tools.

Once these five conditions are resolved (they are seldom completely satisfied) there will usually be some extraneous material that may still have to be removed. The amount of material removal required to finish a particular item may dictate that alternatives to casting or molding be explored.

IV. MATERIALS REMOVAL TECHNOLOGY

Materials removal processes are used to make incomplete castings conform to design specifications. In addition, materials removal can be used selectively to allow generic castings to be altered into unique finished products. By cutting away unwanted material, special design features can be added to common castings, effectively converting them into distinctly different end items; examples are different hold patterns, flanges removed, wall thicknesses reduced, surfaces flattened, and so forth. However, quite often materials removal processes are chosen as an alternative to casting and molding processes. When the problems associated with cast shape processing considerations become insurmountable (usually on the basis of cost-effectiveness), materials removal may become the only viable alternative.

The process of cutting a material to design specifications is called machining. Machine processes accomplish materials removal by producing a chip or physically separating the material. The conceptual basis of chip removal had its origin in the industrial revolution of the nineteenth century, and the term "machining" still carries connotations of machine tools making products of wood or metal. Although modern products present a much broader range of product materials, the nineteenth century principles of chip removal have served manufacturing process engineers long, often, and well. Even today, chip-producing methods constitute the majority of materials removal processes.

At the heart of the original materials removal concept is the notion of a sharp single-point tool cutting material like an axe or knife through wood. However, when a metallic material is the medium being shaped, the wood-splitting analogy becomes more simplistic than factual. In actuality, two technical events occur simultaneously during chip removal. First, there is indeed a sharpened tool that is supported to follow, or hold, a definite cutting plane. Next, however, the cutting action results in an area of highly compressed material immediately in front of the cutting tool. For example, even a small shop lathe will have an electric motor of perhaps 1 hp or more. Since very little of the energy produced is used to turn the mechanisms, the balance is focused on the razor-thin area defined by the cutting edge. Allowing for gear ratio advantages, this phenomenon generates pounds-per-square-inch figures more easily recorded in tons than pounds. Naturally, this results in extremely elevated and localized temperatures leading to plastic flow and heat shear. Figure 5 demonstrates the physics involved in chip formation.

Figure 5a illustrates the principle of orthogonal cutting. The tool moves at a velocity VO through the work material with the force of cutting action FC and the force of feed FF. As the tool moves into the work with a depth of cut H, forces FC and FF cause the work material ahead of the

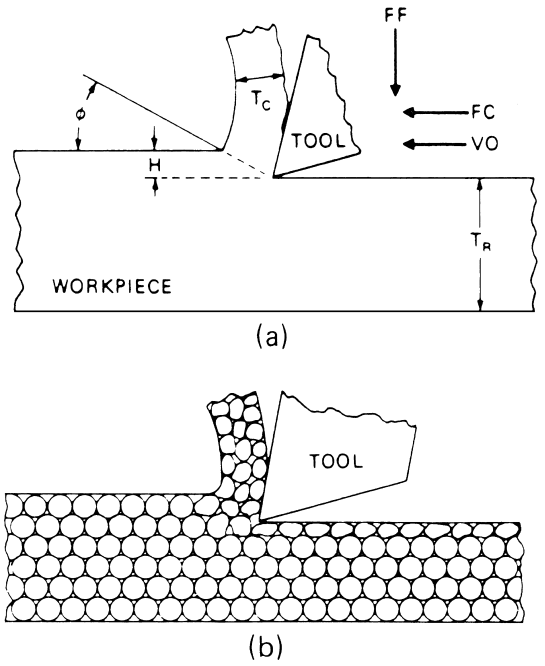


FIGURE 5 (a) Principle of orthogonal cutting. (b) Deformation in the chip and workpiece resulting from machining.

tool to be compressed. The area under compression is in a state of plastic deformation; within this area is a section referred to as the shear zone. Material failure within the shear zone occurs at a plane called the shear plane. The shear plane occurs at the shear angle ϕ, approximately a 45° angle. The work material moves up and over the face of the cutting tool and breaks off the parent material. Cutting action causes the work material in the shear plane to be compressed and flow up and over the face of the tool. The deformation occurs in both the chip and the workpiece below the cutting tool (Fig. 5b). Shear deformation can be compared to a deck of playing cards when pushed, causing a sliding motion. Deformation occurs at different stages in machining, depending on the specific materials.

Failure or separation from the workpiece occurs when a critical value is reached for that material. Cast iron, a brittle material, fractures easily because its resistance to plastic deformation is high. Soft ductile materials have lower resistance to deformation and fracture. Materials with a low resistance to plastic deformation form a continuous chip. Brittle materials, with their high resistance to deformation, form small and easily fractured or segmented chips.

Analyzed as a system, machining consists of three basic elements: the material to be shaped, secured in a way that permits it to be processed; a cutting point or edge that is attached to or part of a tool body; and the machine tool itself, some mechanism for controlling the interaction between the workpiece and the cutting tool. Once one realizes the basic elements involved in the chip removal system, the possible processing applications become fairly obvious.

These operational possibilities are (1) the material to be processed (workpiece) is held static and the cutting tool is moved, (2) the workpiece is moved and the cutting tool is fixed, or (3) both the workpiece and the cutting tool are moved so that the desired configuration is achieved.

Virtually all machine tool designs from the nineteenth century to date are based on one of these three operational approaches to chip removal. Though machine tools are often sophisticated and complex, and the single-point cutting edge is difficult to identify, one can still expect to find it hidden in the overall design. Figure 6 illustrates schematically some basic relationships between materials and cutting tools. Figure 7 introduces variations of machine tool design as the third element in the traditional materials removal system.

Once the basic principles of materials chip removal or cutting are recognized, a number of other design considerations must be addressed. The first two of these concern the nature of the materials involved. First, the material being shaped must be processable—in this case, the ease with which chips can be removed is referred to as machinability. The ultimate question is, can a specific material have reasonable machinability characteristics and still meet its engineering service requirements? Thus, machinability becomes a materials design factor even though it is only indirectly related to the end function of a material. Second, all things being equal, it is necessary to assume that the processing tool has mechanical and characteristic properties superior to those of the material being processed. That is, in the technical battle between the material being processed and the material doing the processing, it is the workpiece that must yield to the laws of physics. Consequently, every new advance in product materials forces a corresponding need for process and tool design engineers to develop even more superior processing (tooling) materials.

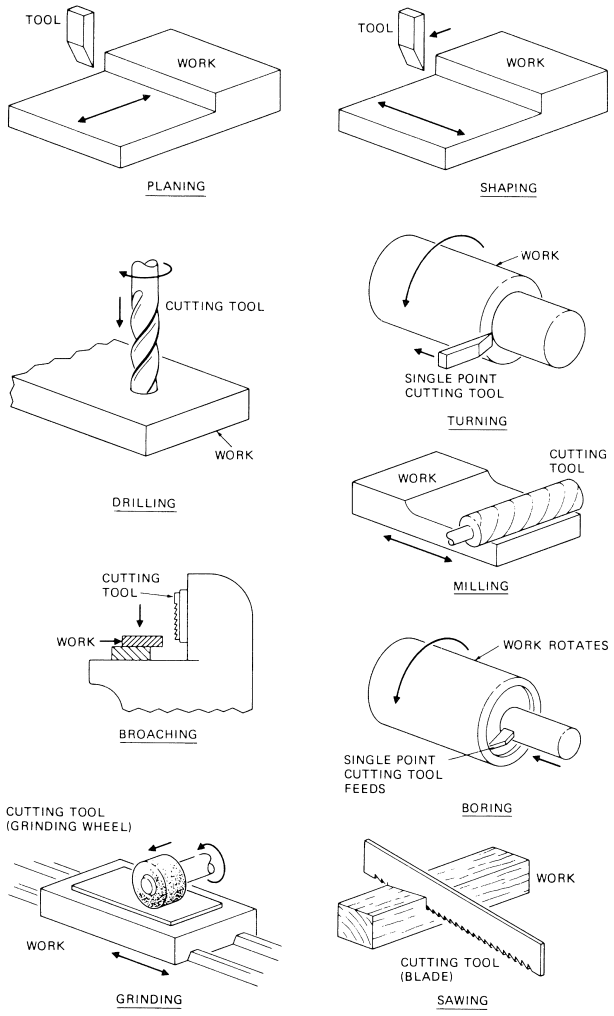


FIGURE 6 Basic traditional material removal processes.

Another body of design considerations has to do with the mechanics of processing as such. Since heat is an immediate result of chip removal operations, there must be some provision for controlling it. This need has led to a whole related technology in cooling, cleaning, and lubrication techniques. Even the question of what to do with the chips, as and after they are formed, is a major consideration. Finally, there is the necessity of ensuring the integrity of the process itself, that is, guarding against altering the properties of the workpiece due to heat or other considerations. Machines must be designed to carry out their functions in a manner that does not impair the quality of the processed items. Vibration, shock, speed, feed, distortion, accuracy, durability, and, of course, affordability are just a few of the factors that must be weighed against product requirements.

While the past few paragraphs have focused on machining processes, it is important to recognize the associated problems. The problems inherent in machine chip removal and materials separation processes, because they are often irresolvable, have led to the development of alternative approaches. Some contemporary approaches to materials removal are by laser beam, electrodischarge machining, and electrochemical principles. These processes tend to use the applied sciences rather than simple physics for the strategy for separating any given material. The basic difference is that they attempt to "take" rather than "tear" a material apart in an effort to avoid process damage.

V. FORMING AND FABRICATION PROCESSES TECHNOLOGY

A. Forming

Forming processes utilize pressure forces to alter the shape and/or size of the material being manufactured. The use of pressure to form materials sets this broad grouping of

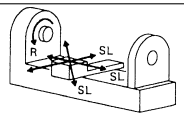
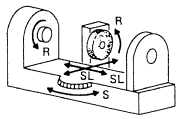
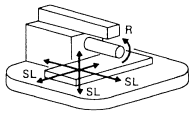
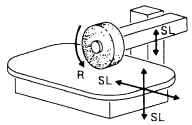
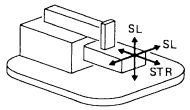
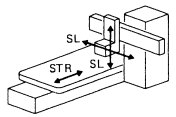
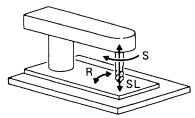
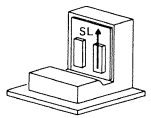
Schematics of Selected Machine Tools	Common Name of Machine Tool	Available Motions at:		Feeding Motion			Cutting Speed Obtained at:	Type of Cutting Action	Type of Cutting Tools Used	Relative Motion Between Work-Piece and Cutting Tool
		Workpiece	Cutting Tool	Applied At:	Type of Motion	When Applied				
	Lathes	Rotation (R)	Three straight line motions (SL)	Cutting tool	Straight line	During cutting action	Workpiece	Continuous	Single point	Rotation and one straight line
	Grinders (cylindrical)	Rotation and one straight line	Rotation and one straight line	Cutting tool or work-piece	Rotation and straight line	During cutting action	Cutting tool (abrasive wheel)	Intermittent	Abrasive wheel	Two rotary and one straight line
	Milling machines (horizontal)	Three straight lines and rotation	Rotation only or rotation and one straight line	Workpiece	Straight line or rotation	During cutting action	Cutting tool	Intermittent	Multiple point	Rotary and one straight line or two rotary
	Grinders (surface)	Rotation or one or two straight line or rotation and straight line	Rotation or rotation and straight line	Workpiece and cut-ting tool	Straight line and rotation	During cutting action	Cutting tool (abrasive wheel)	Intermittent	Abrasive wheels	One or two rotary and one straight line
	Shapers	Straight line or rotation	Straight line (stroke)	Workpiece	Straight line or rotation	Intermittently while not cutting	Cutting tool	Intermittent	Single point	Straight line
	Planers	Straight line	Two straight lines	Cutting tool	Straight line	Intermittently while not cutting	Workpiece	Intermittent	Single point	Straight line
	Drilling machines	Stationary or straight line	Rotation or rotation and straight line	Cutting tool or work-piece	Straight line	During cutting action	Cutting tool	Continuous	Drills	Rotary and straight line
	Broaches	Stationary or straight line	Straight line or stationary	Cutting tool	Straight line	Intermittently while cutting	Workpiece or cutting tool	Intermittent	Multiple point (broach)	Straight line

FIGURE 7 Common characteristics and interrelationships of selected machine tools.

processes apart from materials removal (machining) or casting processes. There is a great variety of both processes and equipment used, and the selection of each depends on such factors as size, material, use, cost, and availability. The materials most commonly formed by the various processes include most of the ferrous and nonferrous metals, plastics, metal powders, wood, cloth, rubber, food products, and some prefired ceramic products. Some plastics are shaped by processes that are similar to or the same as forming processes. However, in the plastics industry, as well as in the rubber, ceramics, and similar enterprises, these processes are often referred to as "molding." As a result, they are categorized as molding processes and considered a part of casting and molding technology.

Table III lists a variety of processes and some of the various subgroupings. It can be seen that there are four broad families of operations: pressing, drawing, bending, and shearing. There are also some special processes that are not normally categorized under any of these classifications. Yet these atypical processes are characterized by their use of pressure and lack of materials removal in forming the designed object.

The major advantage of most forming processes is their effect on the grain structure of the material being formed. Materials removal by machining tends to sever the workpiece material across as well as along its structural grain boundaries. The grain direction of the finished product is interrupted. By contrast, the grain structure of a formed material actually flows into designed shapes and is compressed or stretched at appropriate places (see Fig. 8). The resulting continuous grain structure increases the overall strength of the formed piece. The forming processes also exhibit a similar advantage over cast products, as the grain structure of cast items is usually uniform, which neither adds or detracts from the casting's structural integrity.

Two other major classifications or distinctions are made in reference to forming processes. When metal is the processed material, a distinction is usually made between "hot" and "cold" forming. However, with other materials, such as plastics and ceramics, "hot" is not a valid classification, since heat may not be the factor that causes a material to assume a pliable crystalline structure.

Most materials may be permanently deformed—formed to shape, that is—throughout a range of temperatures between a melting point and a solid, cold point. A given material may be completely solid, as in steels, or a semisolid where a mixture of solid and liquid material exists. An analogy for soluble materials is when the amount of solvent is more (hotter) or less (colder). The cold or "freeze" point in any case is that point where a material will break rather than assume a new shape.

The major advantage of processing materials in a hot (or solvent) state is the ease with which the materials

TABLE III Forming Processes

Category	Process	Variations	Predominant working state[a]
Pressing	Forcing	Smith forging	H
		Drop forging	H
		Upset forging	H and C
		Press forging	H and C
		Coining	C
		Hobbing	C
		Swaging	H and C
	Rolling	Rolling mill	H and C
		Roll milling	H and C
		Shape forming	H and C
	Surfacing	Knurling	C
		Burnishing	C
Drawing (pushing or pulling)	Drawing	Wire	H and C
		Bar	C
		Fluid rubber	C
		Kirksite drawing	C
		Tubular	H and C
		Spinning	C
		Stretch forming	C
		Shell drawing	C
	Extruding	Solid	H and C
		Tubular	H and C
Bending	Braking	Angular bending	H and C
		Folding	C
		Seaming	C
	Rolling	Straightening	H and C
		Cylindrical forming	C
		Conical forming	C
	Roll forming	Forming	H and C
		Flanging	C
	Bending		H and C
Shearing	Shearing	Shearing	H and C
		Notching and nibbling	C
		Blanking and dinking	C
		Piercing	H and C
		Trimming	H
		Shaving	H
	Slitting	Slitting	H and C
		Truing	C
Special	High energy rate	Explosive	C
		Impactive	C
	Magnetic forming		C
	Powder metal		Granular-C
	Compacting		

[a] H, hot; C, cold.

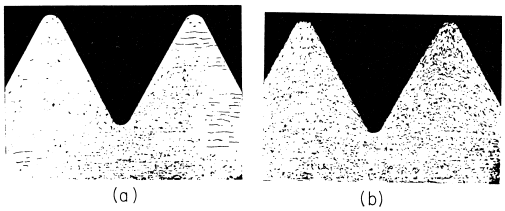


FIGURE 8 (a) Grain structure of threads formed by cutting part of the material away. (b) Grain structure of threads formed by pressing (rolling). In (b) note that the grain has been formed to the configuration of the threads, increasing the ability to withstand shear forces.

can be formed. Less force and time are required; minimum strain hardening occurs while the material is heated to pliability. No significant increase in hardness or decrease in desirable structural qualities occurs during heat forming. Working a material during the hot stage also helps refine the grain structure. Above certain specific temperatures, recrystallization takes place, and voids and imperfections are refined or reduced as a result of new grain growth or crystalline realignment. See Table IV for examples.

At first glance, there might appear to be no advantage to cold forming materials, given the merits of hot-working processes. This is not necessarily the case. A distinct advantage of cold-working is precisely that no heating is involved. Elevated temperatures can result in undesirable surface reactions between the material and its environment. Tolerances may be difficult to maintain because of a material's sag, flow, thinning, or creep and thermal characteristics. Cooling histories can become crucial to the integrity of the finished product. Thus cold processing can result in superior surface finishes and dimensional tolerancing. Since no heating or heat soaking time is involved, overall production rates can often be comparably faster. In many cases, the tendency of a material to work harden can be used to advantage. Cold forming can be used to improve a material's strength when its cold setting

TABLE IV Examples of Recrystallization Temperatures

	Approximate temperature	
Metal	**°F**	**°C**
Aluminum	300	148
Copper	390	198
Iron	840	449
Lead	Below 75	25
Magnesium	300	148
Nickel	1110	600
Tin	Below 75	25
Zinc	75	25

properties are strategically incorporated into its processing. Also, in many cases, cold processing allows existing grain direction and crystalline structure to be redirected to complement known service requirements, thereby eliminating further process engineering.

The four basic forming processes—pressing, drawing, bending, and shearing (which includes blanking and piercing)—all use machines known as presses to exert pressure (see Fig. 9). The basic press is the temporary equivalent of yesterday's village blacksmith. However, today the huge hammers and straining muscles have been replaced by forming machines. Of course, greater forces and more complex operations are possible, but the basic operations are the same. Two types of basic presses are in common usage. The first exerts force slowly with a squeezing action; the second provides pressure combined with a sudden impact. The former are quite appropriately referred to as presses; the latter are classed as hammers (see Fig. 10). Technically speaking, when a combination of impact and pressure is used to form a part, the process is called forging rather than forming.

There is a third classification of forming tools used to bend or fold materials in which machines are given various

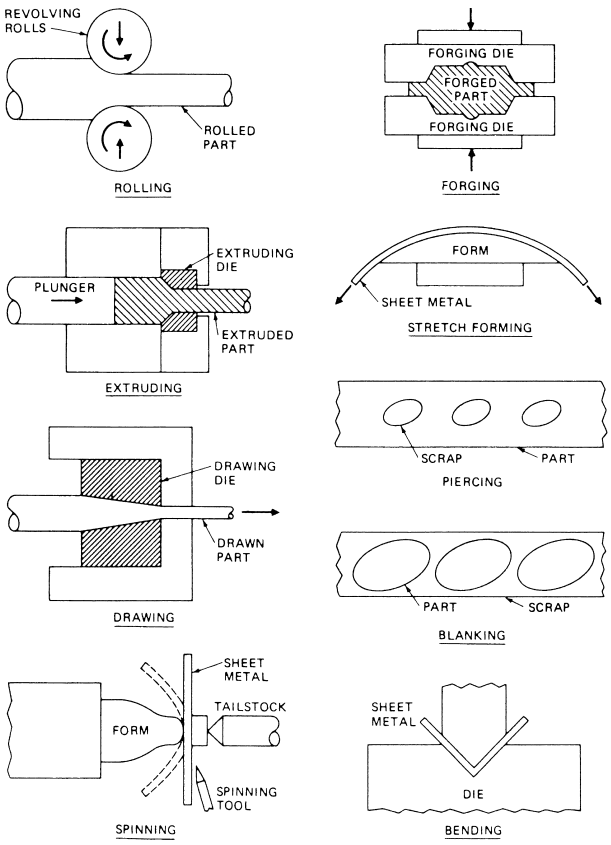


FIGURE 9 Schematics of the most common mechanical forming processes.

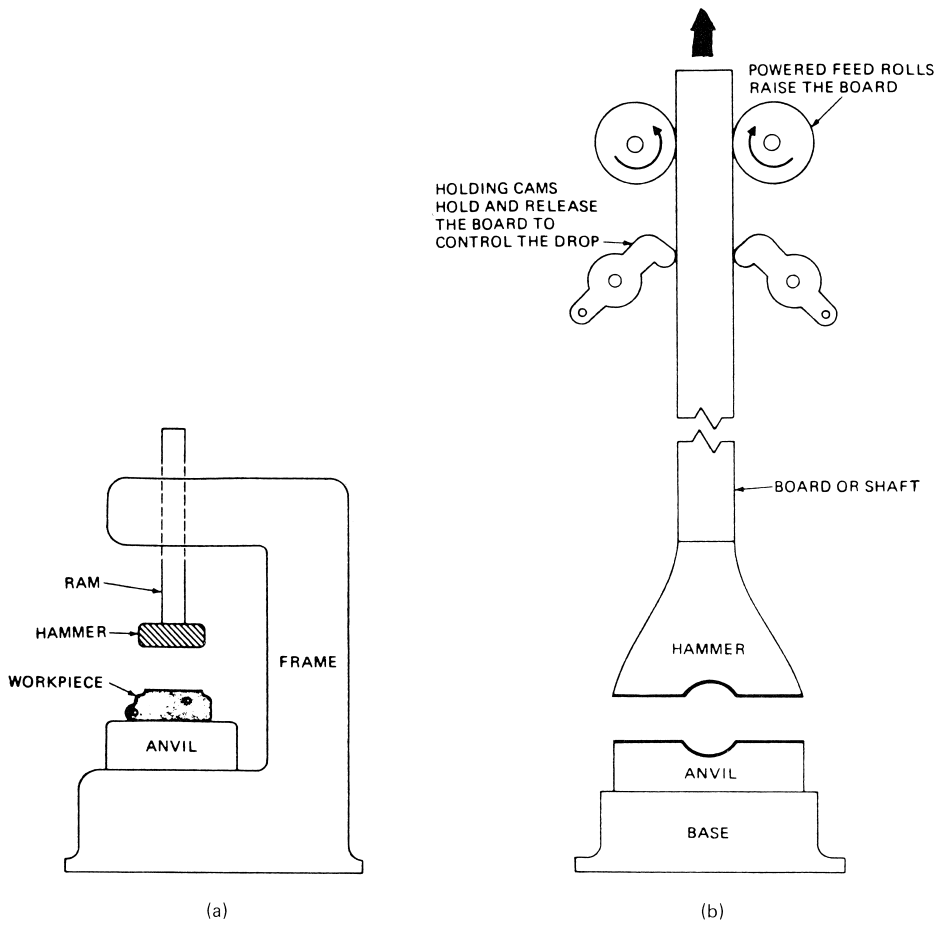


FIGURE 10 Schematics of basic press and hammer design.

names. Some of the machines use obscured applications of press and impact (hammer) principles; others are so remotely related as to make comparisons irrelevant. The machines using obscured applications of squeezing and impact principles tend to be identified with the processes they perform; examples are the rolling mill, roll-press, brake press, and impact extruder. The more remote nongeneralizable forming processes are often referred to as special processes. However, the "special" in this case is valid only from a historical perspective. The s pecial forming processes are now in common use and can be subdivided into three major alternative forming techniques: (1) magnetic forming, (2) high-energy-rate forming, and (3) powder metallurgy processing (also known as sintering).

The magnetic forming processes induce sudden magnetic forces about the object to be shaped, causing it to assume the shape of the mold surrounding it. Magnetic forming is used primarily in the aerospace industry to form extremely large sheets of thin materials and also extensively to form unusually complex tubing shapes. High-energy-rate forming is a rediscovery of the process used to form nineteenth century brass cuspidors. The process uses a contained high-energy charge (or explosion) to generate pressure on the surface of a liquid. The liquid, in turn, provides a multidirectional uniform blast that forces a sheet of material into the mold cavity, allowing intricate shapes to be formed quickly and without complex drawing or forming operations. Powder metallurgy techniques encompass a broad range of applications, but with reference to forming, they consist of pressing metallic powders into some compacted solid shape. The process is a true technological anachronism, it has been discovered and rediscovered in a number of historical contexts ranging from ancient Egypt to the Incan Empire to Victorian England. Powder metallurgy is used in modern industry to form an extensive array of pressure-formed small parts.

B. Joining Processes

Up until now, manufacturing processes have been discussed as if the end product consisted of a single material that is cast, molded, machined, or formed into some marketable item. This concept helps facilitate and simplify the discussion of the basic processes involved in

manufacturing. However, in actuality, almost every product on today's market is made up of more than one piece. Most processed materials become mere parts in a larger assembly. A modern automobile is made up of several thousand parts and pieces that have been fabricated into a single unit product. Thus, the way in which materials are fabricated or joined has led to methods and techniques that qualify as processes in their own right. These fabrication processes tend to fall into three broad initial categories: (1) mechanical fastening, which includes all methods employing mechanical devices to hold two or more parts together; (2) adhesion, which relies on surface irregularities and interstitial adhesion to glue or cement parts together; and (3) welding, which uses fusion or cohesion to actually produce a single monolithic joint.

The process engineer or designer should be familiar with the characteristics, advantages, and limitations of the basic fabrication processes in order to select the ones most appropriate for a given product. The basic consideration is whether the joined assembly is to be regarded as a permanent or nonpermanent unit. Permanent joining or fastening processes do not allow for future disassembly or separation of component parts once they have been fabricated. Nonpermanent processes allow for the eventual removal or replacement of parts that may be damaged or worn in service. The need to adjust parts to compensate for normal wear may also dictate nonpermanent solutions to joining problems.

A compromise approach to the issue of permanent versus nonpermanent fabrication is to produce a number of permanent subassemblies that are in turn assembled, by nonpermanent methods, into a larger assembly or end product unit. This approach is generally called "modularized" or "componentized" product design because entire subassemblies rather than individual parts are replaced in service.

Mechanical fastening can be either permanent or nonpermanent, depending on the fabrication methods or techniques used. Mechanical fastening is composed of two broad groups of processes, involving joints and fasteners. Joints, which feature a mechanical configuration such that parts are supported or retained by physical interference, are classified as either seaming or joinery. Mechanical fasteners includes all types of screws, pins, nails, staples, rivets, clips, springs, bolts, and other mechanical devices. Usually, mechanical fastening includes all those processes which join two or more parts by using a separate piece of material that serves as the joining agent (see Fig. 11).

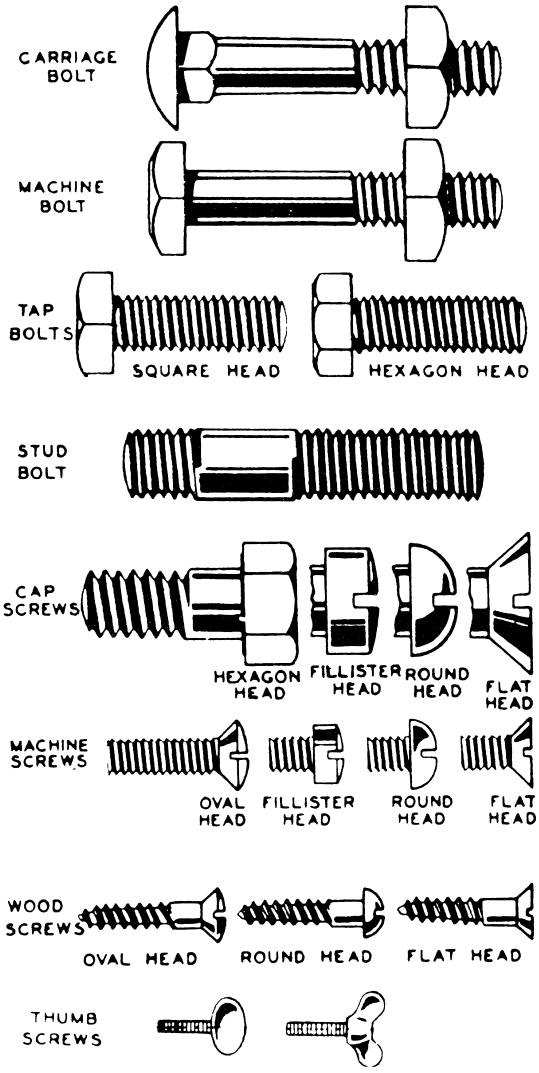
Adhesive fabrication is usually regarded as a permanent or, more technically, semipermanent process. Adhesion processes rely on the fact that certain materials will adhere and attach themselves to the surfaces of other materials. In these processes, the material used as the joining agent is not the same as the material or materials being joined. Adhesives attach themselves to the surfaces of other materials in a phenomenon commonly thought of as gluing. There is no mechanical connection involved in adhesion, nor is there a physical fusion of the base materials. Consequently, while such joining may appear to be permanent, its stability is dependent on the integrity of the adhesive medium and the materials being joined. This characteristic semipermanence can be a design handicap, but it can also be used to advantage. Within known service parameters, fabricated units can be made to function as permanent assemblies. Yet, by using special de-adhesion techniques, the same units can be disassembled at will. The manner in which windshields and certain panels are attached to modern automobile bodies is a good example of strategic adhesive joining.

Welding or cohesive processing is done by both heat and pressure techniques and, as a method, is considered to be permanent in nature. Welding of metals relies on fusion of the base parts into a single material by some combination of heat, pressure, and external or internal melting, plus some form of controlled resolidification. Plastics can be joined by similar processes. In addition, plastics can be welded with solvents that will dissolve the base materials, melting them together and leaving a fused joint on evaporation. The array of welding processes in common use has become so extensive that they are now addressed as a separate study. Figure 12 is a schematic diagram that provides some idea of the scope of welding as a separate technology. (Note: This diagram includes all welding processes commonly used with metallic materials. Some can be used with plastics, but Fig. 12 does not necessarily include those fusion processes which are unique to the joining of plastics.)

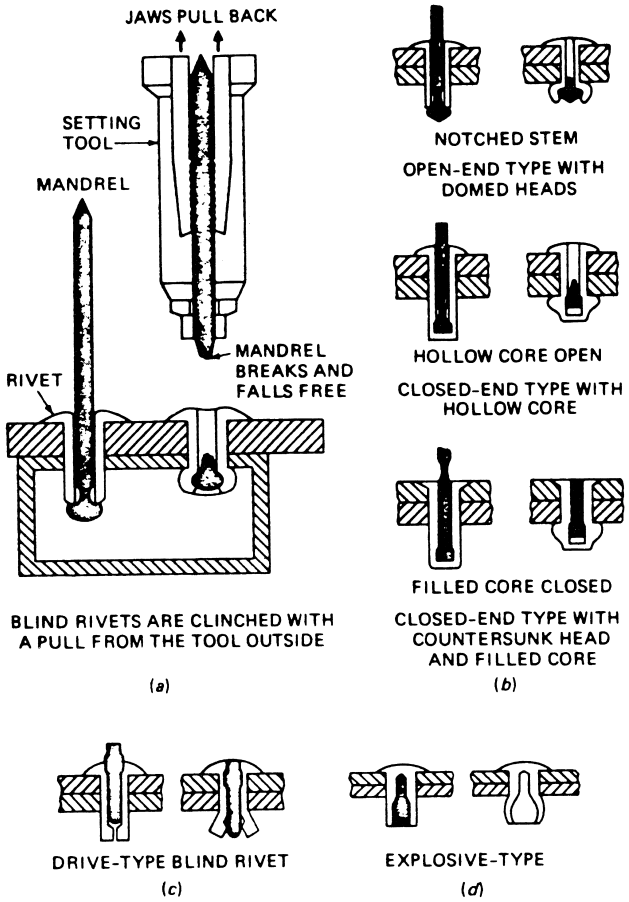
VI. MISCELLANEOUS PROCESSES TECHNOLOGY

At this point, source materials "x," "y," and "z" have been cast, molded, machined, or formed, and then fabricated by mechanical, adhesive, and welding processes into end product "P." In the abstract, this sequence follows the basic manufacturing processes from raw materials to finished product. But while these processes may be basic, they are not necessarily all-inclusive or even all-important. There are a number of ancillary and related processes that are vital to successful manufacturing. Three of these are heat-treating, cleaning, and finishing.

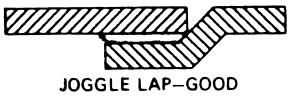
During manufacturing all metallic materials and most nonmetallic materials are subject to heating as a result of process operations. Materials may be heated for melting, forming, or fabrication; for vaporizing, baking, or curing;

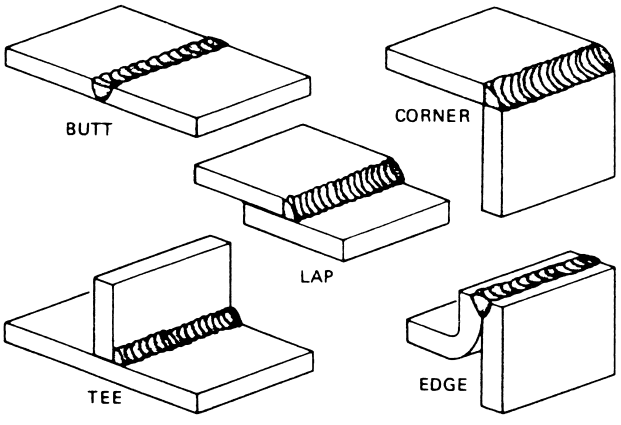


Types of Bolts and Screws

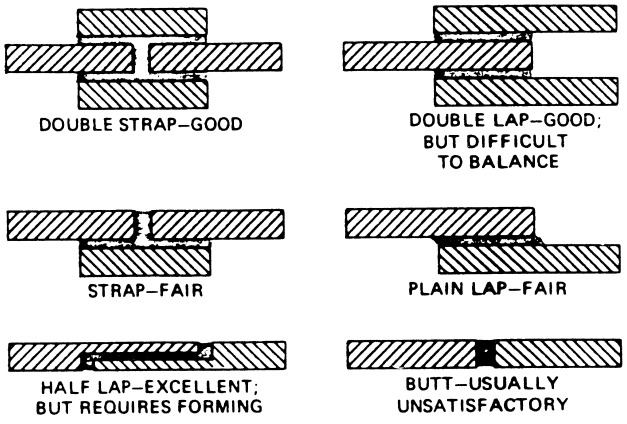


(*a*) The "mandrel" type blind rivet. Blind rivets are clinched with a pull from the tools outside. (*b*) Shapes of mandrel-type blind rivets. (*c*) Blind rivet. (*d*) Explosive type.



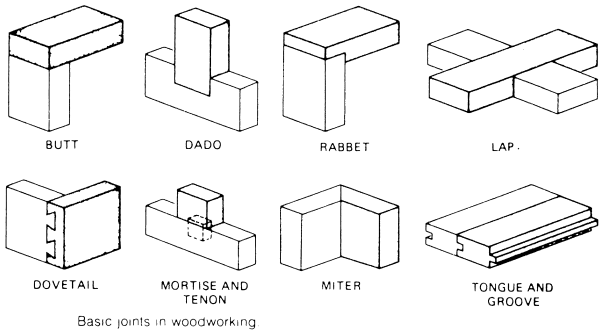


Basic types of joints.

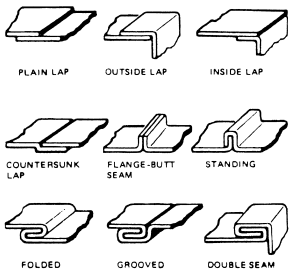


Types of adhesive-surface joints.

FIGURE 11 The technology of joining processes.



Basic joints in woodworking.



The most commonly used sheet-metal seams.

FIGURE 11 (*continued*).

or for what is formally known as heat-treatment. Although all heating operations may have heat-treatment effects on a material, controlled heating and cooling to obtain specific properties is a highly specialized ancillary process. It is done to establish or reestablish certain physical and characteristic material properties that may be critical to product's in-service performance.

Heat-treating is a broad generic term that includes softening, hardening, and surface treatment processes. Softening includes such processes as stress relieving, annealing, normalizing, and spheroidizing. Hardening includes such processes as martempering, austempering, and hardening. Among the processes included in surface treatment are nitriding, cyniding, carburizing, and induction and flame hardening. These processes are performed for the purposes of increasing strength strength and hardness, improving utility, changing grain size and chemical composition, improving machinability, relieving stresses, modifying the electrical and magnetic properties of materials, and filling a number of design-specific requirements. Heat-treat processes involve changes or modifications in the atomic- and microstructure of a material, and a complete understanding of them requires extensive knowledge of metallurgy or physical chemistry. But, in ordinary terms, heat-treating attempts to accomplish one or more of the following objectives: increase hardness, reduce hardness, or remove internal stress. All these objectives involve atomic, chemical, and microscopic changes in the internal structure of a material. They are changes that are not visible to the naked eye. As such, they require the manipulation of time and temperature in strict accordance with relevant scientific theories and practices.

Two of the basic requirements most industrial products must meet are aesthetic appeal and resistance to deterioration. Aesthetic appeal pertains mainly to surface finish and is an important factor in customer reaction. An attractive finish is usually more desirable to a potential buyer and thus makes the product more salable and competitive in the market. Resistance to deterioration is a measure of a product's surface durability against environmental attack. Some materials have a natural resistance to environmental deterioration; others need surface treatments or coatings to protect them from external elements. The three major objectives of finishing processes are to preserve, protect, and, in most cases, beautify the base product. A wide variety of process operations have been developed to satisfy these three basic requirements. They are generally classified as surface cleaning and surface finishing processes. However, these broad classifications are not mutually exclusive, and a number of processes serve both purposes, cleaning and finishing. Also, these processes tend to be made up of a number of complete operations that are themselves complete subprocesses. For example, a material may go through a cleaning operation that effectively prepares it for a number of alternative finishing options.

Generally, surface cleaning can be accomplished by mechanical and chemical processes. Abrasive techniques, using the mechanisms of free forming materials such as metallic grit, aluminum oxide, silicon carbide, or sand, are perhaps the most common form of bulk cleaning processes. These abrasive agents can be used dry, in combination with high-velocity air streams, or suspended in some fluid medium. They provide the combined benefits of both surface cleaning and finishing, depending on the materials involved. However, mechanical approaches such as belt sanding and wire brushing machines and tools are also widely used, providing similar benefits. In cases where the objective is merely to polish or buff a surface, mechanized wheels and pads are quite common. Various electrochemical methods are often used to clean or polish surfaces. In addition, a number of purely chemical cleaning methods can be used, including acid baths, alkaline or salt solutions (which are often heated), or any of several petroleum-based solvents. And no list of potential cleaning processes would be complete unless it considered the use of steam or other high-pressure fluids, often just water to which detergents or solvents may or may not have been added.

Surface finishing processes can be classified as organic, metallic, or miscellaneous finishing operations. The major purpose of a surface finish is to protect and/or decorate the

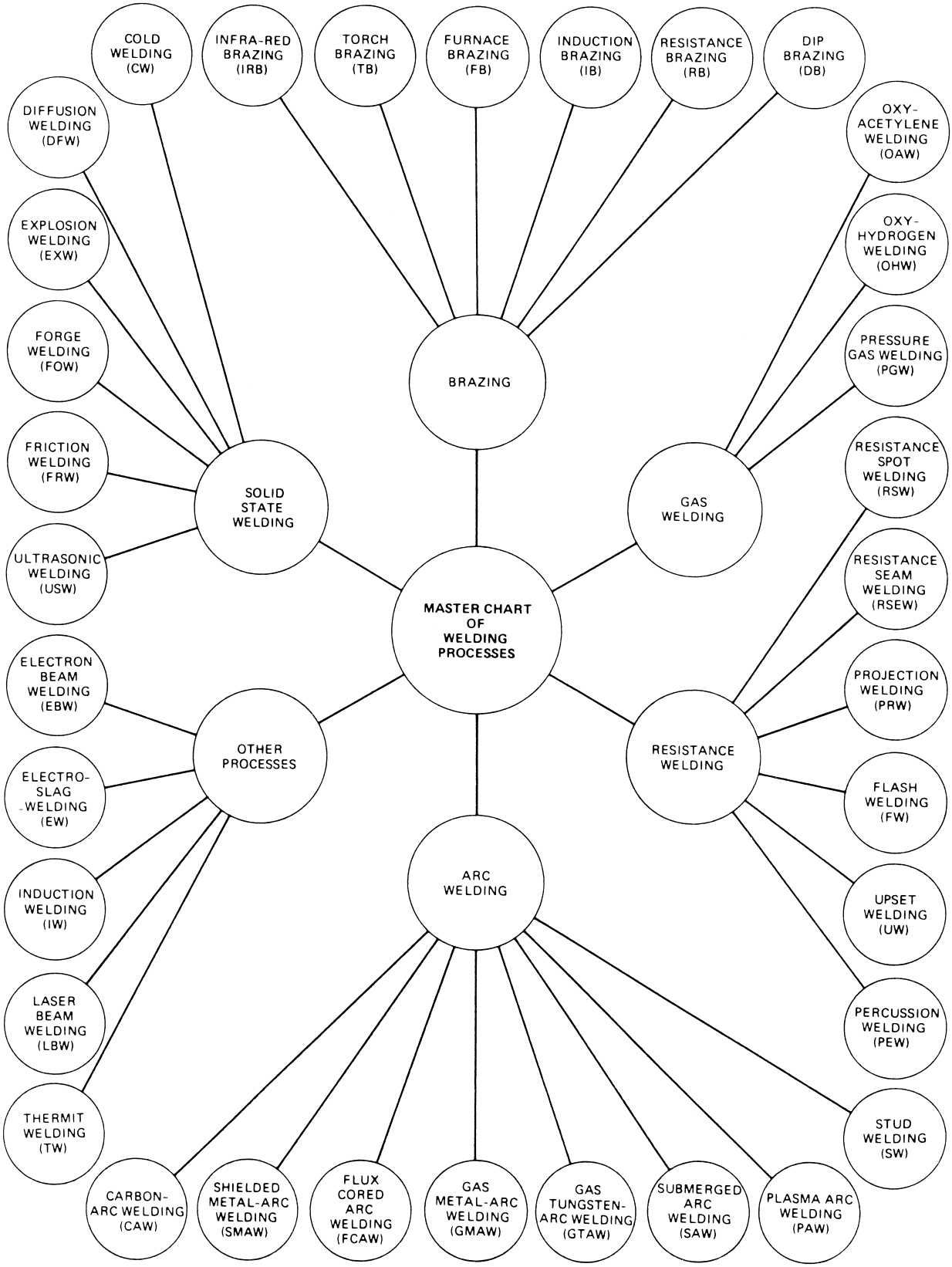


FIGURE 12 Schematic diagram of welding processes technology. [Courtesy of the American Welding Society.]

underlying product surface. This can be done by adding material to or removing it from the underlying surface, or by some combination of both methods. When material is added to a surface it is called a coating. Organic finishes are used as protective and decorative coatings. Organic finishes protect the base material by providing a continuous film between the environment and the material to be protected. These finishes depend on adhesion to hold the film on the surface of the base material. Therefore, the host surface must be thoroughly cleaned, and it usually requires an intervening chemical treatment called a primer before the organic coating will adhere satisfactorily. Organic coatings are what the layperson normally thinks of as paints, varnishes, lacquers, shellacs, and enamels. While there are important technical differences between these various coatings, they are all formed in whole or part with carbon compounds. They can be applied by brushing, spraying, dipping, or rolling, or by more sophisticated electrocoating or powder coating techniques. Regardless of the process, the end result is a film of material that serves as a protective coat over the surface of the base product.

Unlike organic coatings, which are considered relatively soft, metallic coatings are regarded as hard coats. Among the most commonly used metallic coating processes are hot dipping, electroplating, electrolysis plating, electroforming, anodizing, vacuum metalizing, and metal spraying. The most frequently used coating materials are chrominum, nickel, copper, zinc, lead, cadmium, tin, silver, gold, and platinum. Like the organic coatings, metallic coatings can be used to protect or beautify the host surface, but metallic processing can also be used to restore or build up the base material. Metallic coatings can also be used as barrier coatings between materials that must function together but are chemically incompatible.

Like organic finishes, most metallic coatings rely on adhesion or pseudo-bonding to maintain their functional integrity. Since metals are involved, host surfaces must usually be chemically as well as physically clean. In most cases the coating material is solidly dissimilar from the base material. Since this is true, it is proper to speak of an actual physical coat being formed on the product. Some oxide coating processes are notable exceptions. For instance, in the anodizing of aluminum the protective coat is formed by an electrolytic action that produces a natural aluminum oxide on the surface of the parent material. Often an inert dye is introduced to provide visual verification that the process has, indeed, occurred. A side benefit of this dyeing technique is that the colors can be used to improve the appearance of the material. In some cases, material is actually etched away by various chemical or electrolytic means to ensure surface integrity before the coating process. Even if this is not done, the fact remains that no new materials were introduced by the process. Thus, it can be argued that anodizing is a surface treatment rather than a coating.

The use of porcelain enamels and ceramic coatings constitutes the bulk of the miscellaneous surface finishing processes and their related preparation and cleaning techniques. Porcelainizing and ceramic coating methods date back to antiquity. Generally, these vitrifying processes involve extreme heat applied in a manner that glazes the surface or surface coating of a product. Historically, these processes have had a large but fixed range of product applications. With the advent of a new family of coating materials that do not require excessive and potentially damaging heating operations, porcelain or ceramiclike finishing processes have become more common. Most of these new materials are plastics and, as such, could be classified as organic coatings; however, the tendency is to process them like vitreous materials.

VII. NEW AND EMERGING PROCESSES TECHNOLOGY

In any contemporary discussion of new and emerging processes, there is a temptation to focus on the computer as the harbinger of the times. The past decade alone has generated dozens of computer-related terms that have become commonly known by their acronyms. Examples are computer-aided design (CAD), computer-aided engineering (CAE), computer-aided manufacturing (CAM), computerized numerical control (CNC), management information system (MIS), and computer integrated manufacturing (CIM). The rapid generation of such terms lends an atmosphere of awe to the emerging reality of omnipresent computerization. Indeed, when the phrase "high technology" is used, it is nearly always in reference to computers and computerization. While the computer phenomenon is truly impressive, it may not be as singularly significant as it is momentarily held to be.

From the historical perspective of process engineering, the computer represents a logical extension of the evolution of manufacturing. At an ever accelerating rate, manufacturing has gone from artisanship to mechanization to mass production, which underwent automatization (automation) in an effort to increase productivity. Viewed from that perspective, computerization is the next logical step in the quest for a fully self-regulating process facility. The steam engine was revolutionary in that it replaced repetitious processes using muscle power. In most cases, computers are revolutionary simply because they replace repetitious mental processes. In both cases, the results are regarded as awesome because of the expanded capabilities and speeds involved. It is perhaps the question of what human beings will do with this new leisure, the holiday from repetitive thinking, that is truly significant, in the

same sense as the transition from brute labor to skilled technicians.

Until quite recently, the primary justification for process automation was high-volume production. The automated production machines of 30 years ago were almost exclusively special-purpose machines. They were designed to do specific processes needed to complete a particular product, item, or part (e.g., boring the engine blocks of one designated make, type, and model of automobile). This type of automation is regarded as "hard" automation because it is difficult to convert it to accommodate model or engineering design changes. Most of the process functions are absolute, and they are usually controlled by cams, profile templates, or other mechanistic control devices, many of which are marvels of kinematic design. The feedback sensors that control the parameters of envelope of the processes tend to be stop-action switches—safety devices designed to prevent machine damage rather than monitor product quality. In hard automation, product quality. In hard automation, product quality is normally a function of accurate part loading as verfied by postprocess inspection.

A newer generation of automated process machines falls in a category defined as semifixed (or semihard) automation. The advent of numerical and computer technology has enabled machine designers to extend the capability of product-specific machines. For example, the automated engine boring machine can now be designed to bore several types of similar block configurations rather than one specifically cast item. Improvements in cybernation and machine design allow these machines to better sense, analyze, actuate, and move the workpiece through the envelope of processing operations. Semihard automation is usually equipped to monitor process quality in addition to controlling and safeguarding machine behaviors.

Although both hard automation and semihard automation are still very much in use today, both concepts are already thought of as classic archetypes. These concepts evolved out of the metals processing industries (specifically, the automotive industry); they were and are designed to meet the need for mass production at lowest cost and are both pre-high-technology concepts. That is, they both represent electromechanical improvements based on historic design concepts dating back to the notion of interchangeable automation represent the most sophisticated evolution of past practices rather than new technology per se.

Contradistinctive to hard or semihard automation are more recent types of processing based on high-technology concepts. One of these concepts is the flexible automation system. In the past decade, this has developed to the point where it is conceived of as a flexible manufacturing system (FMS). The difference between so-called classical automation and flexible automation is threefold: (1) automation is seen in terms of a whole system rather than a single machine, (2) flexible automation strives for the advantages of unmanned processing without the need for volume quantities, and (3) flexible automation relies on scientific programming of the relationships between processing tools and processed materials. As a concept, FMS applications look to new science-based solutions to processing problems rather than to historical precedents.

Closely related to the FMS concept, and perhaps eventually to merge with it, is CIM theory. CIM suggests that computers can be integrated into manufacturing and production functions in a way that allows processing to be self-directed and self-regulating from ideation to end product. It is an outgrowth of CAD concepts merged with CAM practices. A more restricted approach to CIM, also evolved from CAD/CAM technology, is computer-integrated production (CIP). CIP limits itself to integration of the historical secondary or on-line processes. In practice, if not in theory, CIP and FMS are similar concepts. Both forms have met with impressive successes. However, to date, these successes have fallen far short of the ideal of a totally automated factory.

The ultimate in emerging process technology is the attempt to apply computers and computerization in a way that integrates manufacturing as a total, singular, and complete system. Computer integrated manufacturing, as a processing concept, offers the prospect of going from ideation to finished product without human assistance. In theory, CIM could integrate all the functions shown in Fig. 2 and make them self-regulating in terms of the total system's needs. Many of these individual functions can already be supported or controlled by various "stand-alone" CAE, CNC, CAM, or MIS computer system. However, the links between these discrete functions are still heavily dependent on human intervention in the form of thousands of interdepartmental memos, data sheets, letters, reports, pictographs, drawings, work orders, and other paperwork; intervention in the form of necessary but often arbitrary or uncoordinated decision making; and intervention in the form of manual overriding to ensure that the end product meets design expectations. Hence, the challenge to CIM, as the ultimate process, is to supersede the vital human presence in the manufacturing system.

A major step toward computer integrated manufacturing is the perfection of the technology of robotics. Early attempts to create robots that would fulfill processing needs were frustrated or confused by basic cybernetic fallacies. Early robotics, dating back to nineteenth century science fiction, was based on the study of so-called human–machine relationships. As a consequence, most of the robots actually produced tended to be nothing more than electromechanical approximations of people; that is, they emulated critical human capabilities. Although these robots provided exciting basic research opportunities,

they proved less than promising in the real world of manufacturing.

The current trend in robotics (and one with immediate practical potential) is to approach the problem on the basis of functional robot-process interactions without reference to human–machine processing traditions. With this concept, a two-factor approach to process solutions is possible: (1) robots are designed specifically to meet process needs and (2) processes and even the products themselves are or may be redesigned to complement robotic capabilities. This robot-process design orientation, together with existing computer technology, has advanced to the point where some companies now claim CIP, if not CIM, capabilities. A core of firms in the United States alone are currently making this claim.

For the process engineer, the real revolution is in the reversal of the approach to manufacturing technology. Historically, the nature of available raw materials determined how, or if, a functional end product could be processed and produced. Modern process engineering begins with a firm grasp of the service requirements that must be satisfied, an extensive knowledge of the materials available or being developed that will meet those requirements and a complete conceptural understanding of the ways in which any given material might be successfully processed. In these areas the computer has become a process in itself. Brute-force experimental development, which once took years to study and months to summarize, can now be computer-simulated in weeks and concluded in days.

Another major influence on emerging processes is a new relationship that sometimes exists between materials science and the generation of engineering materials. Historically, materials scientists and engineers have determined what materials best meet product needs. Now that relationship is often reversed. Materials are often designed to serve scientific and engineering needs. The materials revolution ushered in an era in which materials could be "built" directly from the elements in the periodic table. Experience with plastics, ceramics, and cermets, for example, has led to processes that allow materials to be constructed of elements as small as a single molecule. Thin-film technology, with its micromanipulation and microproduction techniques, has made the microprocessor possible. The microprocessor, combined with the older thick-film processes used to make printed circuits, has made the microcomputer possible. The process technology generated by this materials revolution has led to whole new engineering and scientific support industries which, for lack of a better phrase, are referred to as high-tech industries. The most obvious of these industries are the electronics and computer firms. However, these firms are only representative of the hundreds of companies producing materials by microproduction processes.

The new, complex relationships among processes, materials, engineers, and scientists—many of which are systematic paradoxes—have greatly expanded the list of atypical processes. Ultrasound chemical milling, electrical discharge machining, high-velocity fluids, laser beam technology, and the thick- and thin-film processes used in the electronics industry are just a few of these new processes. Most of them are unique to a particular processing need. What they have in common is that they are all based on microtechnologies, applied sciences that more precisely control the effects of processing on the material being processed.

It seems safe to say that data feedback, for the purpose of quality control, will continue to expand as a process in itself. Monitoring of ongoing processes is becoming increasingly crucial to predicting in-service performance. Again, the sort of data and information being compiled can be adequately tracked only by computerization. Quality control processes, which have not been adequately discussed in this article, are emerging as vital rather than ancillary functions. As the relationship between quantity and quality (in terms of engineering expectations) becomes less and less ambiguous, the need for quality assurance will grow. Methods, techniques, instruments, and theories of measurement will become more sophisticated.

Attempting to predict the future of the new and emerging processes is a frustrating exercise in expert opinion. Manufacturing is experiencing an epoch of change. However, it also marks the beginning of an era that is as exciting as it is frustrating. The challenges beg for creative innovation, new ideas, and fresh solutions. At no time in history have humans been more capable of converting abstract concepts into concrete products. It is a process in which engineers and scientists are more apt to be supported, rather than threatened, by computers and computer technology.

SEE ALSO THE FOLLOWING ARTICLES

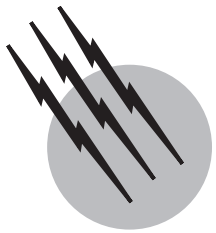
ADHESION AND ADHESIVES • BONDING AND STRUCTURE IN SOLIDS • COATINGS, COLORANTS, AND PAINTS • CORROSION • EMBRITTLEMENT, ENGINEERING ALLOYS • LASER-MATERIALS INTERACTIONS • METAL FORMING • PLASTICITY (ENGINEERING) • POLYMER PROCESSING • RELIABILITY THEORY • SURFACE CHEMISTRY • VIBRATION, MECHANICAL

BIBLIOGRAPHY

ASM International. (1999). "ASM Handbook on CD ROM: Manufacturing Processes Collection, Volumes 7, 14, 15, 16, and 17."

Beddoes, J., and Bibby, M. J. (1999). "Principles of Metal Manufacturing Processes," Wiley, New York.

Creese, R. C. (1999). "Introduction to Manufacturing Processes and Materials," Dekker, New York.
DeGarmo, E. P. (1997). "Materials and Processes in Manufacturing, 8th Ed," Wiley, New York.
DuVall, J. B. (1996). "Contemporary Manufacturing Processes," Goodheart-Wilcox, South Holland, IL.
El-Wakil, S. D. (1998). "Design for Manufacturing Processes, 2nd Ed," PWS PUB-ITP.
Fellers, W., and Hunt, W. (2001). "Manufacturing Processes for Technology, 2nd Ed," Prentice-Hall, New York.
Groover, M. P. (1996). "Fundamentals of Modern Manufacturing: Processes and Systems," Wiley, New York.
IEEE. (2001). "Intelligent Robots and systems," 2000 IEEE/RSI International Conference.
Machover, C. (1995). "CAD/CAM Handbook," McGraw-Hill Professional, New York.
Moon, F. C. (ed.) (1997). "Dynamics and Chaos in Manufacturing Processes," Wiley, New York.
Nof, S. Y. (ed.) (1999). "Handbook of Industrial Robotics, 2nd Ed," Wiley, New York.



Masonry

Gregory R. Kingsley
KL&A of Colorado Structural Engineers

I. Introduction
II. Component Materials
III. Elements and Building Systems
IV. Construction
V. Properties
VI. Structural Design
VII. Testing
VIII. Further Study in Masonry

GLOSSARY

Bed joint Horizontal mortar joint in masonry.
Efflorescence Deposition of white salts on the surface of masonry following the migration of moisture through the wall.
Head joint Vertical mortar joint in masonry.
Load-bearing masonry Masonry subject to or designed to carry loads in addition to its own dead load, except a well element subjected only to wind or earthquake loads in addition to its own dead load. Also termed "structural masonry."
Movement joint Break in masonry construction, either horizontal or vertical, to accommodate expansion or shrinkage of the masonry.
Plain masonry Masonry without steel reinforcement, except that which may be used for bonding or reduction of the effects of dimensional changes due to variations in moisture content or temperature.
Reinforced masonry Masonry in which steel reinforcement is embedded in such a manner that the two materials act together in resisting forces.
Running bond Masonry pattern in which units in alternate courses overlap each other.
Stack bond Masonry pattern in which the head joints form continuous vertical lines.
UBC Uniform Building Code.
Wythe A continuous vertical section of a masonry wall, one unit-width in thickness.

THE TERM "MASONRY" is applied to all building systems that are constructed by stacking relatively small units of stone, clay, or concrete, joined by mortar, into the form of walls, columns, arches, beams, or domes. Masonry, in one of its various forms, was used in some of the earliest examples of semipermanent or permanent construction known to man, and it continues in use today. Early systems were quite literally massive, using the weight of the masonry itself to stabilize structures against lateral

loads and to span great distances using pure and elegant arches, vaults, and domes perfected by the Romans. In modern times, with the need for lightness and economy, and with the advent of engineered masonry, masonry systems have become relatively thin, and are often reinforced with steel to carry the flexural stresses formerly eliminated by the shear bulk material. While some attempts have been made to mechanize the process, masonry construction has always been a manual process or craft, creating great structures by hand, one brick at a time. From this comes both its beauty and its strength.

I. INTRODUCTION

Masonry is one of the oldest known building materials still in use as a modern building system, although modern masonry has evolved considerably from its ancient origins. With the continuing search for economy and new forms in the built environment, the massive stone masonry arches, vaults, and domes of the Roman era have been replaced by steel-reinforced clay and concrete masonry elements much lighter and thinner than their historical precedents. Other building materials such as steel, concrete, and timber are now widely used as structural materials even when the aesthetic richness of masonry is still desired; as a result, a significant percentage of masonry used in buildings today is only as a veneer. Considerable research efforts at the end of the 20th century have elevated the potential structural uses of masonry—even in high seismic zones where masonry has traditionally behaved poorly—to compete with more modern materials.

This article explores the basic concepts of masonry as a modern building system, addressing the most common engineered forms of construction using clay and concrete units in both structural and veneer applications. Characteristics of the materials and elements of typical masonry construction are described in some detail, while methods of design are addressed relatively briefly and qualitatively.

II. COMPONENT MATERIALS

A. Masonry Units

Masonry units used in modern construction, commonly referred to as bricks or blocks depending on their geometry, are made of clay, concrete, or stone. Clay or concrete units can be "solid" or hollow. So-called solid units are rarely truly solid, having up to 25% core space resulting from the extrusion manufacturing process. The voids also serve to decrease their mass, improve bond to the mortar, and ease the brick-laying process. Hollow units are those with greater than 25% core space.

Masonry units come in a limitless array of shapes and sizes, although most are based on a predetermined module that allows orderly layout and dimensioning of walls. In the United States, the standard modular unit for clay bricks 3.75″ thick by 2.25″ high, by 8″ long (95 mm × 57 mm × 203 mm). In this module, a stack of 3 bricks with 3/8″ (10 mm) mortar joints is 8 in. × 8 in. square. Typical hollow concrete units are nominally 8 in. thick by 8 in. high by 16 in. long (203 mm × 203 mm × 406 mm). Typical hollow clay units are nominally 8 in. thick by 4 in. high by 16 in. long (203 mm × 102 mm × 406 mm). "Nominal" dimensions in masonry include the unit dimension plus a standard 3/8″ mortar joint.

1. Clay Units

Clay units are made from clay or shale (the rocklike form of clay). The material is finely ground, mixed together with minor additives to control efflorescence and porosity and to influence color, and then tempered with water (if required) into a mass of either stiff or sort consistency, depending on the process of forming to be used.

In the *stiff-mud* process, the most common manufacturing process, the prepared material is extruded through a die by means of augers. This produces a continuous ribbon of material, which is then scratched, scored, indented, or otherwise marked to produce a particular texture on the face surface prior to being "wire-cut" into individual units. In the *soft-mud* and *dry-mold* processes, the prepared material is formed directly into individual units under pressure in steel molds. In all processes, the units are then dried and subsequently fired.

The firing process sinters and partially fuses the minerals of the clay to produce a hard stable body having both strength and durability. Many of the clays and shales used to manufacture clay units burn to a red color, although some burn to a buff color. The manufacturer can influence color, however, by changing the atmosphere of the kiln, by adding coloring admixtures, or by applying a coloring material such as glaze to the surface.

Requirements for brick units are covered in several ASTM documents, including ASTM C62 (building brick), ASTM C216 (facing brick), and ASTM C652 (hollow brick).

2. Concrete Units

Concrete units are made from a mixture of Portland cement, sand, gravel, and water. If the sand and gravel used in their manufacture is natural crushed rock, the units are called normal weight units; if the aggregates are expanded shale or other low-density material, the units are called lightweight units. The engineering properties of both kinds of units are the same, as they both must satisfy the same ASTM standard specification; the lightweight units are less expensive to ship and construct, but they are more

permeable to water and have greater shrinkage than normal weight units.

The process of manufacture involves forming individual units under pressure in steel molds. Concrete units are self-hardening, but their rate of gain in strength and ultimate strength are influenced by their curing, which may be either by a low- or high-pressure steam (autoclave) process.

The normal color of a concrete unit is gray. Colored concrete units are produced either by adding a pigment to the concrete during the manufacturing process or by later surface application. A popular surface texture for concrete masonry units is the "split face" obtained by breaking or chipping the surface during the manufacturing process.

3. Stone

Although the world is dense with examples of load-bearing masonry structures built entirely of stone that have lasted hundreds or thousands of years, stone is very rarely used as a structural material in modern buildings. Stone is treated as a masonry veneer product, and generally need only support its own weight.

B. Mortar

Mortar is spread on the bed and head joint surfaces of masonry units to bond the units together, and, as importantly, to keep them apart. Structurally, mortar distributes gravity loads uniformly from one unit to the next, and increases the shear strength of the assemblage. Mortar also decreases the moisture permeability of the assemblage. Requirements for mortar are specified in ASTM C270 *Mortar for Unit Masonry*.

Conventional mortar is a mixture of cement, lime, and sand, with the proportion of cementitious materials to sand typically near 1:3. In some localities, a proprietary product called "masonry cement" is used to replace the Portland cement and lime in the traditional mix. In the Uniform Building Code (UBC), the use of masonry cement is prohibited in all but the lowest seismic zones in the United States.

Mortar is typically mixed on site, although recently the use of "ready-mixed" mortar has increased. Ready-mixed mortar includes admixtures for set-control, and may be used for periods in excess of 2.5 hr after initial mixing.

C. Grout

In the context of reinforced masonry construction, grout is a specialized concrete mix used to fill the cores of hollow masonry units or separation space in double-wythe construction to bond reinforcement to the masonry units and allow all to act together as a composite material, thus dramatically increasing the resistance of the masonry to bending and tension. In so doing, it also increases the compressive and shear strengths of the wall, decreases sound transmission through the wall, increases the fire resistance, and increases the thermal mass of the wall. Requirements for grout are given in ASTM C476 *Grout for Masonry*.

Grout is a mixture of Portland cement, a small amount of lime, sand, small coarse aggregate, and water. Admixtures are often used to (1) counteract the volumetric shrinkage of grout due to water loss to the surrounding masonry, (2) plasticize the liquid grout to allow placement in confined spaces and minimize the water content, and (3) retard the setting time to allow time for consolidation. Grout can be ready- or site-mixed and added by bucket or pump to masonry walls up to 5 ft (1.5 m) high (low-lift grouting). With the addition of special procedures for inspection and the cleaning of core spaces, it may be pumped into walls with pour heights as high as 24 ft (7.3 m) (high-lift grouting).

D. Connectors

Connectors are used to join structural elements together, for example, the two wythes of a cavity wall, two intersecting walls, a floor and its supporting walls, a wall and a pilaster, a frame and its infill wall, or a veneer and its backup. They may be broadly classified as ties, anchors, and fasteners. *Ties* hold the wythes of a wall together or connect a veneer to its backup. *Anchors* connect masonry walls at their intersections, or connect walls to their supports or to other structural members, floor systems, or roof systems. *Fasteners* generally connect nonstructural elements such as equipment or fixtures to masonry walls.

Historically, masonry consisted of only units and mortar; however, in modern masonry construction, connectors are an essential element of reinforced structural wall systems and veneer systems. They carry gravity, wind, and seismic loads, and must be designed with due consideration to corrosion and building movements. Commonly, steels used for connectors are galvanized or copper-clad. Stainless steel and bronze connectors are increasingly common in harsh, corrosive environments.

III. ELEMENTS AND BUILDING SYSTEMS

A. Categories of Masonry Construction

1. Unreinforced Masonry

Until the latter half of the 20th century, all masonry was unreinforced, with only a few notable exceptions. Today, unreinforced masonry is still common in low-rise buildings in zones of very low seismic activity. However, the

modern push toward thinner, lighter, and taller building systems has severely limited the applicability of unreinforced systems, and advanced the development of efficient reinforced systems. This is because unreinforced systems can carry little or no tension force without causing cracking and, ultimately, failure of the masonry. Early unreinforced structures relied upon the mass and thickness of heavy masonry systems to counteract tension forces and to ensure that effective lines of thrust due to gravity and lateral loads remained within the wall system. Even so, such structures have proven time and again to be hazardous and potentially unstable under strong seismic loads.

Recent trends in the interpretation of seismic risk suggest that potentially large-scale catastrophic results may be associated with a moderate seismic event in areas of relatively low seismic activity such as eastern North America. These developments in our understanding of seismic risk make the use of unreinforced masonry less acceptable than in the past.

2. Reinforced Masonry

While there are a few selected and fascinating case studies of the early use of steel reinforcement to improve the ductility and resistance to tensile and shear forces of masonry, wide scale use of reinforced masonry increased dramatically in North America following the 1933 Long Beach, CA earthquake. In reaction to the poor performance of unreinforced masonry buildings in that earthquake, building codes required the use of reinforcement in all but the lowest seismic zones.

The type and amount of reinforcement used varies with the demand on a component, but typically masonry is reinforced with a grid of both vertical and horizontal reinforcement to resist flexural tensile stress and shear stress, leaving the masonry units and mortar to carry the compressive stresses. Grout serves to increase the compressive resistance of masonry, and to bind the steel to the bricks or block to ensure composite behavior of the whole.

3. Prestressed Masonry

Prestressing is the process of applying an initial tensile force into bonded or unbonded steel reinforcement—thus an initial compressive force into the masonry. In a sense, it performs like unreinforced masonry in that cracking is prevented by the presence of preexisting compressive stresses in the brick or block, except that in prestressed masonry, the compressive force is achieved in relatively thin, light systems through the introduction of active forces rather than the addition of material to increase the thickness and mass of the masonry. Prestressed masonry is a relatively recent development compared to the development of similar systems in concrete structures in the 1950s and 1960s, and it is not yet widely used.

B. Elements

The basic elements used in structural masonry construction are walls, beams, and columns. Masonry floor systems, structural arches, and vaults are uncommon in modern masonry systems.

1. Walls

Because of the wide variation in the materials and geometry of masonry units and building systems, there are a wide variety of masonry wall systems in use throughout the world. The list of wall types below is limited to those in common use in North America and the U.K.

Single-wythe walls: Single-wythe walls are solid walls one masonry unit in width. This is the simplest of all wall constructions. Walls are typically hollow-unit, reinforced masonry, but unreinforced systems can be used with the provision of sufficient cross-walls or bracing to carry the lateral loads which thin, unreinforced walls cannot carry easily. Insulation is generally applied to the inside or outside of the masonry. Insulating the cores of hollow units is possible, but relatively ineffective in cold climates.

Composite or solid walls: Composite walls are solid walls consisting of two or more wythes of masonry joined together by interlocking masonry units (*headers*) or metal ties, and a completely mortar-filled vertical joint known as a *collar joint*. The two wythes may be of similar units, or of different units. Traditionally, the exterior wythe was built with *face brick*—a hard-fired, high quality brick—and the interior wythe or wythes used *common brick*, a less highly fired and irregular brick. A common contemporary system is to use clay brick units on the exterior wythe with concrete block on the interior wythe of the wall. Composite walls are commonly unreinforced, but may be reinforced.

Cavity walls: Cavity walls are constructed in much the same way as composite walls, except that the exterior wythe is separated from the interior wythe by an air space, with metal ties connecting the two. The interior wythe(s) act as the load-bearing section of the wall, while the exterior wythe acts as the building envelope. The cavity space generally contains some insulation, and is detailed with flashing and weep-holes to allow moisture that penetrates the exterior skin to exit the building.

Diaphragm walls: Diaphragm walls are a special type of cavity wall in which masonry headers or "webs" are used to connect the inner and outer wythes to achieve composite structural action in the resistance to lateral loads. This system is more common in Canada and the U.K. than in the United States.

2. Columns and Pilasters

Masonry columns carry primarily vertical loads, and may be solid, hollow, reinforced, or unreinforced. When a column is incorporated in a masonry wall, it is known as a *pilaster*. Wide columns that appear to be small wall segments are often referred to as *piers*, but the distinction is entirely subjective.

3. Beams and Lintels

A beam or lintel is a horizontal element used primarily to carry gravity loads across a span. All beams are reinforced, since horizontal unreinforced elements could not carry the flexural tensile stresses in a beam element. In masonry, a distinction is often made between beams, bond beams, and lintels. Beams are reinforced, rectangular section elements that carry floor or roof systems. Bond beams are horizontally reinforced elements that are integral to a wall system that act not so much as beams but as horizontal chords for floor or roof diaphragms, and as tension force collectors to distribute lateral loads to shear walls. Lintels are beams within walls that span door or window openings.

4. Arches and Domes

It is said that the great architect Louis Kahn, asked a brick what it wanted to be, and it told him "an arch." It is the ideal application for masonry, since an assembly of bricks is well suited to carry compressive stresses, and a well-proportioned arch has only compressive stresses when it is uniformly loaded. A dome accomplishes the same purpose in three dimensions. Ironically, masonry arches and domes are not in common use in modern construction for much the same reason that unreinforced masonry is falling from use: they rely on the thickness and mass of the masonry system to ensure that they remain in the desired state of full compressive stress. The move to lighter, thinner, longer-spanning arch systems has necessitated a turn to other materials than masonry for arches.

C. Structural Building Systems

1. Single-Story Buildings

The vast majority of load-bearing masonry buildings are single-story, single-wythe wall buildings bearing wood or steel roof trusses, where the masonry serves as the primary gravity and lateral load system. When walls are very tall and slender, they must be reinforced. The economy of this building system is evidenced by the vast number of schools, commercial, light industrial, and even residential buildings that utilize this method of construction.

2. Multistory Buildings

The advent of reinforced masonry construction has made possible the idea of economical multistory buildings, even 20–30 stories in height. The tallest contemporary unreinforced masonry tower is the Monadnock building built in Chicago at the end of the 19th century. It is 16 stories tall, and the walls at the base are nearly 6 ft (1.8 m) thick. Very tall masonry buildings remain uncommon, primarily because the structural system is best laid out in the form of multiple shear walls and cross walls or "cells" which do not allow a great deal of freedom or openness in the architectural plan.

3. Foundation and Retaining Walls

In addition to the building systems identified above, masonry is commonly used in foundation systems such as frost walls, basement walls, or site-retaining walls. *Cantilevered walls* are reinforced vertically and cantilever up from the footings. *Counterfort walls* span (and are thus reinforced) horizontally between perpendicular wall supports on the exterior of the wall. When the vertical wall supports are on the interior, opposite the retained earth, the walls are known as *buttressed walls*. Basement walls may span vertically from a slab on grade to the first floor diaphragm.

D. Non-Structural Building Systems

1. Masonry Veneer

Masonry veneer is a nonstructural cladding tied to a backup system with metal ties, typically with an air space of 1 or 2 in. (25–50 mm). The backup may be masonry (typically concrete block), wood, or light-gage steel, and serves to carry wind loads to the foundation and floor or roof diaphragms. Veneer systems are not expected to carry load, so they must be detailed carefully to ensure that load is not inadvertently transferred from the primary structural system through motions associated with elastic deformation, creep, moisture movement, or thermal movement. The backup must also be carefully designed, not only for strength, but also to limit deformations of the veneer that may cause cracking and subsequent water penetration and corrosion.

Veneer is typically supported on a foundation, and, for buildings more than three stories tall, on intermediate structural supports. Veneer may be *anchored* as described above, or *adhered* directly to the backup.

2. Prefabricated Veneer

Tall buildings often incorporate prefabricated veneer panels that are constructed off-site and lifted into place on the structure. Such panels are invariably reinforced, and often are designed to carry wind loads to the support points.

E. Other Applications

Masonry is used in numerous other, generally nonstructural applications. Fireplace chimneys are still routinely constructed of masonry, with firebrick adjacent to the fire, and stone or brick masonry on the exterior. While the masonry itself is nonstructural, it must be designed to carry seismically induced inertia loads—falling chimneys are a well-known hazard in high seismic zones. Masonry is also often utilized to provide thermal mass in applications such as passive solar energy collection where dramatic temperature fluctuations can be "damped" by the thermal mass of the brick or stone. Masonry is naturally suited to act as a firebreak and a sound damper as well, and is often used in schools for this reason.

IV. CONSTRUCTION

A. General

Masonry derives a great deal of its economy from the very nature of its construction. It is typically made of small, locally produced units (brick or block) that may be handled manually, and mortar or grout made of common, locally available materials. Reinforcing generally consists of small, straight pieces with a minimum of required manufacturing. No formwork is required, and the labor force need be skilled, but not excessively skilled. In recent times, the relatively low cost of manufacturing and shipping in large quantities has resulted in the use of masonry units from a single source over a radius of 1000 miles and more.

B. Workmanship

Perhaps more than any other construction material, the profound effect of the quality of workmanship on the final quality of masonry is universally recognized. Workmanship includes numerous factors that can influence the strength, permeability, and durability of the product, among them:

1. Incorrect proportioning and mixing of mortar
2. Use of excessively thick bed joints
3. Incomplete filling of bed joints
4. Failure to fill vertical joints with mortar
5. Bridging of mortar across cavity spaces
6. Disturbance of units after laying
7. Improper construction and preparation of the grout space
8. Improper mixing, placement, and consolidation of grout
9. Improper placement of reinforcement
10. Failure to build walls plumb and level
11. Failure to protect new work from adverse climatic effects

Building codes address this situation in several ways, specifically:

1. Through the specification of special inspection requirements for particular types of masonry construction
2. Through the imposition of severe limitations on allowable stresses in work which is not inspected by an independent party during construction

The more refined and sophisticated the design or design method, the greater is the reliance on accurate and careful workmanship. Consequently, the criterion for a well-designed building is not only that it be efficient and strong in a material sense, but also that the design details be simple and easily constructed with good workmanship under average job-site conditions. A building design that will only perform well if constructed with excellent workmanship performed under ideal conditions will not be successful in any building material, but especially so for masonry buildings.

C. Mortaring

The strength, quality, and uniformity of the application of mortar can have a significant effect on the strength and performance of masonry, although this is primarily true for brick masonry and less so for grouted hollow masonry. Incompletely filled joints have an adverse affect on the flexure and shear strength of masonry, and increase the water permeability of the assembly. Doubling the joint thickness can result in as much as a 30% decrease in compressive strength of brick masonry.

In grouted masonry, if the grouting is of good quality, the physical properties and quality of the mortar joints have much less influence on the behavior of masonry than in brick masonry. Mortar which protrudes from the joints should be struck off, as the "mortar fins" can adversely affect the quality of the grout or the water permeability of the masonry.

D. Grouting

Proper grouting of masonry is essential to good performance, but it can be difficult to achieve. To achieve good bond of reinforcement to the surrounding masonry:

1. The grout space must be clean and free of mortar fins and obstructions.

2. Grout should be very fluid relative to concrete of like strength to allow penetration into confined spaces and to help to counteract the loss of water due to absorption into the surrounding brick or block.
3. Reinforcement should consist of relatively small bars distributed uniformly rather than bundled at the ends of walls.
4. Clean-out holes should be provided at the base of walls where high-lift grouting methods are used to ensure that a good bond of grout the base of the wall can be achieved.
5. Consolidation by mechanical vibration should occur soon after filling cells with grout. A second consolidation within 15 min of the first is often recommended for concrete masonry, but may be detrimental to the grout quality in highly absorptive hollow clay masonry.

V. PROPERTIES

A. Compressive Strength

When brick masonry fails in compression, the failure actually occurs by tensile splitting of the brick. This is the natural failure mode for unconfined brittle materials in compression, but in masonry the natural lateral expansion is magnified by the lateral expansion of mortar in the bed joints due to Poisson's effect. As a result, the compressive strength of the assemblage is typically less than that of the brick alone, but greater than that of the mortar alone. The compressive strength of the mortar itself has less affect on the compressive strength of masonry than might be expected. Hollow unit masonry fails in a similar fashion, with splitting of the webs occurring in the center of the units. In grouted masonry, the lateral expansion of the grout induces additional tensile stresses in the surrounding block or brick—the net area strength of grouted masonry is typically therefore less than the net area strength of ungrouted masonry even if the gross area strength is much higher.

B. Tensile Strength

In most practical situations, the tensile strength of masonry is adequately described by reference to tensile stress either normal to or parallel to the bed joints.

For tension normal to the bed joints, the tensile strength of ungrouted masonry is controlled entirely by the bond strength between the mortar and the unit: cracking usually occurs at the interface between the mortar and the unit. The tensile strength of the masonry is significantly increased by the presence of grout, because the tensile strength of the grout is directly engaged.

For tension parallel to the bed joints, the tensile strength of stack bond masonry is approximately the same as that for tension normal to the bed joints. For running bond masonry, however, the failure crack usually follows the mortar joints around the units: tensile-bond strength is effective along the head joints, while shear-bond or shear-friction strength acts along the bed joint. As a consequence, the tensile strength parallel to the bed joint will usually be two to three times greater than that normal to the bed joint. The presence of grout in *partially* grouted walls may have little effect on the tension strength parallel to the bed joints, but *fully* grouted walls will have much-improved strength.

The bond strength of mortar to brick relates directly to the tensile strength in either direction. It depends upon the absorption properties of the masonry unit, the cement, lime, and water content of the mortar, and the general quality of the joint. Since this quantity is so workmanship dependent, and because it is numerically very small relative to the masonry compressive strength, tensile strength is often ignored in the design of masonry systems.

C. Shear Strength

The shear strength of masonry is also related to the strength and extent of bond between unit and mortar. It is improved by the keying of units and mortar that takes place at core spaces in the units where mortar is squeezed into the cores when the unit is laid, and it is a function of the amount of compressive force acting normal to the bed joints in the masonry. In grouted masonry, it is more a function of the shear strength of the grout than of the bond strength of mortar to the unit.

D. Movement

Masonry deforms under load and also expands and contracts due to effects such as shrinkage, moisture expansion, and temperature.

1. Elastic Deformation

The elastic deformation of masonry is typically assumed to be linearly related to applied stress (i.e., it obeys Hooke's law) for low levels of stress. Traditionally, the elastic modulus of masonry is in turn assumed to be linearly related to strength. For many years, the relationship $E = 1000\ f'm$ has been assumed, (where $f'm$ is the design compressive strength of masonry), but recently it has become clear that the relationship $E = 750\ f'm$ is more appropriate for

concrete masonry, and E = 500–600 f′m is appropriate for clay masonry. The scatter in the experimental data is great, considering the wide variety of materials and testing conditions used for masonry tests.

2. Creep

The magnitude of creep in masonry depends primarily on the applied stress level and is influenced by the creep characteristics of is components. Clay bricks do not exhibit significant creep at the stress level typical in masonry; concrete blocks creep considerably more, and mortar also creeps. Creep strain for clay masonry is slightly less than its corresponding elastic strain, while the creep strain for concrete masonry is in the range of two to three times its corresponding elastic strain.

3. Shrinkage and Expansion due to Moisture Variation

Concrete units and mortar shrink. Values for shrinkage range from 0.02% to 0.07%, depending on the material composition of the unit and the curing process. In general, lightweight concrete blocks have significantly larger shrinkage than normal weight blocks, and autoclaved blocks exhibit much less shrinkage than low-pressure steam cured blocks. Values of shrinkage for mortar also cover a large range, from about 0.01% for low-flow, low water content mortar to 0.08% for high-flow, high water content mortar.

Whereas vertical shrinkage of concrete units in the wall usually results in vertical shrinkage of the wall itself, horizontal shrinkage of concrete units in the wall tends to result in vertical cracking rather than overall horizontal wall shrinkage.

Clay bricks typically expand due to freezing and moisture ingress. The latter expansion can continue for several years. Therefore, although the mortar may shrink, clay masonry is also likely to expand within the range of 0.025–0.20%.

4. Thermal Movements

It is usually assumed that expansion of masonry due to thermal effects is equal to the sum of the thermal expansions of its component materials. Typical values of coefficients of thermal expansion are $5–7 \times 10^{-6}/^{\circ}C$ for clay bricks, $6–13 \times 10^{-6}/^{\circ}C$ for concrete blocks, and $13 \times 10^{-6}/^{\circ}C$ f or mortar.

Whereas vertical thermal contraction of units and mortar would typically result in vertical wall contraction, horizontal thermal contraction of units and mortar tends to induce fine cracking in the bed-joint mortar and between the units and the head-joint mortar.

VI. STRUCTURAL DESIGN

An in-depth treatment of modern methods of structural design in masonry is well beyond the scope of this article. The next sections qualitatively describe the historical context and current methods of design used in modern engineered construction.

A. Historical Context

Masonry finds itself in a relatively unique position in the engineering community, particularly in reference to the most common structural materials—steel and concrete. Unlike the latter two materials, masonry has been in widespread use in some form for thousands of years. The builders who constructed the pyramids, the Roman aqueducts, and the great cathedrals all did so without the help of structural engineers, using primarily empirically derived rules of thumb in the proportioning of their structures. Despite the changing nature of masonry in the late 20th century—it is thinner and lighter than 100 years ago, and built with products like hollow concrete block that did not exist at that time—there remains a general sense that methods of design and construction that have worked in the past should serve well now. It is further assumed that, unless special measures for inspection are taken, the quality of masonry is difficult to predict due to the personal nature of its assembly (see Section IV.B above). Both of these things have hindered the advancement of masonry as an engineered material.

Because of this historical context, masonry is unique in that building codes still contain prescriptive rules of construction for particular applications. Furthermore, "engineered masonry" still uses primarily the Working Stress Design Method (defined in the following section), an approach originally developed for reinforced concrete structures, but abandoned in favor of Ultimate Strength Design Methods for concrete in the 1960s. The first inkling of Ultimate Strength Design entered the United States masonry codes in the 1991 UBC with the introduction of provisions for slender wall design.

Now there are provisions to allow the limited use of Ultimate Strength Design for masonry, with traditional design methods still forming the core of accepted engineering practice. Development of future codes is clearly moving in the direction of limit states design. This development has also been accelerated in areas of high seismic activity.

Masonry is also unique in that the above trends have followed different courses at different rates throughout the world. Unlike other standardized construction materials, masonry materials and construction practice are strongly influenced by local history and traditions. In New Zealand,

for example, an Ultimate Strength Design approach has been accepted in the codes since the early 1990s. In other parts of the world, "engineered" masonry still does not exist, and the majority of construction is carried out based on oral tradition and rules of thumb.

In the following, discussion is directed primarily toward reinforced masonry.

B. Working Stress Design

In Working Stress Design (WSD), reinforced masonry is analyzed under service loads, and stresses are limited to some level based on the stresses at the elastic limit divided by a safety factor. Materials are assumed to behave linearly elastically. The fundamental assumptions are

1. Plane sections remain plane during bending (shear deformations are negligible compared to bending strains).
2. Masonry carries no tensile stress.
3. Masonry is assumed to carry all compressive stresses, while steel carries all tensile stresses.
4. Stress is proportional to strain which is proportional to distance from the neutral axis.

The WSD method is effective and has guided masonry design since the mid 1930s when reinforcement was first introduced into masonry structures. As of this writing, it remains the primary accepted practice for masonry design in North America. It is the methodology adopted by the Masonry Standards Joint Committee (MSJC) in the development of the Building Code Requirements for Masonry Structures (ACI 530/ASCE 5/TMS 402) and also for the bulk of the UBC.

C. Ultimate Strength Design

In the 1940s and 950s, a method for Ultimate Strength Design (USD) was introduced into the structural engineering community. In this method, the behavior of structural elements is evaluated at their ultimate failure state rather than under service loads. Design loads are factored up to ultimate strength levels in proportion to the degree of uncertainty in the load, and material strengths are similarly reduced in proportion to the level of confidence in the material strength, and in proportion to the consequences of failure in certain modes. This method acknowledges the nonlinear behavior of materials and the importance of applying safety factors in proportion to the degree of uncertainty for specific materials loading conditions. Reinforced concrete has been designed exclusively by USD techniques since the 1960s. Masonry has only recently introduced USD methods, and only in the UBC.

Basic assumptions of mechanics for USD are similar to those for WSD; they are identical to those listed above for WSD, with the exception that stress is not assumed to be proportional to strain. Also, sections are proportioned to ensure that at ultimate load levels, the steel reinforcement has exceeded its yield strain, and the masonry has just reached a state compressive failure, i.e., the compressive strain in masonry at the moment of crushing is assumed to be equal to 0.003.

D. Empirical Design

Empirical design methods for masonry are allowed only in areas of low seismic risk (zones 0 and 1 in the UBC) and where basic design wind speeds are less than 80 mph. Masonry elements are proportioned primarily by prescriptive requirements for maximum heights, thickness, and slenderness ratios. Reinforcement is not required.

In the UBC, a special appendix is provided for using empirical rules for masonry buildings in zones 0, 1, or 2 with basic wind speeds between 80 and 100 mph. Such buildings must be reinforced.

E. Seismic Design

The most significant advancements in masonry design and construction in the latter half of the 20th century have been driven by the poor performance of unreinforced masonry structures in earthquakes. A significant proportion of the experimental and analytical research effort in masonry since the late 1930s has been devoted to improving the understanding of the seismic response of masonary buildings and improving the seismic design philosophy. Building codes for high seismic zones now incorporate provisions intended to ensure that masonary structural elements behave in a ductile manner—absorbing seismic energy rather than resisting it—while maintaining their capacity to carry gravity loads.

F. Crack Control

In the previous section, the movement characteristics of masonry due to moisture and thermal variations were discussed. While such cracks are rarely a structural hazard, prevention of unsightly cracking is addressed in the structural design by the judicious placement of expansion joints and control joints in masonry walls. Both vertical and horizontal movements must be considered. Control joints to control shrinkage cracking in concrete msaonry are typically placed at a maximum spacing of about 30 ft (9 m).

VII. TESTING

The testing of component materials and small masonry assemblages is essentially a quality assurance exercise; the testing of large- or full-scale masonry assemblages is essentially a research exercise, aimed at improving the understanding of the behavior of masonry and advancing the building codes. Only the most commonly used quality assurance testing procedures are highlighted here; these and other test methods are described in detail in standards of the American society of Testing and Materials (ASTM).

A. Masonry Units

Masonry units are typically tested for compressive strength, initial rate of absorption, saturation coefficient, and moisture content.

The compressive strength of masonry units is normally determined on full-size units, and is expressed in the form of either gross area strength (solid units) or net area strength (hollow units).

The initial rate of absorption, IRA, for clay units is measured by the change in weight of a unit that occurs due to water absorption when the bedding surface of the unit is placed in contact with water to a specified depth for one (1) minute. The test provides a comparison of the rates at which different clay bodies absorb water from mortar spread on them, and is related to the shear and tensile bond strength between the mortar and the unit as well as the potential difficulty of working with a particular unit in the fields. Units with very high IRA values are often pre-wetted prior to being laid in a wall.

The saturation coefficient for clay units, also referred to as the cold-to-boiling or C/B ratio, is calculated as the ratio of the moisture content of the unit after a 24-hr soak in water to that after a subsequent 5-hr soak in boiling water. The test relates indirectly to the capacity of saturated units to accommodate freezing temperatures.

The moisture content of concrete units is quantified as the amount of water retained in a unit after its manufacture and curing are complete. A unit is weighed, dried at a specified temperature for a specified length of time, and reweighed, the difference in weight being expressed as a percentage of the weight of the dry unit. Units having high moisture content are unsuitable for use in a dry climate, being prone to drying shrinkage after having been placed in the wall, which results in cracking of the wall.

B. Fresh Mortar and Grout

Fresh mortar is not routinely tested during construction. Grout, on the other hand, is often tested for consistency using the standard slump test. The apparatus consists of a mold in the form of a truncated cone and a tamping rod. The mold is filled in layers and tamped. The cone is then removed and the measurement of slump is the resulting drop in height of the grout. In order to improve the placement of grout in congested spaces, a slump in the range of 66–80% of the slump cone height is often specified, a range in which the slump test is not particularly accurate.

C. Hardened Mortar and Grout

Mortar and grout are routinely tested for compressive strength. Mortar is tested in cylinders or cubes formed in imprevious molds.

In recognition of the fact that much of the water in grout is absorbed into the surrounding masonry units—thus decreasing the water cement ratio, a key determinant of concrete strength—grout samples are formed on site with four hollow masonry units forming a rectangular prism. Mortar is subjected to the same absorption in place, but the mortar compressive strength test is recognized as a relative quality control measure rather than an indication of actual masonry assemblage strength.

D. Masonry Assemblages

The compressive strength of masonry is determined by the testing of prisms, normally five units high for brickwork of standard size bricks, and two units high for concrete masonry of standard size blocks. When testing is being performed to establish or verify the compressive strength of masonry for the purpose for design, the prisms are built in the same bond pattern (stack or running bond) with the same type of units and mortat (and grout) and brick-laying technique (full-bed or face-shell mortaring) as will be used in the actual construction.

VIII. FURTHER STUDY IN MASONRY

The information presented in this article provides an overview of masonry materials, construction methods, and design techniques used primarily in North America. Further study may be pursued through the literature listed in the References at the end of the article. Active societies for the education, research, and dissemination of information about engineered masonry include The Masonry Society, The International Masonry Institute, the National Concrete Masonry Association, The Brick Institute of America, and the Masonry Association of California and Nevada. Engineering Journals for masonry include *The Masonry Society Journal*, published by the Masonry Society, and *Masonry International*, published

by the University of Edinburgh, Edinburgh, Scotland. National and International conferences on masonry are held regularly, most notably the North American Masonry Conferences, The Canadian Masonry Symposia, and the International Conferences on Brick and Block Masonry.

SEE ALSO THE FOLLOWING ARTICLES

CONCRETE, REINFORCED • EARTHQUAKE ENGINEERING • FOUNDATIONS • FRACTURE AND FATIGUE • MECHANICS OF STRUCTURES • SANDWICH COMPOSITES

BIBLIOGRAPHY

Masonry Standards Joint Committee, (1999). "Building Code Requirements for Masonry Structures." ACI 530/ASCE 5/TMS 402.

ASTM Committees. Annual Book of ASTM Standards, American Society for Testing and Materials, Philadelphia.

CSA Committee S304 (1984). "Masonry Design for Buildings" (CAN3-S394-M84). Canadian Standards Association. Rexdale, Ontario, Canada.

Hendry, A. W. (1981). "Structural Brickwork," MacMillan Press, London, England. International Conference of Building Officials (1997). Uniform Building Code, Whittier, CA.

Drysdale, R. G., Hamid, A. A., and Baker, L. R. (1994). "Masonry Structures: Behavior and Design," Prentice Hall, Englewood Cliffs. NJ.

Amrhein, J. E. (1994). "Reinforced Masonry Engineering Handbook, Clay and Concrete Masonry," 5th ed., Masonry Institute of America, Los Angeles.

Ekwueme, C. G., and Uzarski, J. (2000). "Seismic Design of Masonry Using the 1997 UBC," Concrete Masonry Association of California and Nevada, Citrus Heights, CA.

Brandow, G. E., Hart, G. C., and Virdee, A. (1997). "1997 Design of Reinforced Masonry Structures," Concrete Masonry Association of California and Nevada, Citrus Heights, CA.

Paulay, T., and Priestley, M. J. N. (1992). "Seismic Design of Reinforced Concrete and Masonry Buildings." John Wiley & Sons, New York.



Mechanics of Structures

Robert A. Heller
Virginia Polytechnic Institute and State University

I. Structural Mechanics
II. Structural Loads and Stresses
III. Material Behavior and Constitutive Relations
IV. Tensile and Compressive Structures
V. Trusses
VI. Beams and Frames
VII. Two- and Three-Dimensional Structures
VIII. Thin Shells and Space Frames
IX. Structural Safety and Reliability
X. Structural Failures

GLOSSARY

Bending moment Moment that causes curvature in a beam or any structural element.
Brittle Material that breaks with only a small amount of deformation.
Buckling Sudden, large bending deformation.
Compression Force or stress that results in shortening.
Creep Time- and temperature-dependent deformation.
Critical load Load that produces buckling of columns.
Dead load Weight of the structure.
Determinate Structure for which the number of unknown quantities is equal to the number of equilibrium equations.
Ductile Material that can undergo large deformations.
Dynamic load Impact or oscillating load.
Elastic Material that returns immediately to its original shape once loads are removed.
Fatigue Deterioration of structural strength under repeated loading.
Indeterminate or redundant Structure for which the number of unknown quantities exceeds the number of equilibrium equations.
Live load All applied loads other than dead load.
Modulus of elasticity Initial slope of the stress strain diagram; a measure of material stiffness.
Moment Force times distance from a fulcrum.
Moment of inertia Measure of the distribution of cross-sectional areas.
Neutral surface Surface in a beam that remains unstressed.
Parallel system System that fails only if all of its components fail.
Plastic Material that acquired a permanent deformation.
Probability of failure The chance that applied stress exceeds strength.

Safety factor Ratio of strength to applied stress.
Series system System that survives only if all of its components survive.
Stable Structure that does not change its shape significantly under varying loads.
Static load Stationary or slowly applied load.
Strain Change in length per unit length.
Stress Internal force per unit area.
Stress concentration factor Larger than unity multiplier of the average stress.
Stress-corrosion Corrosion accelerated by stress.
Structure Load-carrying part of a building, a bridge, or an airplane.
Tension Force or stress that results in elongation.
Thermal coefficient of expansion Change in length of a unit length material as a result of one degree temperature change.
Thermal strain Change in length of a unit length of material due to thermal expansion or contraction.
Thermal stress Stress that develops when thermal deformation is prevented.
Truss Triangulated structure.
Viscoelastic Material whose deformation depends on temperature and on the speed of load application.
Yield strength Stress beyond which a material is no longer elastic.

THE HUMAN BODY is kept erect against the gravitational pull of the earth by an internal skeleton of bones; a similar purpose is served by the hard outer shell of the lobster. Such "structures" also provide strength and stability to manmade systems. The beams, girders, and frames of a building or an airplane are internal structures while the outer shell of a large, covered sport stadium is an external structure.

I. STRUCTURAL MECHANICS

Structural mechanics, the study of force distribution among the various load-carrying members, is based on Newton's laws of equilibrium which require that the sum of all forces acting on a stationary structure or on any of its components must add to zero. In other words, buildings should not move. Buildings should not overturn either: This requirement is satisfied if the sum of all moments (forces multiplied by their distance from a point) also add to zero. Moving structures must obey similar, dynamic equilibrium laws.

For equilibrium (Fig. 1), the sum of vertical forces adds to zero

$$\sum F_y = 0$$

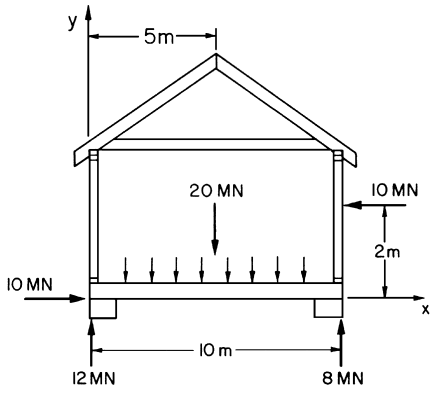


FIGURE 1 Equilibrium of forces and moments.

or

$$12 + 8 - 20 = 0.$$

The sum of horizontal forces adds to zero as follows:

$$\sum F_x = 0$$

or

$$10 - 10 = 0.$$

The sum of moments about the origin (forces multiplied by their perpendicular distances from the origin) also add to zero as follows:

$$\sum M_0 = 0$$

or

$$10 \times 8 - 5 \times 20 + 2 \times 10 = 0.$$

A structure is considered to be stable if load variations are not accompanied by large deformations or changes in shape. Instability may take several forms: Buildings may slide or overturn if not properly anchored to the ground, and a cable or chain changes its shape if the load is moved (Fig. 2a). A slender rod may bend outward, or buckle, under an axial compressive load (Fig. 2b), while a square structure consisting of four bars connected by flexible joints changes its shape into a rhombus under the influence of a horizontal load (Fig. 2c). Wind can produce large oscillations called flutter in a flexible structure such as a tent. Generally, structures should be stable or should be stabilized to reduce unwanted motion.

Cable-supported suspension bridges, for instance, are inherently unstable, yet when properly stabilized, they are favored for spanning large distances. The Tacoma Narrows Bridge in Tacoma, Washington, which was not stabilized, collapsed in 1940 as a result of wind-induced oscillations (Fig. 3).

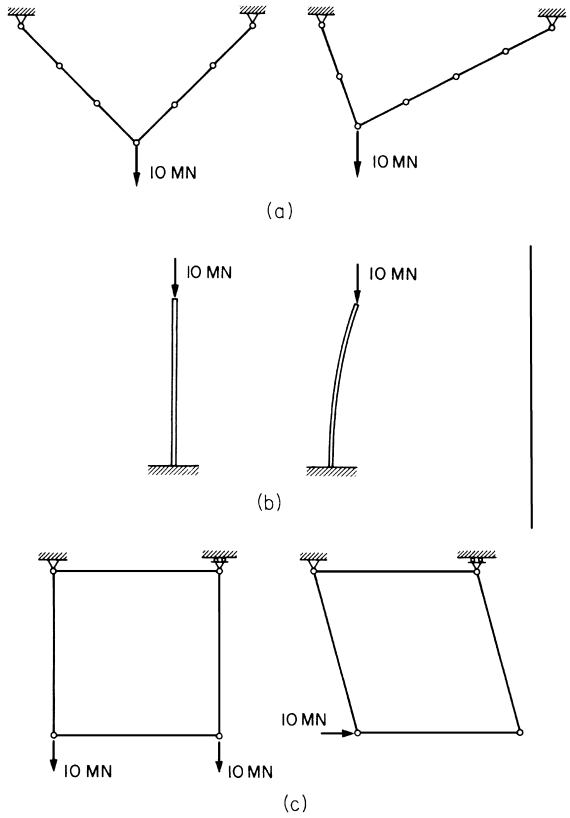


FIGURE 2 (a) Instability of cables. (b) Instability of columns. (c) Instability of structures.

II. STRUCTURAL LOADS AND STRESSES

As stated, the purpose of structure is to carry a variety of loads: the weight of its components, referred to as dead load, and applied live loads or payloads. Into the latter category belong all occupancy loads as well as those produced by the environment: wind, snow, earthquake, and temperature changes.

Some typical dead loads are listed in Table I, while a few important occupancy loads are presented in Table II.

TABLE I Weights of Some Building Materials

Material	Weight (N/m^2)
Concrete floor	230
Wood floor	150
Asphalt shingle roof	140
Clay tile roof	500
Brick wall	1950
Concrete block wall	2100
Stone wall	2700
Glass window	400

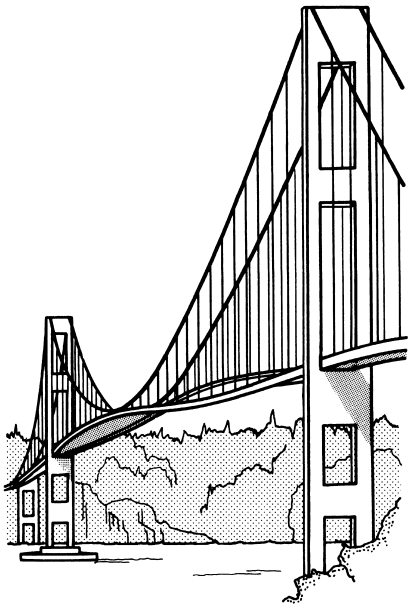


FIGURE 3 The Tacoma Narrows Bridge.

Wind loads may produce pressure on the windward side of a building and suction on the leeward side. Wind pressure/suction, P, measured in kilonewtons per square meter, is a function of the wind speed, V, in kilometers per hour, and a shape coefficient, C, which is of the order of 0.8–1.0 on the windward side and 0.6–0.8 on the leeward side. The pressure/suction is given by the following relationship:

$$P = 0.0000473CV^2.$$

Maximum wind speeds vary with the geography of the terrain as well as with the prevailing wind conditions as indicated on Fig. 4. Wind pressure is also influenced by the height of a structure. Higher buildings experience greater pressures. The devastation wreaked by hurricanes and tornadoes is all too obvious and structures must be carefully designed to withstand their pressures.

Snow loads are also terrain dependent as may be seen on Fig. 5. Roof geometry also influences the effects of snow; steeper roofs are less vulnerable than shallower ones.

TABLE II Occupancy Loads

Type of structure	Load (kN/m^2)
Balcony	4.8
Garage	2.5
Library	7.0
Office building	4.0
School room	2.0
Shop	5.0
Theater	3.0

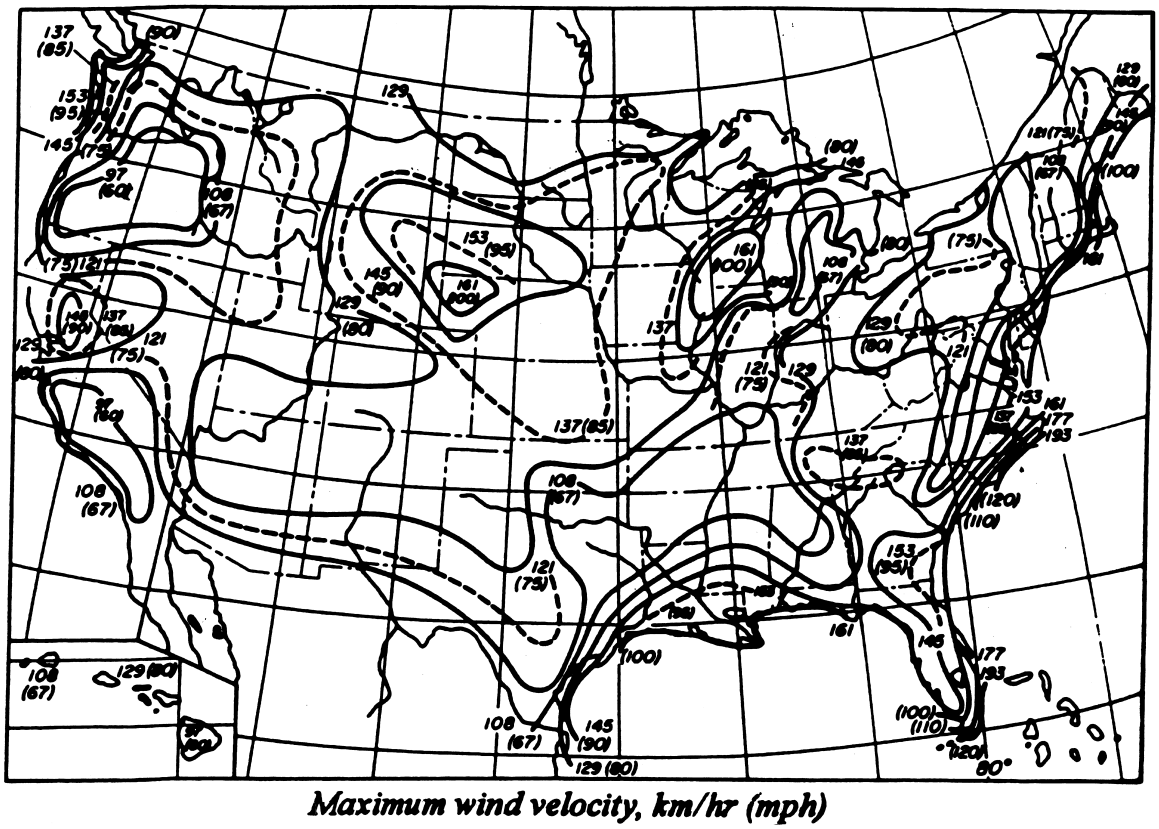


FIGURE 4 Maximum wind velocities in the United States. [From Salvadori, M. G., and Levy, □. (1981). "Structural Design in Architecture," reprinted by permission of Prentice Hall Inc., Upper Saddle River, NJ.]

While some of these loads are static or quasi-static (stationary and slowly applied), others may be dynamic or oscillatory in nature.

The intensity of dynamic loads (e.g., blast, impact, and earthquake) is considerably greater and hence more damaging than that of the same force applied slowly.

Recent catastrophic earthquakes in Turkey, Greece, and Taiwan are reminders of their destructive effects. During such tremors flexible, steel structures fared much better than unreinforced, brittle, concrete and masonry buildings. Current design of earthquake-resistant structures includes passive and active damping systems that reduce the intensity and the frequency of the transmitted disturbances. Areas of the United States at risk due to earthquakes are presented in Fig. 6.

Oscillating loads are dangerous because they are repetitious and can result in a gradual weakening of the structure.

The natural frequency of a structure is the number of times it swings from one extreme position to the other per second. For instance, the 410-m tall World Trade Center buildings in New York City have a natural frequency of 0.1 Hz (hertz, cycles/sec), that is, it takes 10 sec for its top to complete one cycle. (Inman)

If the frequency of the oscillating load coincides with the natural frequency, the structure may resonate. Under such conditions the amplitude of oscillations will increase significantly and may eventually destroy the structure. Wind-induced resonance was responsible for the destruction of the Tacoma Narrows Bridge (Fig. 3).

Applied forces produce internal stresses as well as deformations. A force P acting on a cross-sectional area A may be resolved into components perpendicular, P_n, and parallel, P_t, to the surface (Fig. 7a). Using trigonometric relations, we obtain the following:

$$P_n = P \sin \alpha$$

and

$$P_t = P \cos \alpha.$$

The perpendicular component results in a normal stress, σ_n, measured in newtons per square meter, also denoted as Pa (Pascal) as follows:

$$\sigma_n = P_n / A,$$

which is load per unit of cross-sectional area, while the force parallel to the cross section produces a shear stress,

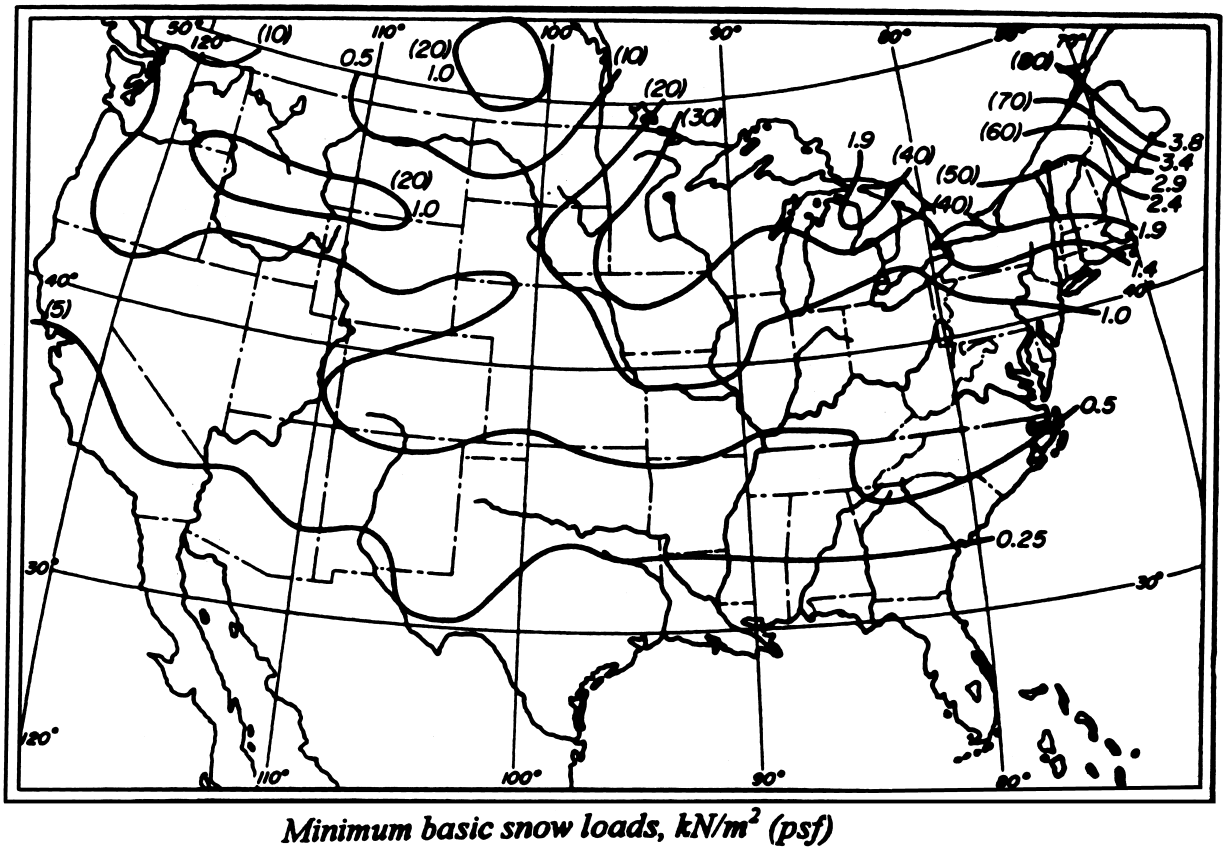


FIGURE 5 Snow loads on flat surfaces in the United States. [From Salvadori, M. G., and Levy, □. (1981). "Structural Design in Architecture," reprinted by permission of Prentice Hall Inc., Upper Saddle River, NJ.]

τ, as follows:

$$\tau = P_s/A.$$

Tensile normal stresses result in elongation and compressive normal stresses in shortening of structural members. The accompanying normal strain, ε, is defined as the change in length per unit length of material (see Fig. 7b) as follows:

$$\varepsilon = \Delta l/l.$$

Shear stresses change right angles into acute or obtuse angles, as seen in Fig. 7c. To maintain force and moment equilibrium, shear stresses must be equal and opposite on four sides of a unit cube of material. The resulting shear strain is the angle change, γ.

Tension, compression, and shear are basic states of stress. When only one of these stresses acts on a component, the state of stress is termed pure tension, pure compression, or pure shear. Such stresses are uniform over the cross section.

The three basic stresses may occur in combinations. The most important such case is bending. When a beam (Fig. 8a) supported on its ends is bent by vertical loads P, originally straight, parallel, horizontal lines become curved and parallel vertical lines between the applied loads form radial ones. The upper half of the beam has been shortened and is therefore in compression, while the lower half is extended and is in tension. The middle surface of the beam has not changed its length: It is a neutral surface.

The stress distribution varies linearly from maximum compressive stress on top through zero stress on the neutral surface to maximum tension at the bottom. These stresses, acting on the cross-sectional area of the beam, may be resolved into a tensile force, T, and a compressive force, C, which produce a moment, called bending moment, M, (in Newton-meters) as follows:

$$M = T \times e = C \times e,$$

where e is the distance between the resultant tensile and compressive forces (Fig. 8b). Cutting away a short segment of the beam near the left support, one sees that the bending moment is necessary to satisfy moment equilibrium requirements for the segment as follows:

$$T \times e = M = PX.$$

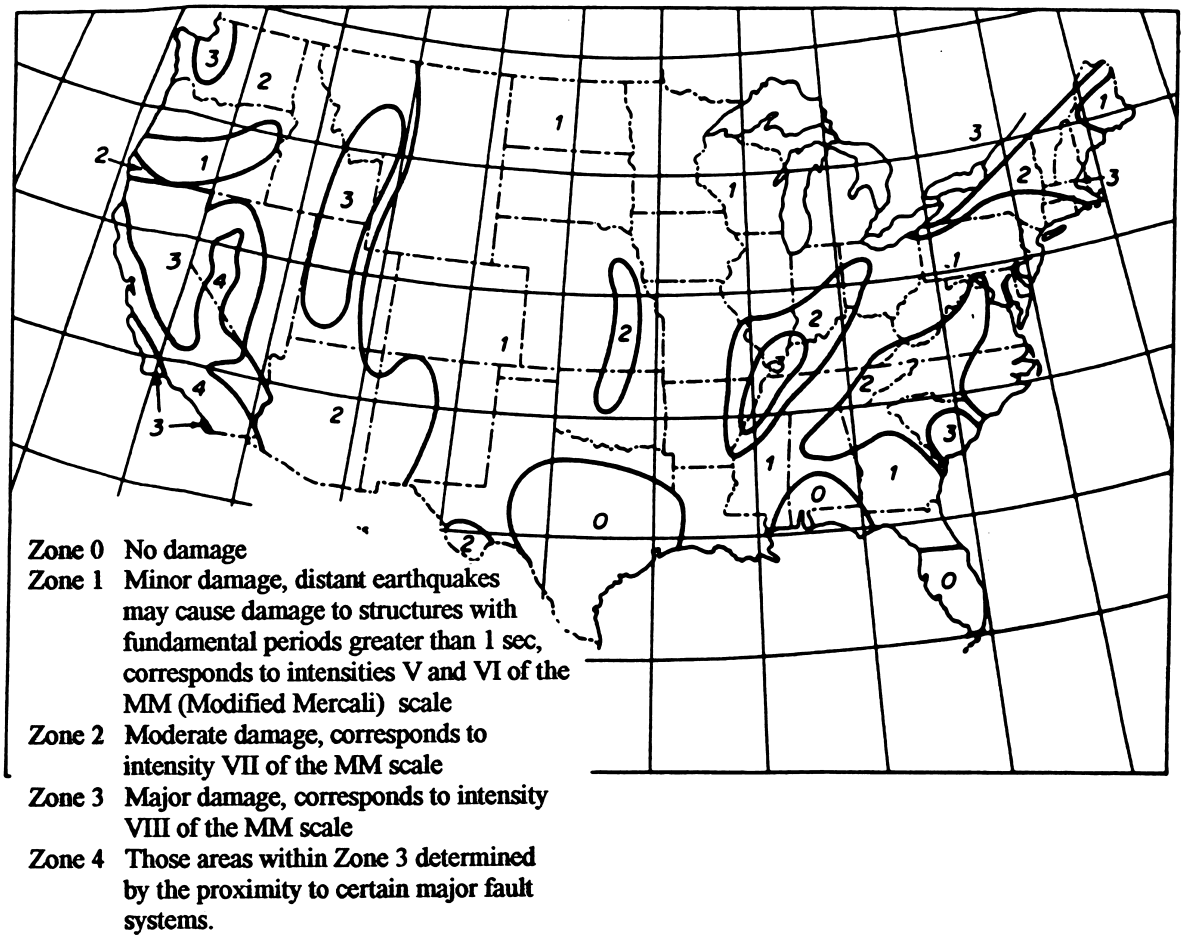


FIGURE 6 Seismic risk map of the United States. Zone 0: no damage; Zone 1: minor damage, distant earthquakes may cause damage to structures with fundamental periods greater than 1 sec, corresponds to intensities V and VI of the MM (Modified Mercali) scale; Zone 2: moderate damage, corresponds to intensity VII of the MM scale; Zone 3: major damage, corresponds to intensity VIII of the MM scale; and Zone 4: those areas within Zone 3 determined by the proximity to certain major fault systems. [From Salvadori, M. G., and Levy, M. (1981). "Structural Design in Architecture," reprinted by permission of Prentice Hall Inc., Upper Saddle River, NJ.]

A vertical force V is also required at the cut for force equilibrium as follows:

$$V = P.$$

This vertical force is a shear force.

It is seen that bending is a state of stress that combines tension, compression, and shear except in the case of pure bending where no shear is involved. Furthermore, in comparison with pure tension or pure compression, where the full cross-sectional area is uniformly stressed, bending is an inefficient state of stress because only the top and bottom fibers are fully stressed while those near the neutral surface are not.

To reduce this disadvantage, material may be removed from the beam cross section near the neutral surface (Fig. 8c); the remaining I-shaped beam is a frequently used structural form.

III. MATERIAL BEHAVIOR AND CONSTITUTIVE RELATIONS

All materials deform under the application of stresses. These deformations may be elastic (deformation disappears upon load removal) or inelastic (the shape change is permanent). Most metals exhibit elastic behavior for a given load range but deform inelastically for larger loads.

To examine the behavior of engineering materials, a tension test may be performed in a testing machine. A tensile normal load is applied to a specimen and is gradually increased; the resulting deformation is carefully measured.

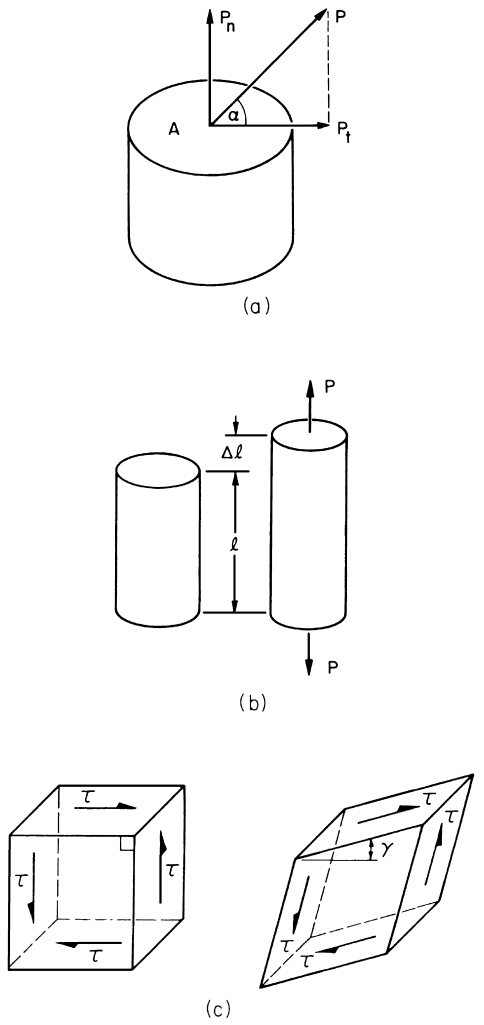


FIGURE 7 (a) Resolution of forces into normal and shear components. (b) Normal strain. (c) Shear stress and shear strain.

The applied load divided by the cross-sectional area (i.e., the normal stress) is then plotted against the normal strain, the elongation divided by the original length. The resulting curve is called a stress–strain diagram and is characteristic for a particular material (Figs. 9a and 9b).

In most metals the initial portion of their stress–strain diagram is a straight line; the material is linearly elastic. In this range the deformations disappear if the load is removed. With increasing stress, eventually the yield strength is reached beyond which elongations increase disproportionally. The material deforms plastically and remains permanently deformed. Finally, after considerable deformation, the specimen reaches its ultimate strength and breaks. Materials that can undergo considerable amount of deformation before failure are termed ductile.

Other so-called brittle materials such as concrete, stone, and glass, fracture with relatively small deformations.

The initial slope of the stress–strain curve, given by the following:

$$E = \tan\alpha,$$

is called the modulus of elasticity or Young's modulus for the material; the units are the same as those of stress (newtons per square meter).

For linearly elastic materials, a simple expression relates stress and strain. This simple constitutive equation, or Hooke's law, given by the following:

$$\varepsilon = \frac{\sigma}{E},$$

indicates that a high-modulus material deforms less than one with a lower modulus (Fig. 9c). More complicated,

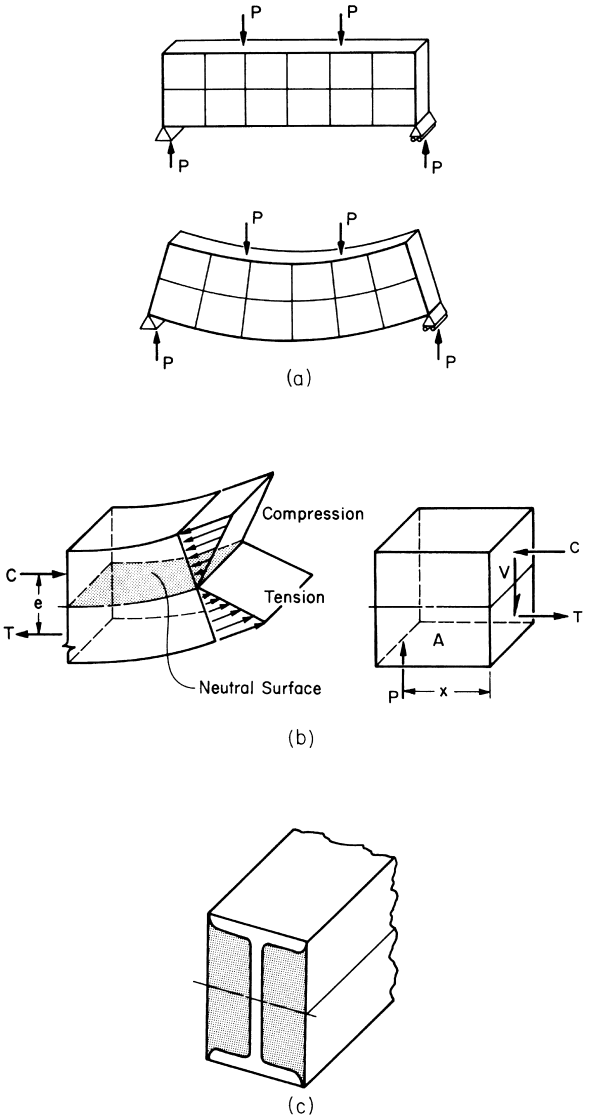


FIGURE 8 (a) Bending. (b) Bending and shear stresses in beams. (c) I-beam.

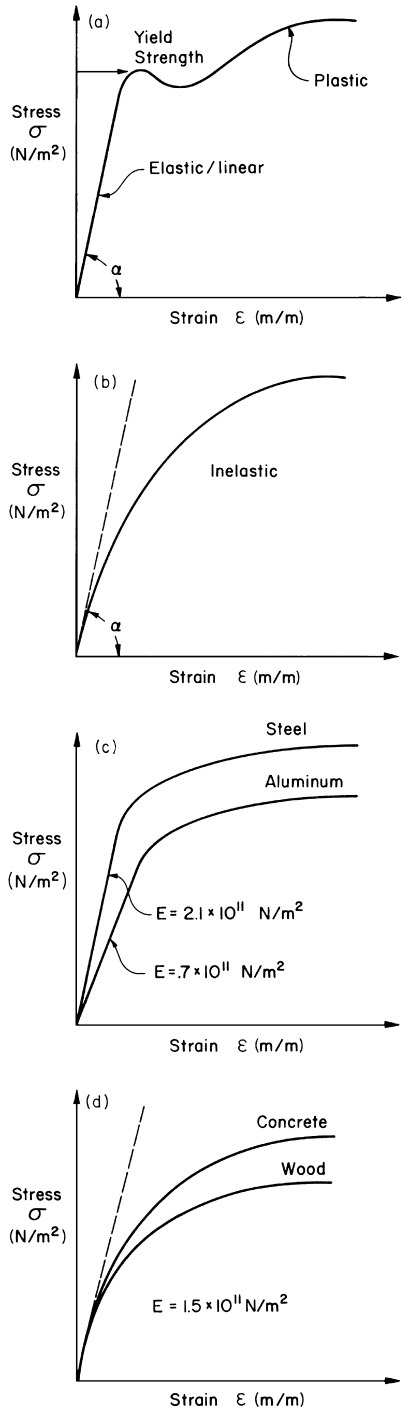


FIGURE 9 (a) Stress–strain curve for an elastic–plastic material. (b) Stress–strain curve for a nonlinear material. (c) Stress–strain curve for aluminum and steel. (d) Stress–strain curves for wood and concrete.

nonlinear, constitutive relations may be written for inelastic substances (Fig. 9d).

For structural materials, elasticity is a desirable property; deformations should disappear after load removal

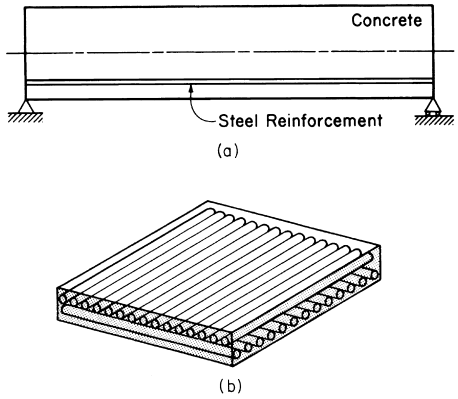


FIGURE 10 (a) Reinforced concrete beam. (b) Laminated composite plate.

lest the structure exhibit permanent changes in shape, though the plasticity of metals, with their ability to withstand large deformations without collapse, is more useful in earthquake-sensitive regions than brittle concrete.

While the majority or ductile metals are equally strong in tension and compression, brittle materials are usually much more resistant to compressive loads than to tensile ones. To improve the tension strength of structural members made of such a brittle material as concrete, for instance, the properties of two materials are combined. A reinforced concrete beam has steel rods cast into portions that are under tension stress, as seen in Fig. 10a. Other types of reinforcement are used to improve the stiffness characteristics of structural members by casting high-modulus fibers into a low-modulus matrix. The reinforcement is usually applied in layers with alternating orientations producing structural shapes with designed strength characteristics (Fig. 10b).

Composites such as graphite-epoxy, boron-epoxy, and carbon-carbon exhibit high modulus of elasticity as well as high strength and light weight. Because they are also resistant to environmental damage, they are used extensively in aircraft and space structures, sporting equipment, and even in buildings and bridges.

Components made of composite materials may be cast or molded into the desired shape without extensive machining and are, therefore, very economical.

IV. TENSILE AND COMPRESSIVE STRUCTURES

It has been stated that the most efficient utilization of materials occurs when stresses are uniform throughout the cross section of the structural member. Components carrying pure tension or pure compression are therefore frequently used to support large loads.

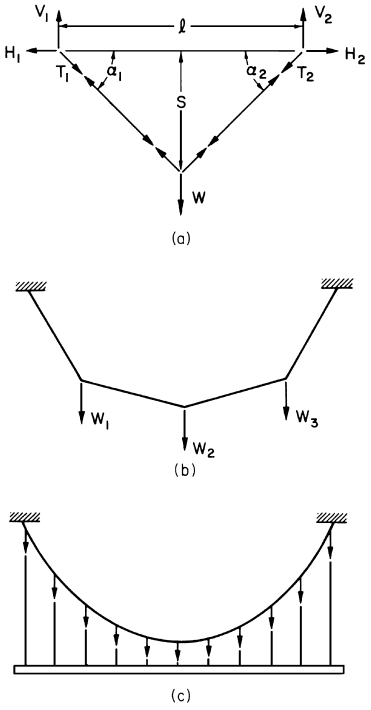


FIGURE 11 (a) Cable tension. (b) Cable under multiple loads. (c) Cable carrying a uniform load.

When a force is applied to a cable, the cable forms a V-shaped trough, both sides of which are in pure tension (Figs. 2a and 11a). At the supports, the inclined cable tension T may be resolved into vertical and horizontal components V and H, respectively. Equilibrium considerations require that the two vertical support reactions add up to the applied load W, while the horizontal, component of the reaction, called thrust, is the same at the two supports.

$$H = T_1 \cos\alpha_1 = T_2 \cos\alpha_2$$

where the angles α_1 and α_2 are indicators of the depth, or sag, of the cable. Because the cosine function increases with decreasing angle, a shallow cable requires larger horizontal thrusts than a deep one. This thrust must be absorbed by a strong anchorage. Additional loads change the shape of the cable, but segments between loads remain straight and indicate the presence of tensile stresses (Fig. 11b). This shape change is referred to as instability. As the number of loads increases, the straight lines between them become shorter, and the shape of the cable approaches a curve that for a uniformly distributed load is a parabola (Fig. 11c).

The high-strength steel cables of suspension bridges form such curves and are capable of spanning distances of over a mile. The inherent instability of cables requires that such large bridges be stabilized. One method used is the addition of stiffening trusses (Fig. 12a) that form part of

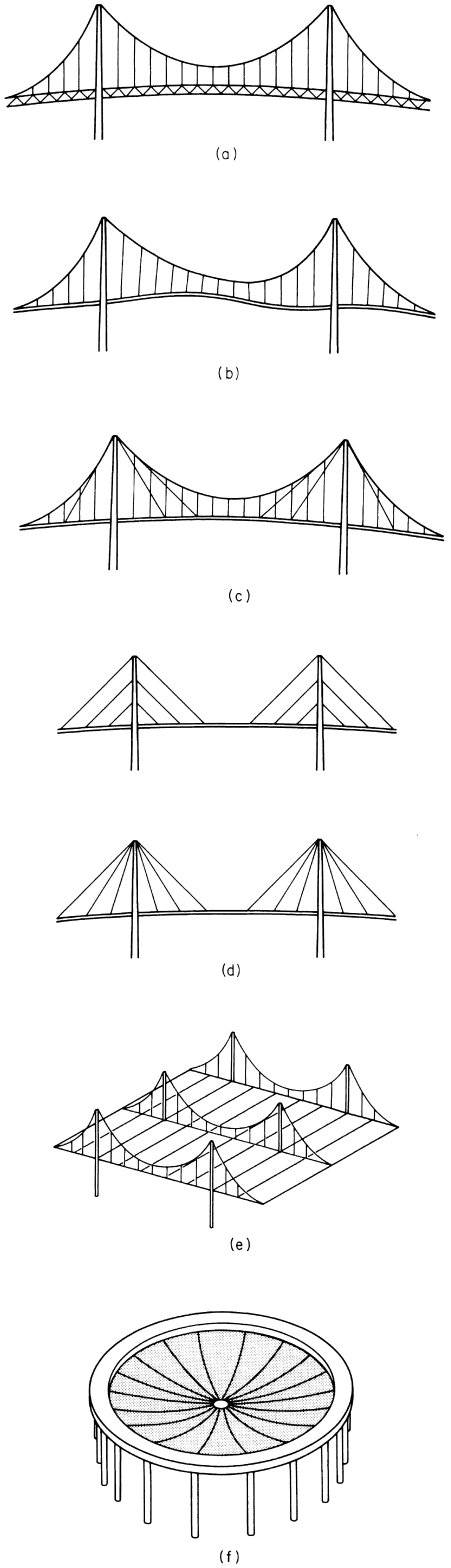


FIGURE 12 (a) Suspension bridge with stiffening truss. (b) Suspension bridge instability. (c) Cable stay stiffened suspension bridge. (d) Cable-stayed bridges. (e) Cable-supported rectangular roof. (f) Cable-supported circular roof.

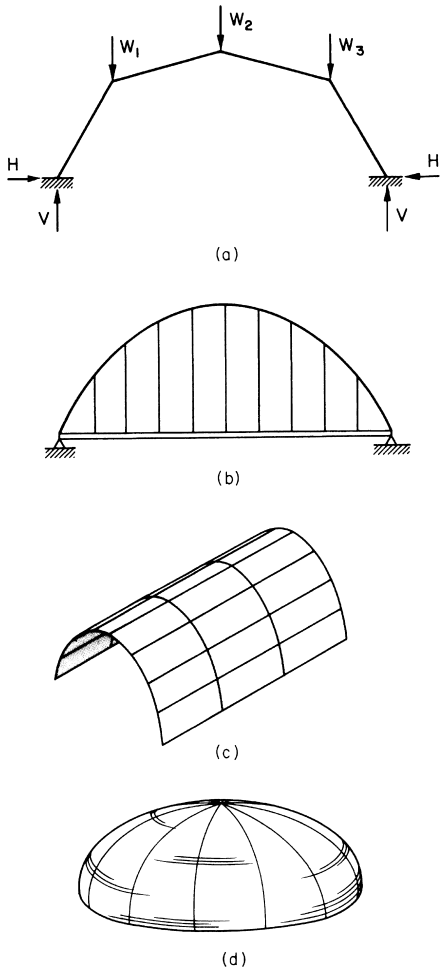


FIGURE 13 Arches for (a) concentrated loads and (b) uniform loads. Arch-supported roofs of (c) parallel and (d) radial arches.

the roadway. Another technique prevents the longitudinal swinging action of the cables (Fig. 12b) with the aid of diagonal cables (Fig. 12c) called stays. In a modern development of cable-stayed bridges with intermediate span lengths, the main cable is completely eliminated, and load-carrying and stabilizing functions are carried out by diagonal stays alone (Fig. 12d). Cables are also used in roof structures, in either parallel or radial arrangements. The two-dimensionality and stiffness of the roofing material have stabilizing effects (Figs. 12e and 12f).

An inverted cable-shape, called an arch, that is made of a stiff material carries loads by means of pure compression (Figs. 13a and 13b). Such structures are ideally made of masonry with high compressive strength. Here too a horizontal thrust H is developed, and heavy buttresses must be used to prevent the arch from opening up. As is the case with cables, a shallow arch produces greater thrust than one with a high rise.

Slender arches like slender rods have a tendency to buckle under compressive loads and are considered to be unstable structures. Arches are frequently used as bridge supports for moderately long spans and as roof structures. A set of parallel arches may support a vault while dome-like structures result when arches are arranged in a radial pattern (Figs. 13c and 13d).

V. TRUSSES

Another type of structure called an ideal truss consists of slender rods with flexible hinges on their ends. Such rods are two-force members (Fig. 14a); the two forces applied at the hinges always act along the longitudinal axes of the bars, either in pure tension or pure compression. When three rods are arranged in a triangular configuration, a stable structure results (Fig. 14b). A four-bar structure (Fig. 2c) was shown to be unstable.

It should be noted that at a hinged support, such as A in Fig. 14b, movement is restricted in two directions, horizontal and vertical; hence two reaction force components are developed. A roller support restricts motion only in the direction perpendicular to the rolling surface; therefore, at B in the figure, only a vertical reaction force is required.

The forces in the bars are calculated with the aid of equilibrium equations. Two equations may be written for each joint, one in the horizontal and one in the vertical direction (Fig. 14c). For instance, at joint A, the two bar forces AB and AC must be in equilibrium with the unknown reaction forces R_{AH} and R_{AV}. Therefore:

$$\sum F_{\mathrm{H}} = AC - AB\cos\alpha + R_{\mathrm{AH}} = 0$$

$$\sum F_{\mathrm{V}} = R_{\mathrm{AV}} - AB\sin\alpha = 0.$$

A total of six such equations may be written and solved at the three joints for the six unknown forces: the three reaction components and the three bar forces.

For the truss of Fig. 14b, if the number of bars, b, and the number of unknown reactions, r, equal twice the number of joints, j; that is:

$$b + r = 2j$$

or

$$3 + 3 = 2 \times 3,$$

there are exactly as many equations $(2j)$ available as unknowns. Such a truss is said to be statically determinate and stable. If the left side of the equation is less than the right, as is the case for the four-bar truss of Fig. 2c, then we have the following:

$$b + r < 2j$$

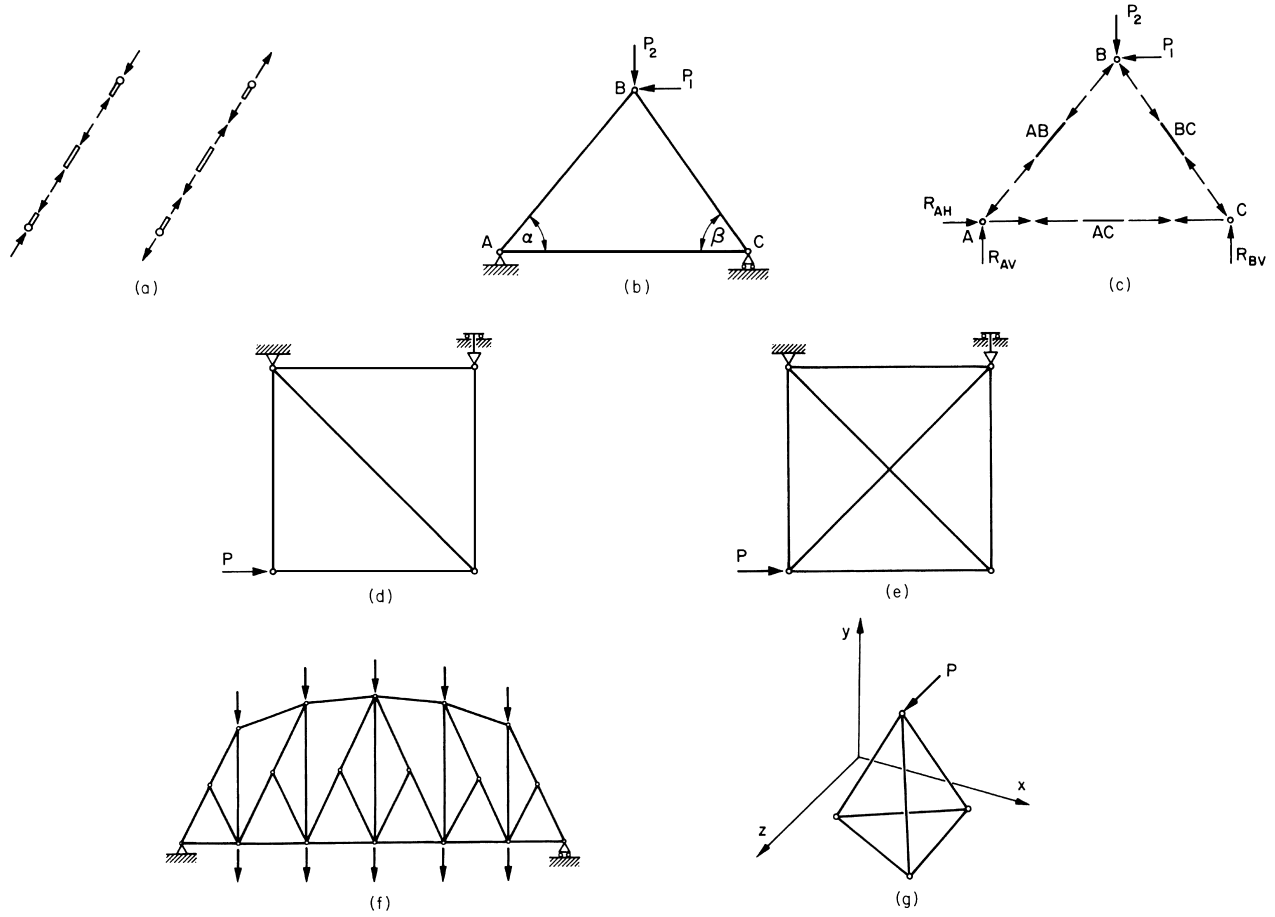


FIGURE 14 (a) Two force members. (b) Stable, triangulated structure. (c) Force distribution in a truss. (d) Statically determinate truss. (e) Indeterminate or redundant truss. (f) Two-dimensional truss structure. (g) Three-dimensional truss.

or

$$4 + 3 < 2 \times 4.$$

This structure is unstable because there is an insufficient number of bars. The addition of a diagonal will, however, stabilize the structure (Fig. 14d). On the other hand, a second diagonal (Fig. 14e) not only will stabilize the truss but will also make it statically indeterminate, or redundant. In this case,

$$b + r > 2j$$

or

$$6 + 3 > 2 \times 4.$$

There is one more unknown quantity than the available number of equations of equilibrium. To determine the forces in such a redundant structure, additional relations involving the flexibility of bars are required.

The six equations of equilibrium that can be written for the three-bar truss of Fig. 14b may be arranged in matrix form:

$$\underbrace{\begin{bmatrix} -\cos\alpha & 0 & 1 & 1 & 0 & 0 \\ -\sin\alpha & 0 & 0 & 0 & 1 & 0 \\ \cos\alpha & -\cos\beta & 0 & 0 & 0 & 0 \\ \sin\alpha & \sin\beta & 0 & 0 & 0 & 0 \\ 0 & \cos\beta & -1 & 0 & 0 & 0 \\ 0 & -\sin\beta & 0 & 0 & 0 & 1 \end{bmatrix}}_{[\mathbf{A}]} \underbrace{\begin{Bmatrix} AB \\ BC \\ AC \\ R_{\mathrm{AH}} \\ R_{\mathrm{AV}} \\ R_{\mathrm{B}} \end{Bmatrix}}_{\{\mathbf{X}\}} = \underbrace{\begin{Bmatrix} 0 \\ 0 \\ P_1 \\ P_2 \\ 0 \\ 0 \end{Bmatrix}}_{\{\mathbf{P}\}}$$

Denoting the matrix of the coefficients (i.e., of the direction cosines) by $[\mathbf{A}]$, the vector of unknown forces by $\{\mathbf{X}\}$, and the vector of applied loads by $\{\mathbf{P}\}$, the solution of the truss is obtained by solving the following matrix equation:

$$[\mathbf{A}]\{\mathbf{X}\} = \{\mathbf{P}\}.$$

This is accomplished by premultiplying both sides of the equation by the inverse of $[\mathbf{A}]$:

$$[\mathbf{A}]^{-1}[\mathbf{A}]\{\mathbf{X}\} = [\mathbf{A}]^{-1}\{\mathbf{P}\}$$

or

$$\{\mathbf{X}\} = [\mathbf{A}]^{-1}\{\mathbf{P}\}$$

While the above six equations may easily be solved for the six unknowns without resorting to matrix algebra, a more complicated system, such as the truss shown in Fig. 14f consisting of 33 bars, 18 joints, and three reactions, requires the solution of 36 simultaneous equations. This and much larger structural systems are, however, quickly analyzed with modern matrix methods and high-speed computers.

The trusses examined so far have been two dimensional. For a space truss (Fig. 14g), three equations of equilibrium are written at each joint: $\sum F_x = 0$, $\sum F_y = 0$, and $\sum F_z = 0$. Hence the determinacy–stability relation for three-dimensional trusses becomes the following:

$$r + b = 3j.$$

In practical trusses the hinges at bar ends are replaced by riveted gusset plates that induce some, usually small, secondary bending stresses in the structure. These are superimposed on the pure tensions and pure compressions in the idealized, hinged structure. It should also be noted that the compressed elements of trusses are prone to buckling and must usually be designed to prevent this form of instability. Trusses are extremely versatile structural elements and are frequently used in a great variety of applications. A trussed bridge is shown in Fig. 14f.

VI. BEAMS AND FRAMES

Beams are structural components that carry loads primarily by bending and shear stresses. They are the most frequently used elements for spanning relatively short distances.

Beams may be supported in a variety of ways. A cantilevered beam has one of its ends rigidly attached to a wall (Fig. 15a) and may not rotate at this point. A simply supported beam (Fig. 15b) usually has one hinged and one roller type support, which permit rotation, while a fixed-end beam (Fig. 15c) has both ends rigidly attached to walls. Combinations of supports and multiple supports are also possible.

Bending stresses in beams are greatest where the bending moment and hence the curvature is largest. For the cantilever and the fixed-end beam, this occurs at the built-in ends, while in the simply supported case stresses are greatest near the center of the beam.

As discussed above (Fig. 8b), bending stresses vary linearly through the depth of the beam from maximum compression at the top through zero at the neutral surface to maximum tension at the bottom. The triangular stress

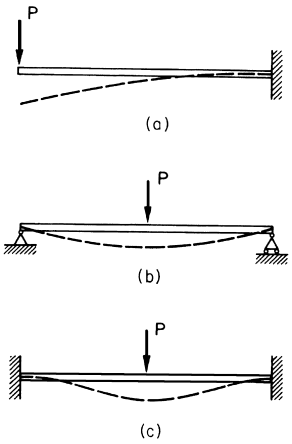


FIGURE 15 (a) Cantilevered beam, (b) simply supported beam, and (c) fixed-end beam.

distribution is equivalent to two equal and opposite horizontal forces acting at the centroids of the triangles, that is, at one-sixth of the depth measured from the top and bottom of a rectangular beam.

The two forces produce a resisting moment that equilibrates the externally applied bending moment. Because the distance between these forces is $\frac{2}{3}d$, a deeper beam requires smaller tension and compression forces and consequently lower stresses than a shallower one (Fig. 16).

Bending stresses are also influenced by the distribution of the cross-sectional area. The greater the area that is located far from the neutral surface, the smaller the bending stress. A measure of the area distribution is the moment of inertia or second moment of area, defined mathematically as the following equation:

$$I = \int_A y^2\, dA.$$

For a rectangular cross section this integral becomes equal to the following:

$$I = bd^3/12.$$

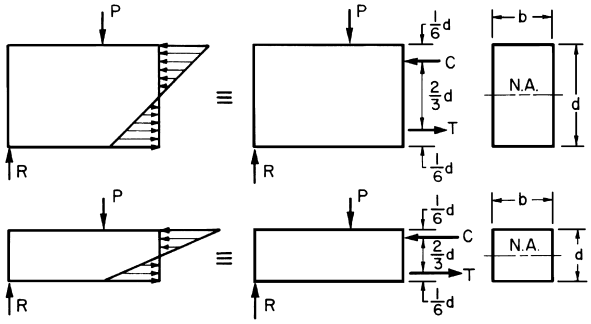


FIGURE 16 Stresses in deep and shallow beams.

Wide-flange beams or I-beams have most of their areas concentrated in their flanges, away from the neutral surface, and hence have large moments of inertia (Fig. 8c). They are the most frequently used shapes in beam design.

The maximum bending stress, S_{max}, in a beam is proportional to the maximum bending moment, M_{max}, and to the distance, c, between the neutral surface and the extreme fiber and is inversely proportional to the moment of inertia, I, as follows:

$$S_{max} = M_{max}c/I.$$

Horizontal beams carry vertical loads to their supports. A popular vertical support member is the column, which is subjected mostly to compressive forces although some secondary bending stresses may also be present.

A long slender column is prone to buckling. The critical load, P_{cr}, at which a simply supported column buckles is proportional to the modulus of elasticity, E, of its material and to the moment of inertia, I, of its cross section and is inversely proportional to the square of its length, l according to the Euler formula as follows:

$$P_{cr} = \pi^2 EI/l^2,$$

where $\pi = 3.1416$.

Because the buckling load is proportional to the moment of inertia, structural shapes favored for beams are also useful in column design.

The critical load is also dependent on support conditions. The force required to buckle a column with both ends fixed is, for instance, four times greater than in the simply supported case, while a cantilevered column has a carrying capacity one-quarter that of the simply supported column (Fig. 17).

Beams and columns are commonly interconnected to produce frame-type structures and usually form the internal skeleton of buildings. Multicell frames in three dimensions are used in such cases. It should be noted that the interaction between beams and columns may produce deformations and hence bending stresses in members not directly loaded (Figs. 18a and 18b).

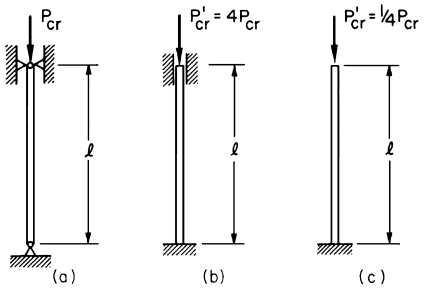


FIGURE 17 (a) Simply supported, (b) fixed-end, and (c) cantilevered columns.

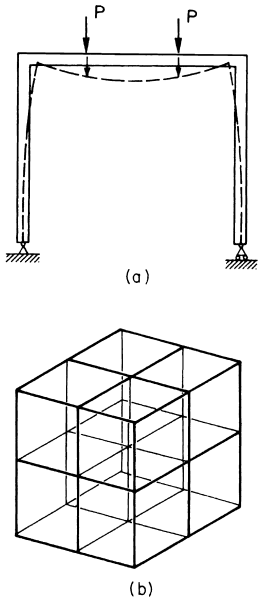


FIGURE 18 (a) Simple and (b) multicell frames.

Combined stresses may also result from such beam–column interactions. The downward force, P, in Fig. 19a produces compressive stresses in the column with superimposed bending. Because bending alone results in tensile stresses on one side of the column and compression on the other (Fig. 19b), the combined effects are an increased

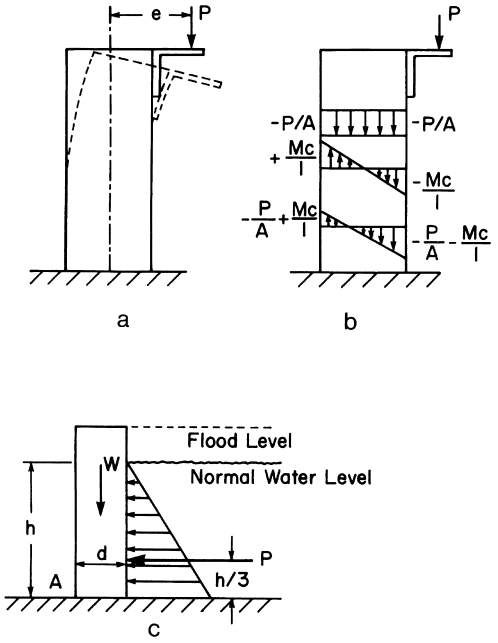


FIGURE 19 (a) Excentrically loaded column; (b) stress distribution in the column; (c) forces producing combined stresses in a dam.

compression on the right side and a decreased tension on the left

$$S = -\frac{P}{A} \pm \frac{Mc}{I},$$

where $M = Pe$ is the bending moment produced by the excentric load, P.

Similar combined effects are present in dams and retaining walls, where the horizontal pressure of the water or the retained earth and the vertically acting weight of the structure induce both bending and direct compression in the wall (Fig. 19c). Earth and fluid pressures vary triangularly with depth; the resultant pressure load, P, is acting at one-third the total height of the structure. Hence, stresses in the wall are again obtained as the following:

$$S = -\frac{W}{A} \pm \frac{Mc}{I},$$

where $M = Ph/3$.

Frequently overlooked in large structures are thermal stresses produced by temperature changes. It is a well-known fact that a temperature increase results in an expansion, while a reduction is responsible for contraction of structural materials. The thermal coefficient of expansion, α, is a measure of the change in a unit length of material due to a 1° temperature change. As an example, for steel we have the following:

$$\alpha_s = 11.7 \times 10^{-6} \text{m/m°C}.$$

Consequently, the total increase in length, $\Delta\ell$, of a bar ℓ in. long subjected to a temperature change of ΔT degrees is given by the following:

$$\Delta\ell = \alpha_s \ell \Delta T,$$

where the corresponding thermal strain is given by the following:

$$\varepsilon_T = \Delta\ell/\ell = \alpha_s \Delta T.$$

Should such an expansion (contraction) be prevented by inserting a steel bar between two rigid walls, the resulting elastic thermal stress would be given by the following:

$$S_T = \varepsilon_T E_s \quad \text{or} \quad S_T = E_s \alpha_s \Delta T.$$

For instance a temperature rise of 50°C in a built-in steel bar with $E_s = 210$ Gpa (210×10^9 N/m^2) produces a compressive stress of 123 Mpa (123×10^6 N/m^2), the additional bending stress due to a transverse load (Fig. 15c) may result in a combined stress that exceeds the allowable strength of the material or might buckle the beam. Such thermal stresses are consequently dangerous and should be avoided by permitting the structure to expand (contract). This is accomplished by freeing one end or supporting it on a roller (Fig. 15a,b). Similar freedom of movement is provided by temperature expansion strips placed periodically in concrete highways.

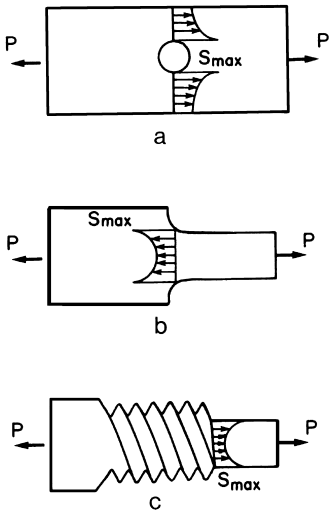


FIGURE 20 Stress concentrations around (a) holes, (b) fillets, and (c) threads.

While direct forces applied to components with uniform cross sections produce uniform tensile or compressive stresses as indicated in Section II, bending stresses vary throughout the depth of a beam (Fig. 8b) and result in stress gradients. Structures under direct loads are also subjected to stress gradients in the vicinity of abrupt changes in geometry. Holes, sharp corners, thread roots, and fillets are responsible for such concentrations of stress as shown in Fig. 20. At the edge of these "stress raisers" the maximum stress, $k_t P/A$, can be several times greater than the average value of P/A. Charts are available for the calculation of the "stress concentration factor," k_t.

VII. TWO- AND THREE-DIMENSIONAL STRUCTURES

Cables, arches, beams, and columns are one-dimensional structures; load transfer to the supports takes place along a single straight or curved line.

The umbrella of a parachute, the spinnaker of a sailboat, the skin of a hot-air balloon, and the roof of a tent are membranes, which are the two-dimensional equivalents of cables. Loads are carried along plane or curved surfaces. Like cables, membranes are also unstable tensile structures but are useful in lightweight systems when properly stabilized.

Internal-pressure-stiffened pneumatic structures are frequently employed as roofs for tennis courts and swimming pools (Fig. 21).

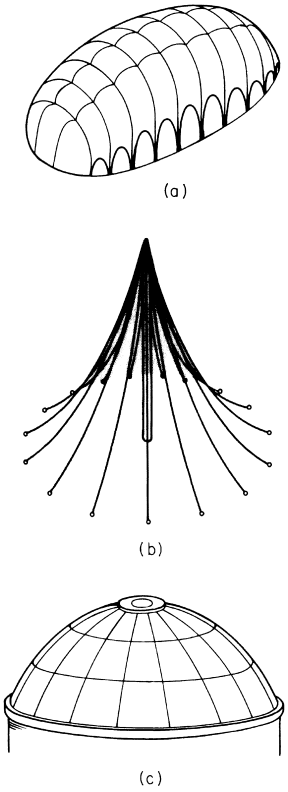


FIGURE 21 (a) Pneumatic structure, (b) membrane structure, and (c) cupola.

Tents and umbrellas are stabilized by applying tension to the membrane with the aid of tent poles, internal frames, and spokes (Fig. 21).

The two-dimensional equivalent of the compressive arch is the dome. Because only compressive stresses exist in properly constructed domes, such structures may consist of individual stones. The ancient Romans made use of these techniques in the construction of the Pantheon, built by Hadrian nearly 2000 years ago. It is still standing.

Just as for arches, buttressing is also required for domes, to prevent the outward movement of the circumference. Hadrian used heavy stone rings to provide this restraint (Fig. 21). A similar modern structure based on this principle is the Jefferson Memorial in Washington, DC.

Plates are the planar equivalents of the one-dimensional beam (Fig. 22). Plates carry loads primarily by bending, although midplane tension and twisting are also present. To have high bending resistance, plates require large moments of inertia also. This is frequently accomplished with sandwich-type construction: Relatively weak materials such as styrofoam, paper honeycomb, and wood are sandwiched between two layers of strong flange materials.

The weak core separates two strong layers and gives greater depth and moment of inertia to the plate. Modern

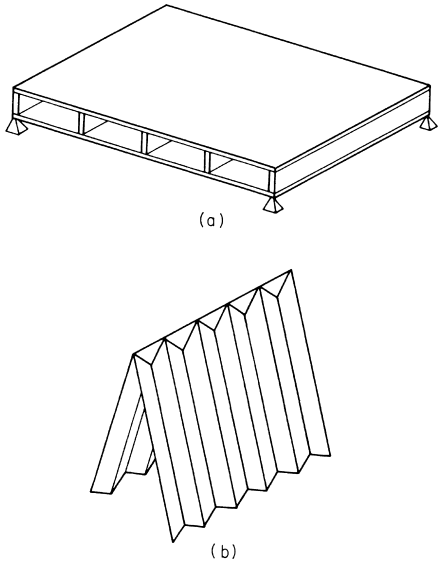


FIGURE 22 (a) Ribbed plate and (b) folded-plate structures.

reinforced materials, among them fiberglass and graphite epoxy as well as metals, are used as skin in such sandwich structures. Ribs may also be employed to separate the outer layers producing an I-beam effect (Fig. 22a).

Plates do not have to be horizontal; vertical and slanted or folded plates frequently form the sides and roofs of interesting structures (Fig. 22b). A folded plate has a greater moment of inertia than a flat one and acts as a deep beam because some of the material is moved away from the neutral surface (Figs. 23a and 23b).

Plane trusses are also two-dimensional structures because they carry loads to their supports in the plane of the bars.

VIII. THIN SHELLS AND SPACE FRAMES

When a sheet of paper is held horizontally, it bends down under its own weight, but given a slight curvature, it forms a cantilever beam (Figs. 24a and 24b). The curved surface behaves similarly to the folded plate: Some of the material is moved away from the neutral surface, thereby increasing the moment of inertia.

Cables carry loads by tensile stresses. An inverted cable forms an arch that transmits loads through compressive stresses. Analogously, the two-dimensional equivalent of the cable, the membrane, is in tension, while an inverted membrane shape is compressed. Because compressed elements are prone to buckling, such structures, called thin shells, are made of stiffer and somewhat thicker materials than membranes.

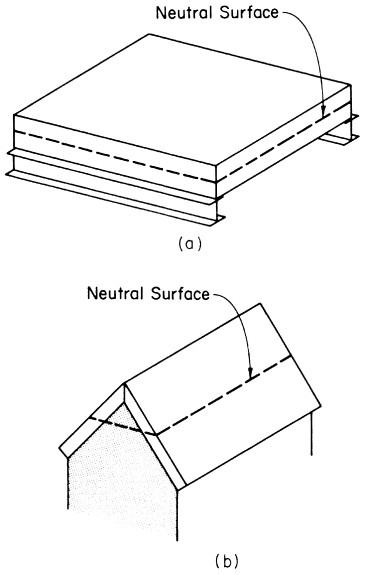


FIGURE 23 Increased moment of inertia of (a) flat and (b) folded plates.

Thin shells derive their strength from their curvatures. Curvatures, which are the reciprocals of the radii of curvature, are measured in two perpendicular cutting planes through a point on the surface. If both are downward, as in a dome (Figs. 25a and 25b), or both upward, as in a dish antenna, the surface is known as synclastic. On the other hand, an anticlastic or saddle surface has one curvature downward while the other one is upward (Fig. 25c).

One may imagine a hemispherical dome to consist of compressed meridional arches held together by tensile hoops along the horizontal circles near the bottom (Fig. 25b). Similarly, a saddle-type surface such as a hyperbolic paraboloid (Fig. 25c) may be considered to consist of compressed, downward-curving arches carried on upward-curving cables. Hence thin shells must be constructed of materials capable of carrying both tension and compression. The most frequently used material is reinforced concrete, but modern fiber-reinforced plastics are finding new popularity in such structures.

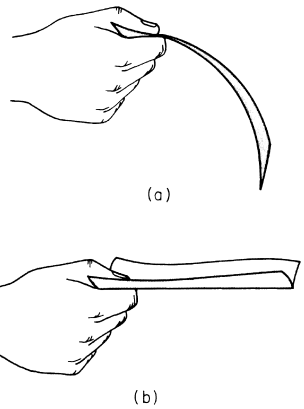


FIGURE 24 Increased moment of inertia of (a) flat and (b) folded sheets.

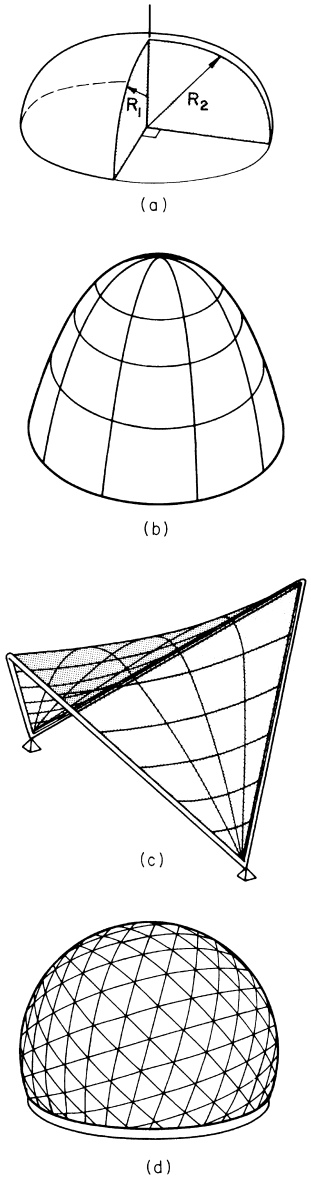


FIGURE 25 (a) Radii of curvature in a dome, (b) elliptic paraboloid dome, (c) hyperbolic paraboloid, and (d) geodesic dome.

Structures having the same general shape as thin shells are also constructed with the aid of a lattice work of triangulated bars. An example of this type of space frame is the geodesic dome. Such structures are in essence three-dimensional trusses (Fig. 25d).

A building in which beams, frames, and columns are arranged in three directions (Fig. 18b), and have significant

dimensions in all three, is obviously a three-dimensional structure, as is the truss of Fig. 14g. The analysis of such systems required six equations of spatial equilibrium at every joint: Three summations of forces along the three coordinate axes must add to zero as well as three summations of moments about the same axes. Such an analysis leads to information about the forces and bending moments carried by individual components, which are then treated as one-dimensional structures for purposes of stress analysis.

To calculate the forces and moments in members of a complicated structure or structural system, computer-based methods are used. The matrix method was discussed above.

Continuous structures, such as an airplane fuselage with numerous window cutouts and aerodynamic shape, require the use of differential equations to determine stresses and strains at all points. While such equations can be formulated, frequently they cannot be solved in closed form. Instead, approximate methods of numerical analysis must be used, in which the actual structure is subdivided into smaller, simpler components interconnected (mathematically) at their boundaries. Such finite elements are piecewise approximations to the continuous structure. These elements are connected at nodes where simultaneous equations are solved.

As an example, the area of a circle may be calculated as the sum of the areas of the inscribed triangles (finite elements) in Fig. 26. This approximation results in an error because the shaded areas are neglected. The accuracy of the calculation may be improved by increasing the number and reducing the size (mesh size) of the triangles. The smaller mesh size, however, requires a larger number of computations.

By using appropriate equations derived from variational methods, one can calculate stresses and strains for the finite elements that approximate the structure. The finer the mesh size the better the approximation, but the cost of computer time increases sharply. A mesh arrangement for a plate containing several holes is shown in Fig. 27.

Several multipurpose finite element computer codes are available for the solution of complicated structural problems.

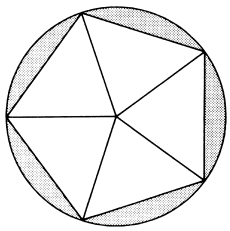


FIGURE 26 Area of circle approximated by finite elements.

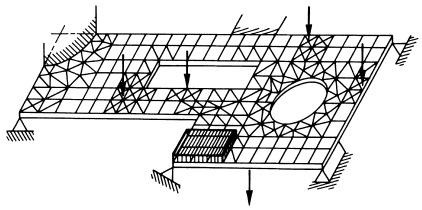


FIGURE 27 Finite element mesh for a plate with cutouts.

IX. STRUCTURAL SAFETY AND RELIABILITY

To insure that structures are safe for public use it is necessary to know the worst loading conditions a structure may be subjected to. While it is not difficult to estimate occupancy loads such as the number and weight of people in a theatre or in an aircraft (Table II), natural forces are unpredictable. Earthquakes occur at random, and the intensity of the quake can only be guessed (Fig. 6). Wind, rain, and snow loads are also variable loads though worst conditions may be gleaned from historical records (Figs. 4 and 5).

With the aid of structural analysis, member stresses can be calculated, which are then compared with the strength of the component. Structural safety is then insured with the use of safety factors. The safety factor is the ratio of strength to applied stress and can range from 1.25 for some aerospace structures to 6–7 in sensitive nuclear installations.

Basing structural reliability on worst-case scenarios is, however, neither perfectly safe nor always feasible. For instance, aircraft cannot be built to the same standards of safety as buildings because such requirements would make airplanes too heavy to fly.

New techniques based on statistics and probabilistic methods are being developed to predict the chance occurrence of heavy loads acting on weak structural components and thereby to predict the probability of failure. By setting this probability at a low level, one can design structures for desired levels of safety.

The strength, R, of structural components as well as the applied stresses, S, may be statistically distributed as illustrated on Fig. 28. In a well-designed structure the two probability density functions, $f_R(r)$ and $f_S(s)$, are separated and have only small interference (overlapping regions). The size of the overlap is related to the probability of failure, i.e., the probability that S is greater than R as follows:

$$P_{\mathrm{f}} = P[S > R].$$

If this probability is too high, stresses may be reduced by increasing the cross-sectional dimensions (area, moment of inertia, etc.) of the component.

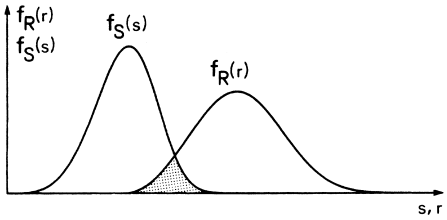


FIGURE 28 Stress–strength interference.

The design usually relies on the initial distributions of material strength and applied loads; these may change during the life of the structure. The strength of the materials may be reduced by environmental effects (corrosion), fatigue, creep, and so on. On the other hand stress redistribution may increase the initial stresses. As a consequence the two density functions may move closer to each other and the overlap (cross-hatched region, Fig. 28) may increase, indicating a greater probability of failure. If this rate of increase is known and by predetermining the acceptable maximum probability of failure, the safe service life a component may be calculated.

The various structures described in the foregoing consist of systems of interconnected beams, bars, cables, plates, and shells. Each of the components is designed to withstand the applied stresses, but arrangement of these components and their interactions greatly influence the reliability of the complete structure.

Two basic systems and their combinations are examined. In a "series system," all components must survive for the system to survive. The prototype of such an arrangement is a chain consisting of several links as presented in Fig. 29a. The chain fails if any one of its links fails and survives only if all its links survive.

Denoting the reliability of a link by L_i, the reliability of the series system is the product of the reliability of its components as follows:

$$L_s - L_1 \times L_2 \times \cdots \times L_i \times \cdots \times L_n$$

and its probability of failure is given by the following:

$$P_\mathrm{f} = 1 - L_s.$$

Statically determinate structures such as the trusses discussed in Section V are series structures. The removal of any one of their bars renders them unstable (Fig. 2c). If, for example, each bar of the 33-member truss shown in Fig. 14f is designed to a reliability $L_i = 0.9999$ (that is, only 1 bar in 10,000 is prone to failure) ($P_{\mathrm{f}i} = 0.0001$), the reliability of the truss is $L_s = L_i^{33} = 0.9999^{33} = 0.997$. The probability of failure for the series structure is consequently much greater ($P_\mathrm{f} = 0.003$, or 3 in a 1000 such trusses may fail) than that of the individual bars.

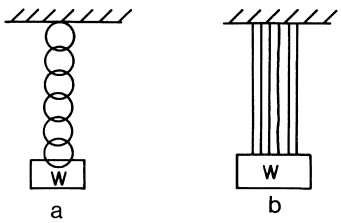


FIGURE 29 (a) Series and (b) parallel system.

"Parallel systems," on the other hand, are capable of carrying their loads as long as a single component survives and fail only when all members fail. A load supported by several cables is such a system (Fig. 29b). If the probability of failure of each member is $P_{\mathrm{f}i}$, the probability of failure for the system is the product of the component failure probabilities as follows:

$$P_\mathrm{f} = P_{\mathrm{f}1} \times P_{\mathrm{f}2} \times \cdots \times P_{\mathrm{f}i} \times \cdots \times P_{\mathrm{f}n}$$

and its reliability is as follows:

$$L_i = 1 - P_\mathrm{f}.$$

The four engines of an airplane are in such a parallel arrangement. If each engine has a probability of failure, $P_{\mathrm{f}i} = 0.001$, i.e., 1 in a 1000 engines may fail, the probability of failure for the airplane becomes the following:

$$P_\mathrm{f} = P_{\mathrm{f}i}^4 = 0.001^4 = 1 \times 10^{-12}.$$

The probability of failure for parallel systems is much smaller than those of its components.

Statically indeterminate or redundant structures behave essentially as parallel systems, though in these only a limited number of members may fail before the structure itself fails. In the redundant truss of Fig. 14e, any one bar may fail without endangering the structure, but the failure of a second bar will render it unstable.

Most real structures are combinations of series and parallel systems. Their reliability analysis becomes more complex.

X. STRUCTURAL FAILURES

Although engineering structures are designed and built with utmost care, failures do occur in some instances with attendent loss of human life.

Failures are due either to excessive deformation or to fracture. Large irrecoverable "plastic" deformation in structures made of elastic-plastic materials such as steel (Section III) is usually the result of unexpected overloads. When a brittle material (concrete or glass) is overloaded, it usually breaks without excessive deformation. The concrete columns of the Nimitz Highway broke and allowed the roadway to collapse during the 1989 San Francisco Earthquake. Sixty-six people lost their lives. The 1999

earthquakes in Turkey and Taiwan also destroyed mostly concrete and brick structures with thousands of lives lost. In contrast, several steel structures showed excessive deformations but did not break.

Creep is another form of permanent deformation. Creeping, or "viscoelastic," materials deform even under small loads, and the deformation continues to grow as long as the load is applied. Creep is accelerated at elevated temperature. Most polymers and even metals at high temperature creep. Turbine blades in jet engines run at high temperatures and are subjected to centrifugal forces. If they creep excessively they may touch the turbine's casing and break off with consequent serious damage to the engine.

Several forms of instability are classified as excessive deformations. Sliding, overturning, buckling, loss of redundancy, motion of cables and membranes are among such deformations.

Structures built on the edge of the Pacific Palisades near Los Angeles have been known to slide down into the ocean when their rain-soaked foundations gave way. Dams may fail by overturning where, during a flood, the overturning moment of the water pressure ($M_0 = P \times h/3$ in Fig. 19c) becomes greater than the resisting moment ($M_R = W \times d/2$) produced by the weight of the dam. The leaning tower of Pisa, built on an uneven soil foundation, is unstable and is in danger of overturning. Efforts are underway to stabilize it, though a straightened tower would probably not be as interesting as the slanting one.

Buckling of roads during hot summer days is a frequent occurrence. The flexibility of the cables of the Tacoma Narrows bridge caused its collapse (Fig. 3) in 1940. The wind-induced oscillations continued for 3 hr before final failure, luckily without loss of life.

Fractures are usually more serious failure modes than large deformations. As stated earlier, structures made of brittle material break without warning if overloaded. Fractures are initiated at regions of high stress, usually at points of stress concentration.

Ductile structures may also fracture but more often as a result of metal fatigue caused by repeated loads rather than due to a single overload. It is a common experience that a piece of wire will not break if bent once, but repeated back and forth bending will break it. The phenomenon is called metal fatigue. Structures may withstand millions of load cycles, but their fatigue lives are finite. Eventually, a small crack will start at a point of stress concentration and will propagate through the structural component until finally the component fractures.

Corrosion is another damaging effect in metal structures. It is accelerated in the presence of stresses and frequently accompanies fatigue crack propagation.

Such stress-corrosion combined with fatigue was responsible for the collapse of the Silver Bridge over the Ohio River between Point Pleasant, West Virginia, and Gallipolis, Ohio, in 1968. This eye-bar suspension bridge, built in 1929, was constructed of individual chain links (eye-bars) (Fig. 30). Stress-corrosion fatigue attacked and eventually broke one of the eyes (Fig. 30), the chain separated, and, being a series-type structure, the bridge collapsed into the Ohio River, causing the deaths of 46 people.

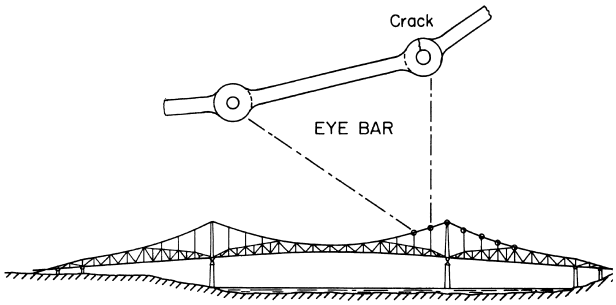


FIGURE 30 Eye-bar suspension bridge.

Metal fatigue was also responsible in 1988 for the in-flight loss of a 40-ft fuselage section in a Boeing 737 jet operated by Aloha Airlines. The cabin of commercial jets is pressurized for passenger comfort and depressurized on each takeoff and landing. This repeated stressing initiated cracks along a series of rivet holes and adhesive joints in the skin of the aging aircraft (Fig. 31). This redundant structure did not, however, fail completely; the aircraft was able to land, and only one person was killed.

The most probable cause of the crash of a United Airlines DC-10, in Sioux City, Iowa, in 1989, is the fatigue fracture of an engine part. The flying shrapnel also severed some hydraulic lines, making the aircraft uncontrollable. The crash landing killed 111 passengers, but 185 survived.

Structural design and analysis encompasses many diverse fields of the arts, sciences, and engineering and requires the cooperation of architects, engineers, material scientists, aero- and hydrodynamcists, and social scientists

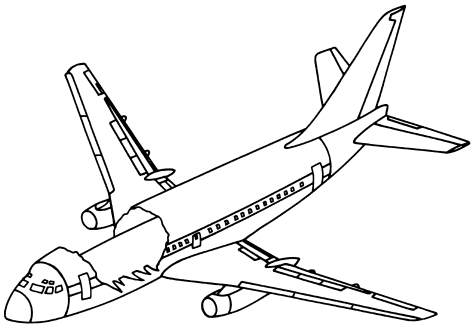


FIGURE 31 Stress-corrosion fatigue failure of aircraft fuselage.

to make structures esthetically pleasing, functional, environmentally acceptable, safe, and economical.

SEE ALSO THE FOLLOWING ARTICLES

EARTHQUAKE PREDICTION • ELASTICITY • EMBRITTLEMENT, ENGINEERING ALLOYS • FRACTURE AND FATIGUE • MECHANICS, CLASSICAL • RELIABILITY THEORY • SANDWICH COMPOSITES

BIBLIOGRAPHY

Barsom, J. M., and Rolfe, S. T. (1987). "Fracture and Fatigue Control in Structures," Prentice-Hall, Englewood Cliffs, NJ.

Dowling, N. E. (1999). "Mechanical Behavior of Materials," 2nd ed., Prentice-Hall, Upper Saddle River, NJ.

Elishaloff, I. (1983). "Probabilistic Methods in the Theory of Structures," Wiley, New York.

"GT STRUDL," Georgia Tech, Case Center, Atlanta, GA.

Heller, R. A., and Pap, A. A. (1987). "Video Tape Series on Mechanics of Materials and Structures," Virginia Tech Intellectual Properties, Blacksburg, VA.

Heller, R. A., and Thangjitham, S. (1987). Probabilistic methods in thermal stress analysis. *In* "Thermal Stress II," (Hetnarski, Ed.), Vol. 2, pp. 189–265, Elsevier, Amsterdam.

Hibbler, R. C. (1995). "Structural Analysis," 3rd ed., Prectice-Hall, Upper Saddle River, NJ.

Inman, D. J. (1994). "Engineering Vibration," Prentice-Hall, Upper Saddle River, NJ.

Ott, J. (1989). "Investigators find reconstructed tail of DC-10 riddled with damage," *Aviat. Week Space Technol.* **Aug. 7,** 22–23.

Ott, J., and O'Lone, R. G. (1988). "737 fuselage separation spurs review of safeguards," *Aviat. Week Space Technol.* **May 9,** 92–93.

Plowden, D. (1984). "Bridges: The Spans of North America," American Society of Civil Engineers, New York.

Reddy, J. N. (1984). "An Introduction to the Finite Element Method," McGraw-Hill, New York.

Riley, W. F., Sturges, L. D., and Morris, D. H. (1995). "Statics and Mechanics of Materials," Wiley, New York.

Salvadori, M. G., and Heller, R. A. (1986). "Structure in Architecture: The Building of Buildings," 3rd ed., Prentice-Hall, Upper Saddle River, NJ.

"SAP 2000" Computers and Structures, Inc. Yorba Linda, CA.

Shermer, C. (1968). "Eye-bar bridges and the Silver Bridge disaster," *Engineer* **Jan.–Feb.,** 20–31.



Mining Engineering

Jaak J. K. Daemen
University of Nevada, Reno

I. Mining
II. Mining Geology
III. Mining Method Selection
IV. Mine Feasibility Evaluation
V. Mine Planning
VI. Mine Unit Operations
VII. Auxiliary Mining Engineering Disciplines
VIII. Mine Utilities
IX. Health and Safety Engineering
X. Environmental Engineering and Reclamation
XI. The Future

GLOSSARY

Adit Horizontal or nearly horizontal excavation (opening, drift, gallery) of limited cross-sectional area, providing access to an underground mine located in the side of a hill or mountain.

Auger mining Coal mining method in which the coal is drilled out of the seam (augered) by means of a large coal auger, a screw-type drill. The driving and loading mechanism is installed on the surface and travels along the seam being mined. The auger almost always operates horizontally. Auger mining is used primarily when the overburden is too thick to be removed economically, and usually follows strip mining or contour mining.

Block caving Underground mining method, primarily for mining of metal ores. A thick block of ore is undermined by a vertically thin but laterally extensive area into which the undercut block is allowed to cave. As the broken ore is withdrawn underneath the undercut, caving propagates upward until the entire block to be mined has been removed.

Contour mining Coal mining method in which the mine excavations proceed along the outcrop of a coal seam on the side of a hill. In conventional contour mining (rim stripping) the overburden is spoiled (cast) down the hill below the coal seam, thus exposing the coal. This enviromentally destructive mining method has largely been replaced by a variety of environmentally acceptable contour mining methods.

Discovery Within the mining context, a controversial and ambiguous term indicating that a valuable mineral deposit has been located. Legal U.S. practice requires that a prudent person would be justified in making further

expenditures of labor and means on the site, or that the mineral located can be marketed at a profit.

Gangue Waste or barren rock. Minerals without economic value, associated and intermingled with economically valuable metal-bearing minerals within most metal ore deposits.

Level System of connected horizontal mine workings at a given elevation. Underground mines that operate over a significant vertical extent usually are subdivided into multiple levels. Each level is the operations base for mining the ore up or down to the adjacent level. (Also a surveying instrument.)

Lode Strictly speaking, a fissure filled with metalliferous ore within a barren rock formation. Widely used for a thin tabular orebody, vein, or ledge containing metal values, with a clearly demarcated boundary from the surrounding nonmetallic barren rock. Acceptable for any zone or belt of mineralized hard rock (as distinguished from sand, gravel, i.e., alluvial deposits) with unambiguously identifiable boundaries or limits.

Mineral deposit Volume of rock within which a mineral of present or of potential future economic value is concentrated to a sufficient degree and extent to make it clearly distinguishable from the surrounding rock, which typically has no foreseeable economic value.

Mountain-top removal mining Coal mining method practiced in hilly or mountainous terrain, where the coal seam is not too deep below the mountain top. As implied by the name, the entire rock mass overlying the coal seam(s) is removed.

Ore deposit, orebody Mineral deposit with geological (grade, composition, geomechanical behavior), geometrical (shape, volume, orientation, depth), and locational (e.g., distance from shipping or railroad lines, environmental impact of mining) characteristics such that it can be mined economically within a given set of technological and marketing conditions. A narrow definition restricts the use of the term ore to mineral deposits from which metals are to be recovered. A broader definition allows the term ore to be used for all economically recoverable solid mineral resources.

Quarry Excavation or working on the surface, that is, in open air (unless specifically referred to as an "underground quarry") for the production of minerals used primarily in the construction or building industry, such as limestone for cement manufacturing; sand, gravel, and aggregate for concrete production; or building stone such as granite, marble, limestone, or slate.

Shaft Vertical (vertical shaft) or inclined (inclined shaft; slope; incline; decline; ramp) opening of small cross-sectional area excavated from the surface down to a mine or prospective mine (shaft sinking) or from a mine level up to the surface (shaft raising).

Surface mining, open-pit mining, open-cast mining, open-cut mining, area mining, strip mining, quarrying Various methods for removing minerals in open air, if necessary after removal of an economically useless overburden consisting of barren or waste rock.

Underground, subsurface, deep mining Recovery of solid minerals underneath overburden, the rock overlying the extracted mineral, the latter being of insufficient economic value to warrant recovery and of too great a volume or depth to warrant the removal that would be required for surface mining.

MINING ENGINEERING is the application of engineering sciences and principles to the extraction of mineral resources from their natural environment. Engineering in support of the mining industry integrates multiple disciplines, ranging from geology and its many subfields to mechanical, systems, and electrical engineering, and including operations research, remote control, and automation. It is customary to limit the use of the term *mining* to the extraction of solid minerals. The extraction of fluid resources such as oil, gas, and groundwater usually is not considered to be part of mining. Recent developments, such as oil-shale and tar-sand mining and the utilization of methane gas recovered from coal mine seams, blur the distinction between mining and petroleum engineering. The broader term *minerals engineering* encompasses exploration, extraction, and processing of minerals in general. Traditionally, the definition of mining is restricted to recovery from the Earth's crust. On-going investigations of extraterrestrial mining and of deep-sea mineral recovery might make such a limitation too restrictive.

Mining nomenclature varies widely. Many terms are far from universally accepted or even understood. Major differences exist among English-speaking countries and between coal mining and metal mining. Many mining districts have their own terminology, developed historically, and frequently in isolation from other districts. The "definitions" given in the Glossary are an amalgam of widely used U.S. coal and metal mining conventions.

I. MINING

Because mining engineering is so predominately and intimately associated with one industry, it is instructive to discuss briefly the industry itself. Its various aspects have a strong bearing on the type of engineering most likely to be associated with different areas of this very diverse industry.

The first and most visible distinction is the one between surface mining and underground mining. In surface

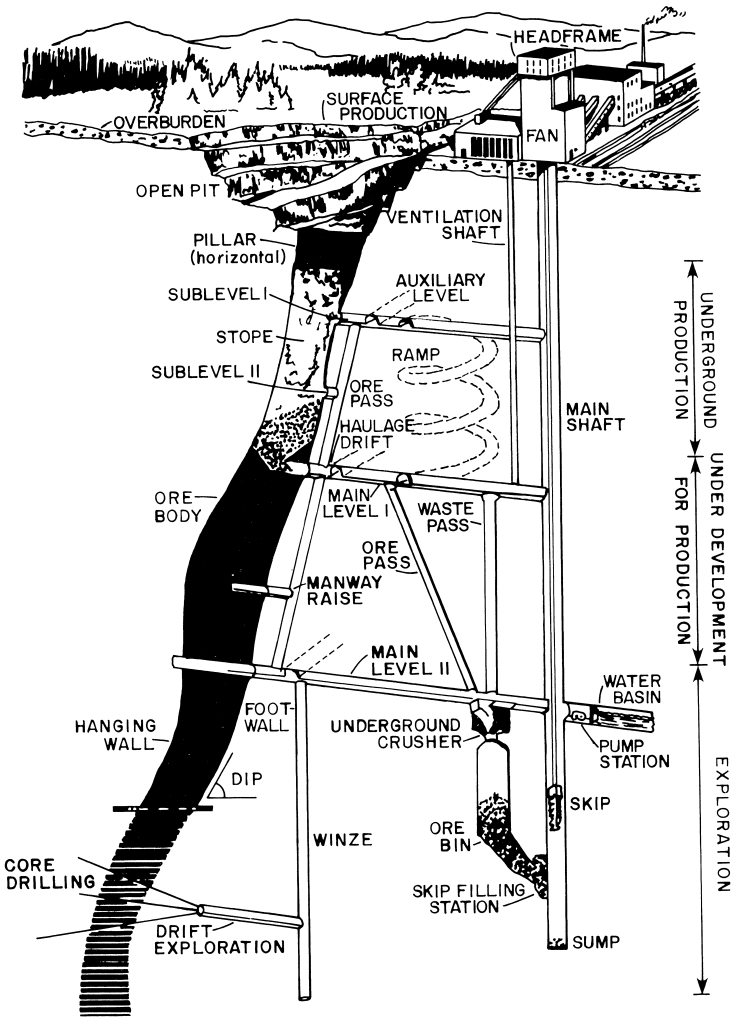


FIGURE 1 Open-pit mining and one underground mining method. Various mining terms and operations are identified. [From H. Hamrin (1982). Choosing an underground mining method. *In* "Underground Mining Methods Handbook" (W. A. Hustrulid, ed.), pp. 88–112. Society of Mining Engineers of the American Institute of Mining, Metallurgical, and Petroleum Engineers, Inc., New York.]

mining, the mineral of interest is removed in open air, frequently after preliminary removal of sterile overburden. In underground mining, access to a subsurface mineral is developed and the mineral of interest is removed underneath overlying ground (Fig. 1). A second important distinction is the one between hard-rock and soft-rock mining. The type of rock strongly affects the mining methods and equipment used. A classification based on the type of resources recovered is between metal and nonmetal mining. The latter category usually is divided into coal, or more broadly fuels, and industrial minerals. Nonmetal mining is, by far, the dominant branch of the industry. It moves much larger rock volumes, its total production has a considerably larger value, it employs many more people, at a larger number of mines, and it has the bulk of the invested capital. A legally particularly important distinction in the United States is that between leasable minerals, primarily energy minerals (coal, oil shale, oil, gas) as well as phosphate and sodium, for which land is leased from the federal government in return for the payment of rentals and royalties, and most other minerals for which federal land can be "claimed."

II. MINING GEOLOGY

Because mining operates in nature, recovering mineral resources from their natural environment, the geological sciences play a fundamental role in many aspects of mining engineering. Geology dominates the investigations preceding actual mine operations, i.e., exploration or the search for an economically recoverable mineral deposit,

as well as site characterization to provide the required detailed information to allow feasibility studies, mine method selection, and mine planning. During operations, mine geologists provide the information for further mine development.

Before mining can start, an economically mineable mineral deposit must be available. Exploration geology, an applied branch of the geosciences, has as its fundamental objective the finding of mineral reserves. Exploration proceeds in different phases, ranging from general and broad reconnaissance exploration to very narrowly focused target characterization. The goal of the broad reconnaissance investigation is to identify large areas or geological regions that are favorable with regard to the likelihood of finding a particular type of mineral deposit. The goals for the general reconnaissance programs usually are determined by market considerations. For example, during a period when precious metals (gold, silver, platinum) demand and prices are high, relative to their typical mining cost, it is likely that exploration groups (e.g., subsidiaries or divisions of mining firms as well as independents) will make an effort to locate potentially highly profitable deposits of precious metals. Conversely, when demand for a commodity decreases, exploration expenditures for that particular mineral are likely to decrease also. A typical example, the change in exploration expenditures for uranium, in response to fluctuations in the projected growth of nuclear power, is illustrated in Table I.

A more recent example of the sensitivity of exploration spending to metal prices is the decline in gold exploration expenditures as a result of the steep decline in gold prices during the late 1990s. In 1998 estimated gold exploration commitments in the United States were US $103 million, down from $160 million in 1997. Similarly, the Latin American gold exploration spending in 1998 decreased to US $148 million, from $241 million in 1997.

TABLE I Drilling Activity and Expenditures for Uranium Exploration in the United States, 1973–1983[a]

Year	Drilling activity [million ft (m) drilled]	Total exploration expenditures (million constant 1983 dollars)
1973	17 (5.2)	102
1974	22 (6.7)	149
1975	25 (7.6)	210
1976	35 (10.7)	279
1977	46 (14.0)	400
1978	48 (14.6)	452
1979	41 (12.5)	419
1980	28 (8.5)	325
1981	14 (4.3)	161
1982	5 (1.5)	77
1983	3 (0.9)	40

[a] Source: Energy Information Administration (1983). "1982 Survey of United States Uranium Exploration Activity," DOE/EIA-09402(82), Washington, D.C., August.

Reconnaissance exploration usually starts with intensive literature studies, certainly in areas where geological information already is available, for example, because of previous investigations by such agencies as the U.S. Geological Survey or state geological surveys, or universities. This preparatory step is followed by field studies that allow rapid coverage of large areas at minimum cost (e.g., by means of remote sensing or airborne geophysics), followed in turn by more detailed geological mapping, as well as field geophysical and geochemical surveys. An important aspect of all exploration is the development and continued refining of a fundamental geological model that integrates and explains to the greatest extent possible all available geological information. The model, in turn, provides guidance for further exploration approaches.

If the reconnaissance exploration proves successful — that is, identifies one or more well-defined mineral deposits that justify deeper investigation — target characterization can be initiated. This exploration phase focuses on relatively small areas, which are studied in detail. Sample collection by means of diamond coring will provide physical evidence of any mineral values present. If the target continues to be promising as more detailed information becomes available, sampling is intensified to provide the data for reserve evaluations, mine feasibility assessments, mining method selection, and initial mine planning.

During exploration vast amounts of data are collected, and the data points are distributed in three-dimensional space. Moreover, some data changes with time. GIS (Geographic Information System) software is an ideal management tool for such large, complex spatial and temporal data, and now is widely used in exploration programs. GIS finds many additional mining applications, e.g., for environmental management, where hydrological, geochemical, biological, archaeological, and socioeconomic data needs to be collected, analyzed, and presented. Detailed topographic data subject to frequent changes, e.g., the geometries of open pits, waste dumps, tailings dams and ponds, and heap leach pads, is an example of structures where GIS is eminently appropriate, effective, and efficient.

Once mining has started, the functions of the mine geologists multiply. Exploration continues, in part to extend the already discovered reserves, and in part to develop a detailed characterization of the reserves. The mine geologist samples the ore as it becomes exposed through mining, refines the understanding of the deposit and improves

the geological models thereof, and assists in identifying and characterizing the structural features of importance for the geomechanical and ground control engineering of the mine.

III. MINING METHOD SELECTION

The selection of the optimum mining method for the recovery of a particular mineral deposit is the first phase in the life of a mine where mining engineering becomes predominant. The selection of the mining method is influenced by numerous factors, including the depth, size, geometry, and value of the mineral deposit.

The first decision to be made is whether to use an underground or a surface mining method. Frequently the choice is obvious: a thick (e.g., over 10 m thick) coal seam at shallow depth (e.g., 10 m overburden, i.e., waste rock above the coal) will be mined by surface mining. A thin (e.g., 1 m thick) metal vein at great depth (e.g., 1000 m) will be mined by an underground method. For many situations the choice is not obvious and requires detailed cost comparisons and engineering analyses. Figure 1 illustrates the not uncommon situation where part of an orebody has been mined by surface mining, while the remainder of the orebody is mined by an underground method.

When the mining method selection is not obvious, cost estimates will be required for the alternatives that are being considered. A simplified approach for an open pit mine is illustrated in Fig. 2. An estimate is made of the total mining cost, of the total cost required to bring a product (e.g., copper) to market (this cost includes, for example, processing, smelting, and transportation), and of the market value of the product. On this basis a decision can be made as to how much waste rock can be removed for each ton of ore. This "stripping ratio" in turn determines what ore can be recovered profitably. Another comparative evaluation that might be made is that between underground and surface mining. In the example shown in Fig. 2, the lower right-hand part of the orebody has been determined to be mineable at a lower cost by underground mining than by open-pit mining.

The two most widely used surface mining methods are open-pit mining (Figs. 1, 2, and 3) and strip mining (Figs. 4 and 5). Quarrying is the third major surface mining method, dredging the fourth. Both in volume and in value of ore and minerals produced, surface mining dwarfs underground mining. Over 90% of all minerals production comes from surface operations. The ratio of surface to underground production has increased steadily for several decades. This trend might be reversed because of depletion of near-surface resources and because of

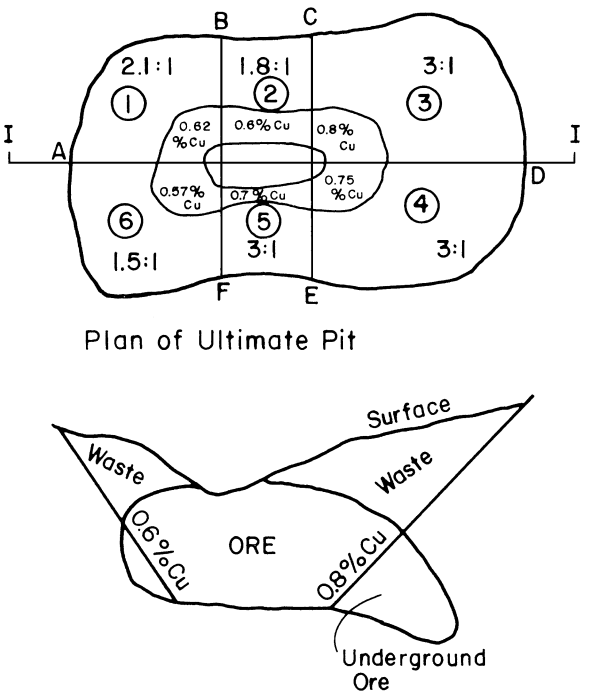


FIGURE 2 Open-pit mine design. The mine is divided into six sectors, indicated by the encircled numbers, on the basis of the ore value (in this case, percent of copper). Based on the total value of the ore and on the total production the stripping ratio, the ratio listed in each sector, is determined. This in turn fixes the final slope angles to which the pit will be mined. [From A. Soderberg and D. O. Rausch (1968). Pit planning and layout, *in* "Surface Mining" (E. P. Pfleider, ed.), pp. 141–165. The American Institute of Mining, Metallurgical, and Petroleum Engineers, Inc., New York.]

increased environmental and reclamation constraints on surface operations.

Numerous underground mining methods have been developed in response to the wide range of conditions encountered. Table II lists underground mining methods and identifies some of the major criteria applied to select a method: geometry of the deposit to be mined and strength of the mineral to be mined as well as of the adjacent rock formations. Many variations exist of virtually all of these methods. The major ones are listed in the last column.

A fairly recently developed variation of backfill stope mining is drift and fill stoping. In this underground mining method drifts are advanced by conventional drilling and blasting or by mechanical (roadheader) mining. The drifts are backfilled, usually with cemented backfill. Drifts immediately adjacent to, underneath, or above previously backfilled drifts are then mined. The method allows safe mining under relatively difficult ground conditions, can be used in orebodies of virtually any geometry, and can be planned to result in very complete mining of deposits. It tends to greatly reduce dilution, i.e., ore contamination by gangue, compared to open stope mining methods



FIGURE 3 Open-pit mine. Typical mining operations: lower front, blasthole drilling; lower left, shovel loading haul trucks; center, shovel loading haul trucks; right, trucks hauling ore out of the pit on the haulage road. The benches excavated to deepen the pit can be seen on the right and in left center. (Photograph courtesy of R. O. Roko.)

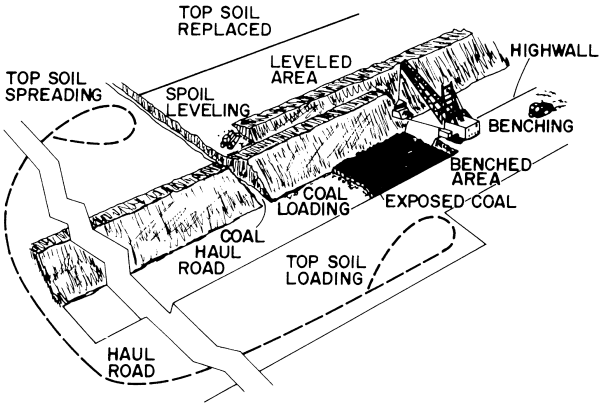


FIGURE 4 Dragline strip mining or area mining. A dragline positioned on the highwall exposes the coal by digging out the waste rock above the coal and dumping it on the spoil pile, in a pit where the coal has been removed during a previous cycle. The exposed coal is loaded (e.g., with a shovel or front-end loader) into trucks and hauled out of the mine. The dragline continues to expose coal until it reaches the end of the pit, and then mines the next pit in the opposite direction. One bulldozer is preparing a bench for the dragline; another one is leveling the spoil pile. Not illustrated is the equipment removing the top soil, in advance of benching, and hauling it to the mined-out area, where it will be used to facilitate reclamation. [From W. C. Morgan and J. C. Schermerhorn (1978). Stripping with scrapers and dozers, *Coal Mining and Processing* **15**(6), 73–77, 124, 128.]

in particular. Mining with fill allows moving back underground a significant fraction of the mined rock, if the rock is suitable for backfill emplacement, thus reducing or minimizing surface space required for waste rock disposal and the resulting permanent footprint left by mining.

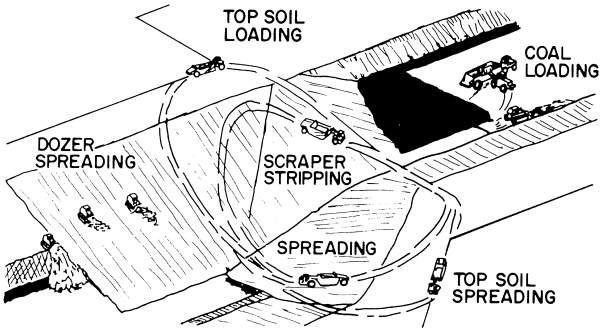


FIGURE 5 Area mining (strip mining) with scrapers and dozers. Bulldozers push the waste rock on top of the coal into a position where it can be loaded efficiently with scrapers. The scrapers expose the coal by hauling the waste rock to the mined (spoil) area. A front-end loader loads coal in a coal truck. Scrapers remove top soil from an area to be mined and spread it over a mined-out area to be reclaimed. [From W. C. Morgan and J. C. Schermerhorn (1978). Stripping with scrapers and dozers, *Coal Mining and Processing* **15**(6), 73–77, 124, 128.]

TABLE II Major Underground Mining Methods

Method[a]	Deposit geometry [DIP*]	Ore (mineral)/ host rock strength	Typical products	Variations
Room-and-pillar	Bedded (sedimentary); metalliferous veins, pockets, lenses; thin to very thick [Flat, medium]	Competent/competent	Coal, salt, potash, limestone, marble, uranium, copper, oil shale	Random (casual) vs regular pillar and/or room patterns Pillared open stoping Pillar recovery (robbing) vs leaving pillars Conventional (drill and blast) vs continuous (mechanical) mining Step mining; pitch mining Partial vs complete extraction Punch mining Full-face vs multiple slicing
Open stoping	Medium to thick uniform thickness [Medium-flat]	Competent/competent	Metals	Room-and-pillar
Sublevel stoping	Regular thick veins [Medium-steep]	Competent/competent	Metals	Long-hole stoping Vertical crater retreat (VCR)
Shrinkage stoping	High-value regular thin veins [Steep]	Competent/competent	Metals	Overhand shrinkage Blasthole shrinkage
Cut-and-fill stoping	High-grade thin or thick veins [Steep]	Competent/weak to competent	Metals	Horizontal vs inclined Transverse stoping
Undercut-and-fill	High-grade veins	Weak/weak	Metals	Longitudinal or transverse
Timber-supported stoping	High-grade veins or pockets of any shape	Weak/weak	Metals	Square set stoping with or without fill Slot or slice mining Stull stoping
Longwall	Thin to thick uniform large extent [Flat-medium-steep]	Weak to competent	Coal, trona, gold	Retreat vs advance vs advance-retreat or alternating retreat Caving vs stowing (backfilling) Single vs multiple lifts Sublevel caving
Shortwall	Uniform thin to medium thickness [Flat]	Competent/competent	Coal	
Sublevel caving	Medium thickness veins [Medium-steep]	Weak to competent/ weak	Metals	Natural caving vs blasting
Block caving	Massive	Weak/competent	Copper, molybdenum, nickel	Block vs panel vs massive caving
Hydraulic mining	Thick beds [Flat-steep]	Weak/competent	Coal	

[a] Flat: 0–20°; medium: 20–50°; steep: 50–90°.

The most widely used underground mining method is room and pillar mining, illustrated in Figs. 6 and 21. Blocks of the mined seam are left in place to support the overlying rock. These pillars are designed to prevent collapse of the openings yet minimize the amount of valuable mineral left in place or abandoned. In areas with great surface value, e.g., because of the presence of buildings, rivers, railroads, or high-quality farmland, it may be necessary to leave sufficiently large pillars to prevent subsidence. Sometimes it is possible to recover ("rob") all or most of the pillars during retreat mining — mining backward toward the point from which mining of a given section of the seam has been started, once the greatest distance has been reached toward which mining was advancing.

Most stoping methods have been developed primarily for metal mining. They also find application for the mining of industrial minerals with a high value per unit volume. Figure 1 illustrates sublevel stoping. A steeply dipping (i.e., vertical or near-vertical) orebody is subdivided into sections at several main levels and sublevels. A stope is created by mining all the ore in between two levels or sublevels. This approach is possible when the waste rock surrounding the ore (i.e., the hanging wall and

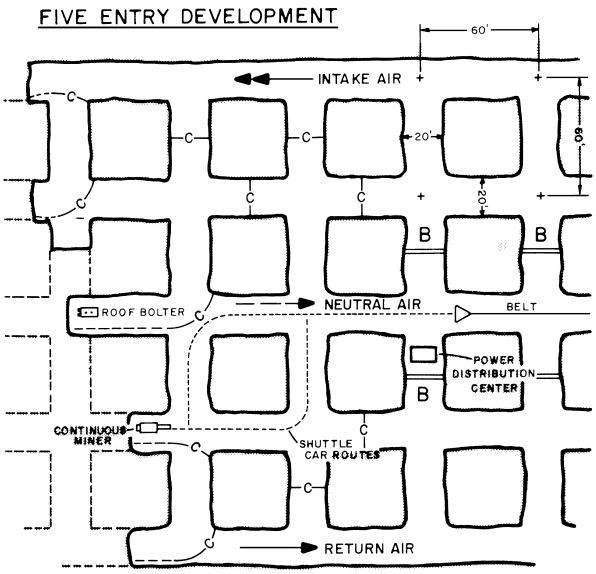


FIGURE 6 Plan view of a typical room and pillar coal mine. Five entries advance in parallel. Squares are coal pillars, left in place to carry overburden (overlying rock). Airflow is controlled by curtains (C) and block walls (B) installed in between the pillars and arranged to assure that fresh intake air sweeps across all faces (edges on the left). A continuous miner cuts coal at one face; two shuttle cars haul the coal to the conveyor belt—along different paths to minimize mutual interference. When the continuous miner reaches the maximum advance length, typically 20 ft (6 m), at the face on which it is working, it will move to the next face (lower entry). Ventilation curtains and shuttle cars (and their power cables) will be adjusted. The roof-bolting machine is installing bolts in the roof of the last place that has been opened up by the miner, and will move to the next face as soon as mining is completed there. [From "Illustrated Underground Mining Methods" (1981). Skelly and Loy, Engineers—Consultants, Harrisburg, PA.]

the footwall) is sufficiently strong to remain stable even though a large opening, the stope, has been excavated. If the rock is not strong enough, a different mining method might have to be considered. In shrinkage stoping, only a small excavation is opened up at any one time. Under even weaker rock conditions, in cut-and-fill mining, the excavations created by mining the ore are backfilled with waste rock. Timbered stoping is used less, because of the high cost of timber, because of its high fire risk, and especially because timbering requires mining expertise that rarely is still available in developed countries. Longwall, shortwall, and underground hydraulic mining are primarily coal mining methods. Longwall mining, a major competitor for room and pillar mining, allows extremely high recovery rates (80 to 90%, and better, compared to 40 to 75%, and frequently less, for room and pillar mining). In 1999 some 62 longwall faces were operating in the United States, all but one in coal. In the United States in 1994, for the first time ever, longwall coal production exceeded coal production by continuous miners. Since then longwall production has continued to grow faster than continuous miner production, and all indications are that the trend toward longwall mining will continue. The productivity gains achieved clearly outweigh the extremely high capital investments (which can range from $30 to $50 million) required to install and develop a longwall. Longwall mining is entirely mechanized and is highly amenable to partial or complete automation. Several prototype automated or semiautomated longwalls have been tried out during the 1990s. It appears likely that such systems will become implemented commerciallly within the near future.

Some mining methods are very specific for particular minerals and geological environments. An example is solution mining of salt. Because salt readily dissolves in water, it can be "extracted" by drilling holes into the salt and circulating water into and out of the formation. Salt is recovered from the brine by evaporating the water. The cavities resulting from the mining frequently are used for storage of hydrocarbons. A technique similar to borehole solution mining of salt is the Frasch sulfur mining system. Hot pressurized water is pumped through holes into a sulfur formation. Molten sulfur is pumped out. Related borehole mining techniques have been under intensive investigation and have found some commercial applications, especially for *in situ* leaching of copper and uranium (Fig. 7), and for *in situ* gasification or liquefaction of coal and shale oil.

Important in mine method selection is engineering judgment based on past experience. Formalized expressions of past experience are tabulated data bases that include mining variables and corresponding results obtained in terms of productivity, capital development and operational costs, manpower and energy costs, safety, and so on. Table III

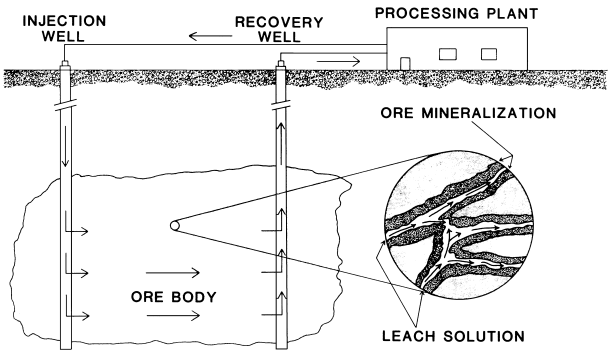


FIGURE 7 *In situ* leach mining. A solvent injected into the orebody dissolves a metal (e.g., copper), which is recovered from the pregnant (metal-bearing) solution in the processing plant. [Courtesy of Jon K. Ahlness, Group Supervisor, In Situ Systems, U.S. Bureau of Mines.]

TABLE III Typical Production Data for Underground Coal Mining Systems[a]

	Seam height of 106–137 cm (42–54 in.), 121 cm (48 in.) average				Seam height of 137–182 cm (54–72 in.), 160 cm (63 in.) average			
	cont.[b]	Conv.	LW	SW	Cont.	Conv.	LW	SW
Production data								
Width of place (face), ft	20	20	400	180	20	20	400	180
Depth of cut (web), ft	20	10	2.5	8	20	10	2.5	8
Tons per cut (pass)	64	32	160	230	84	42	210	302
Minutes per cut[c]	60	29	60	160	65	29	60	178
Work minutes per shift[c]	345	345	345	345	345	345	345	345
Cuts per shift	5.75	12	5.75	2.16	5.31	12	5.75	1.94
Tons per shift	386	384	920	497	446	504	1208	586
Face crew								
Operator	1	2	1	1	1	3	1	1
Helper (snaker)	1	2	1	1	1	1	1	1
Shuttle car Chock operator	2	2	3	3	2	2	3	3
Roof bolters Gate men	2	2	2	1	2	2	2	1
Coal driller Shot firemen		2				2		
Total face crew	6	10	7	6	6	10	7	6
Man minutes per cut								
Continuous miner	120	—	120	320	130	—	120	356
Loading machine	—	58	—	—	—	58	—	—
Shuttle car Chock operator	120	58	180	480	130	58	180	534
Roof bolters Moving gate equipment	120	58	120	160	130	58	120	178
Cutting machine	—	58	—	—	—	58	—	—
Drilling machine	—	29	—	—	—	29	—	—
Shooting	—	29	—	—	—	29	—	—
Total[c]	360	290	420	960	390	290	420	1068
Man minutes at the face per 8-h shift	2070	3480	2415	2074	2071	3480	2415	2072
Man minutes per ton[c]	5.36	9.06	2.62	4.17	4.64	6.90	2.00	3.54
Tons per face man per 8-h shift	64.3	38.4	131.4	82.8	74.3	50.4	172.6	97.7

[a] From R. Stefanko (1981). Table 1, p. 54, *in* "Coal Mining Technology — Theory and Practice" (C. S. Bise, ed.). Society of Mining Engineers of the American Institute of Mining, Metallurgical, and Petroleum Engineers, Inc., New York, New York.

[b] Cont., continuous miner section; Conv., conventional mining section; LW, longwall; SW, shortwall, Metric equivalents: in. × 2.54 = cm; ft × 0.3048 = m; st (short tons) × 0.907 = t (metric tons).

[c] Includes move time to next place but not delays.

shows production achieved for the main underground coal mining methods. Such tables help in making preliminary mining method selection decisions.

Such tables, graphs, or more generally and commonly today, computerized data bases need to be updated regularly in order to reflect the most recent experience, to the greatest extent possible. Changes over time also reflect changes in operating and design practices. For example, most geometrical parameters describing continuous mine section geometries in Table III remain virtually the same in the year 2000, while longwall geometries have changed considerably: longwall panel (face) widths now typically are in the range from 750 to 950 ft, with some significantly exceeding that range; the cut depth typically is 3 ft, the seam height rarely is less than 60 inches, the cut height rarely less than 70 inches, reflecting a concentration on thicker and Western seams. Installed face production capacity usually is at least 5000 tons per shift, most commonly in the range of 8000 to 10,000 tons per shift, with a few installations up to 15,000 tons per shift.

With respect to mining methods, shortwall mining essentially has disappeared from U.S. mining practice, and conventional mining has remained a modest contributor (slightly less than 10% of underground U.S. coal production). Face crew size remains essentially the same. Productivity for longwalls, as a result of the geometry and corresponding equipment changes, has increased dramatically.

IV. MINE FEASIBILITY EVALUATION

An assessment needs to be made, prior to the startup of a mine, as to whether the venture is likely to be profitable. This feasibility study is a coordinated effort by mine geologists and mining engineers, supported by accountants, mineral economists, legal and taxation experts, etc., to make the best estimate of the *in situ* conditions likely to be encountered, of mining methods and equipment that will be used, of the capital and operational costs involved, and of the probable market value of the mineral at the time when it will be produced and marketed. This decision-making process, an effort repeated with further studies on mine geology, mine method selection, and mine planning engineering, ultimately will provide the necessary information to decide whether or not mining of a particular property is feasible, that is, likely to be profitable.

V. MINE PLANNING

Once a mining method has been selected, a detailed schedule is prepared of the sequence of steps to be followed during the excavation of the ore, its removal, and its transport. This plan also includes detailing of auxiliary operations, such as provision of power, ventilation in underground mining, and waste disposal in an environmentally acceptable manner. Mine planning remains a mining engineering responsibility for the life of the mine. Typically, it includes short-term (6 months to 1 yr), intermediate (2–5 yr), and long-term (10–25 yr) plans, prepared in much more detail for the short-term than for the long-term plans.

A simplified example of planning for a surface strip mine is shown in Fig. 8. In this case a plan is prepared for mining a block of coal, commonly several miles wide by several miles (sometimes many miles) long. Mine surveyors have determined the topography of the area. Site exploration has provided the necessary information about depth, thickness, and quality of the coal seam. The primary objective of the mine planner is to provide the most efficient utilization of the dragline (which, given its acquisition cost of several tens of millions of dollars, is likely to dominate mining cost). First the planner has to ensure that the dragline can expose all or nearly all coal. Next a decision needs to be made as to how wide a cut (or pit) will be excavated by the dragline during each pass along the property (this width will depend, for example, on the depth of the coal seam, the room needed on the coal seam for efficient coal loading, and the height and distance to which the overburden waste needs to be dumped). Once the dragline operations have been decided on, access roads for the coal haulage vehicles are designed (Fig. 4), locations for the dewatering sumps and pumps are selected, and a power supply system is designed for the dragline, coal loaders,

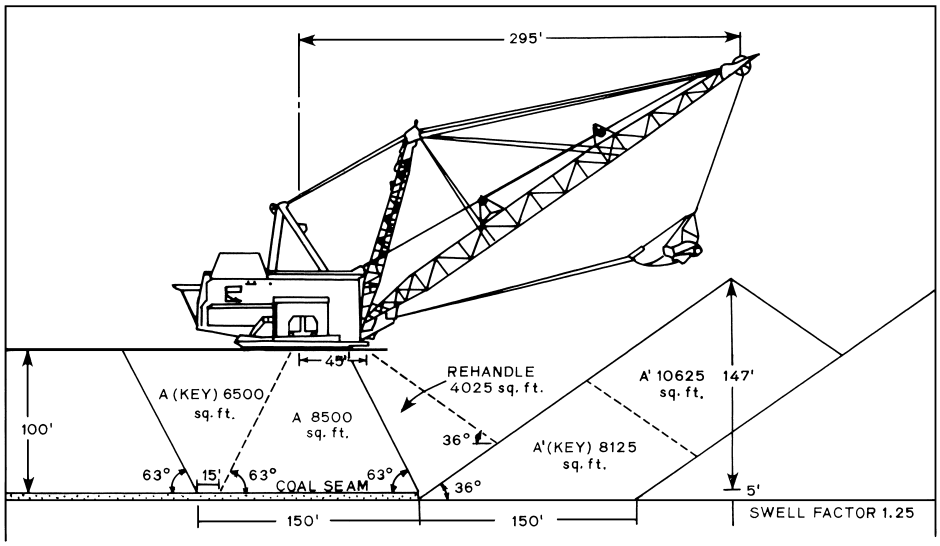


FIGURE 8 Working diagram for the stripping operation of a coal seam by a walking dragline. The working-range diagram is an essential engineering tool for the design and planning of a strip mine, as well as for the selection of the stripping equipment. The diagram shows the depth of the overburden [100 ft (30 m)] that needs to be removed in order to uncover the coal seam, the cuts [A and A(KEY)] that will be dug out by the machine from its present position, and the location [A′ and A′(KEY)] where this spoil will be dumped. The rock increases in volume as it is removed from its initial dense condition to a loose dumped condition, and this volume increase is accounted for by the swell factor, in this example giving a 25% volume increase. Not all overburden can be moved directly to the spoil; some rehandling is required. [From Bucyrus-Erie Company, South Milwaukee, WI.]

pumps, and lights. Blast patterns are designed, and movements of all large (slow) equipment are planned. Numerous variables are involved in the planning: It is likely that a number of different draglines will need to be considered, for instance, and a large number of iterative calculations are required. Virtually all mine planning is computerized and makes extensive use of operations research methods and computerized optimization techniques.

VI. MINE UNIT OPERATIONS

Unit operations refers to separate actions taken as part of the mine cycle to accomplish the overall goal of mine production. In a typical conventional mining cycle, the unit operations include drilling, blasting, loading, and hauling. Drilling provides boreholes within which explosives are emplaced. Detonation of these explosives, or blasting, breaks up the rock into fragments sufficiently small to be loaded and hauled.

A. Drilling of Blastholes

Drilling holes for the emplacement of explosives can be performed by different drilling methods and different types of equipment. The drilling equipment selected depends on whether one is dealing with surface mining or with underground mining. In the latter case space constraints have a significant impact on the equipment to be used.

Commonly used drilling methods are diamond drilling, rotary drilling, and percussion drilling. Diamond drilling is rarely used for blasthole drilling but is a very widely used technique for exploration studies. Both in rotary drilling and in percussion drilling, a drilling tool, the drill bit, is forced against the rock, through a hollow steel cylinder, the drill rod. The force is applied by the drill. As rock is removed by the bit, the machine advances the drill rod. In rotary drilling, the primary rock-breaking mechanism is rotation of the bit against the bottom of the hole. In percussion drilling, the primary mechanism is the repeated impact of the bit against the borehole bottom.

Important for efficient drilling is removal of the drill cuttings from the hole. This is accomplished by means of a drilling fluid, usually air (water in diamond drilling) injected through the drill rod or drill stem, which removes the cuttings along its return path through the annulus between the drill rod and the borehole wall. Reverse circulation drilling, in which the drilling fluid flows down to the drill bit along the annulus, and back up to the surface through the drill string, has become the preferred method for rotary exploration drilling, because it tends to produce more representative "true" rock samples. The drilling fluid also provides the cooling required to prevent overheating of the drill bit. A drill rig for surface mining is shown in Fig. 9.

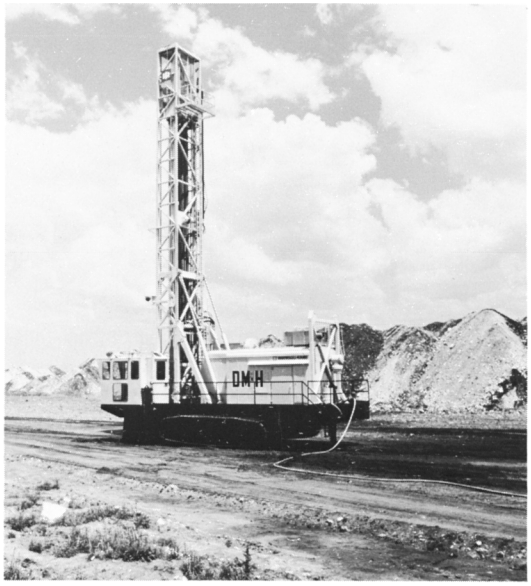


FIGURE 9 Large surface blasthole drill (Ingersoll-Rand DM-H Drillmaster) can drill holes of 9–12$\frac{1}{4}$ inches (229–311 mm) in diameter to a depth of 225 ft (70 m). It can exert a force of 100,000 lb (450 kN) on the drill bit, weighs 200,000 lb (900 kN), and has a 500–700 hp power supply, either electrical or diesel. During drilling the machine is supported on three hydraulic jacks (one can be seen near the back of the drill). For moving to the next hole to be drilled, the jacks are raised and the machine trams to its next position, on its crawlers, at a speed of up to 1 mph (1.6 km/h). (Photograph courtesy of Ingersoll-Rand Company, Rotary Drill Division, Garland, Texas.)

Down-the-hole hammer drilling (DHD) has increasingly gained acceptance, especially in harder rock. In this drilling method the blow piston, or hammer, is located down the hole and directly impacts the drill bit, rather than through the drill string as an intermediary. Particularly for deeper holes this reduces the energy losses otherwise associated with the impact pulse traveling down the drill string. Down-the-hole hammer drilling tends to result in straighter holes and drills faster and more efficiently.

The primary function of the engineer, with respect to drilling, is to select an appropriate drilling method and equipment, including drill bits, rotational speed, axial force, and fluid circulation rates. The appropriate drilling method depends predominantly upon the rock hardness, with rotary drilling the preferred method in soft rock, and percussion drilling in hard rock.

Modern drill rigs tend be highly instrumented, monitored, and computerized. Data collected might include drilling rate of advance, torque and thrust applied to the

bit, rotational speed, power consumption, and various operational variables that indicate the mechanical, hydraulic, and electrical condition of the machine. The resulting data base provides the engineer with factual information on which to base optimized selection of drilling variables, although optimization can require exhaustive data analysis. GPS allows surface drill rigs to be positioned correctly and accurately with minimal external manual supplementary surveying, providing for faster, more economical, and more precise drillhole locations.

B. Blasting

Blasting, or rock fragmentation by means of explosives, is necessary when the rock is too hard or strong to be removed directly with the excavation equipment. Although mechanical equipment exists that can break up almost any rock without blasting, it usually is not economical to remove large volumes of hard rock without blasting.

Explosive rock fragmentation is accomplished by the sudden release of the chemical energy stored in an explosive upon its detonation. In most applications the explosive is installed in boreholes in the rock.

The primary function of the blasting engineer is to design the blasting rounds. A blasting round consists of a number of holes drilled into the rock to be broken. The blast design requires determination of the number of holes needed to break a given volume of rock, determination of the geometry of the holes (i.e., their depths and relative positions and orientations), and selection of an appropriate explosive, of an appropriate detonator or blast ignition method, and of an appropriate detonation sequence (i.e., the sequence in which the holes will detonate).

Blasting patterns for surface mines usually consist of several rows of vertical holes drilled parallel to the "face," the rock surface exposed by previous blasts (Fig. 10). The burden, or distance between the front row of holes and the free face, as well as the spacing between holes within each row, are determined on the basis of the explosive energy packed within the holes, the rock mass strength, and the degree of fragmentation desired. The amount of energy in the holes is a function mainly of the diameter of the holes, which determines the total volume of explosive, and of the type of explosive. A trend has existed for several decades toward the use of increasingly large blasthole diameters. This allows more widely spaced hole patterns,

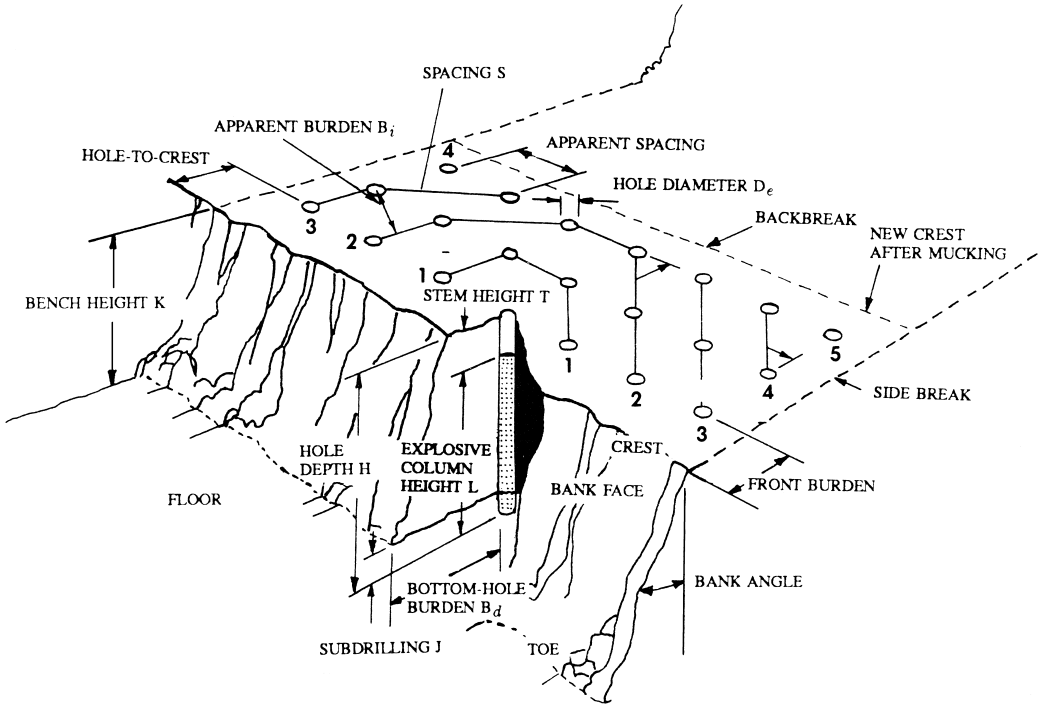


FIGURE 10 Typical blasting pattern for surface mine bench blast. Three rows of seven holes each have been drilled parallel to the bench face and filled with explosives and inert stemming. They will be detonated in the sequence indicated by the numbers on the holes. B. Burden, distance from blasthole to free surface to which it will break rock. L, Explosives column length. J, Subdrilling, depth below the bench floor to which blastholes are drilled in order to assure clean breakage along the bench floor. S, Stemming, length of inert material inserted on top of the explosives column in order to prevent release of explosive energy into the open air. FBF, Free bench face. [Reproduced with permission from Explosive Engineer's Guide, Ireco®, Revised for 1990, Ireco Inc., Salt Lake City, Utah.]

thus reducing the number of holes that need to be drilled. This trend has been facilitated by the parallel trend toward larger loading and hauling equipment, which has made it possible to accommodate the larger rock blocks that result from more widely spaced blasting patterns.

Underground blasting patterns can range from a small number (8–12) of small-diameter (4–6 cm) holes drilled into a coal face, a blast involving not more than 10–20 lb (5–10 kg) of explosives, to a major "coyote" blast, in which entire tunnels or drifts are filled with explosives, up to several million pounds, and detonated simultaneously in order to break up large rock volumes.

Common explosives used in mining are classified in Table IV. The least expensive explosive (ANFO, ammonium nitrate fuel oil) is used for most applications. In bulk form it is simply poured or pumped (vertical surface holes) or pneumatically blown (underground) into the blastholes. It cannot be used in holes where it might come into contact with water, unless it is packaged appropriately. It is not easily detonated; hence, primer and booster charges, cartridges or cans of higher-strength explosives, are included in holes filled with ANFO to assure initiation as well as rapid and complete detonation. Special explosives (permissibles) must be used in underground coal mines. Permissible explosives are designed such that their detonation cannot cause a coal mine explosion.

Several methods exist for initiating the detonation of the explosives column within a hole. The basic initiator is the detonator or blasting cap, a small metal cylinder that contains a small quantity of highly sensitive explosive. The detonator is exploded by sending an electrical current through it, by bringing it in contact with a flame, by means of a blasting fuse, or by impacting it with a shock wave transmitted through a shock tube. Electronic detonators that allow extremely precise blast delay timing have been introduced in the late 1990s. The blasting cap detonation initiates the explosion of the explosive, either directly or through the intermediary of a primer.

The goal of the blasting engineer is to accomplish the necessary fragmentation at a minimum cost (e.g., with

TABLE IV Classification of Explosives[a]

Low explosives, deflagrate rather than detonate, low reaction velocity, no water resistance, highly flammable, very sensitive; e.g., black powder.	High explosives, can be detonated by blasting cap.			Blasting agents, combination of oxidizer and fuel, cannot be detonated by just blasting cap.
Dynamites, composed of nitroglycerine and filler material.	Gelatin dynamites, combination of nitroglycerine and nitrocellulose to form rubber-like material.	Semigelatin, combination of ammonia gelatin and ammonia dynamite, lower strength, good fume rating, good water resistance.	Two component, each individual component is nonexplosive, can be shipped by any means.	Bulk mixed compounds, ammonium nitrate and fuel oil mixed on the job by the consumer; very economical; good fumes if mixed properly; no water resistance.
Straight dynamite, most sensitive commercial explosive; weight strength is equivalent percent of nitroglycerine.	Straight gelatin, excellent water resistance, very cohesive, fumes vary with strength.			Premixed NCN, nitrocarbonitrates, mixtures prepared by manufacturer, light density, no water resistance.
Ammonia dynamites, ammonia nitrate substituted for part of nitroglycerine. Three subclasses: low density, high density, permissibles.	Special gelatin, some nitroglycerine replaced by ammonium nitrates; good water resistance.	Water gels and slurries, can be classed as high explosive or blasting agent, depending on cap sensitivity and chemical composition mixture of oxidizers, fuels, and sensitizers that contain 5–40% water; available in a wide range of densities; commonly packaged in polyethylene tubing; excellent water resistance.		

[a] Reproduced with permission from J. S. Malesky (1981). Blasting, Ch. 7, pp. 140–186, "Elements of Practical Coal Mining." 2nd ed. (D. F. Crickmer and D. A. Zegeer, ed.). Society of Mining Engineers of the American Institute of Mining, Metallurgical, and Petroleum Engineers, Inc., New York.

minimum drilling and explosives cost) while maintaining undesirable side effects at an acceptable level. Typical undesirable side effects that must be controlled include flyrock, the throwing of individual rock blocks to excessive distances, ground and air vibrations (airblast, or blast-induced noise), and excessive overbreak, or damage to the rock intended to remain in place. Flyrock clearly is a safety hazard. Ground and air vibrations are primarily an annoyance, although they can be damaging under extreme conditions. If the rock to remain in place is extensively damaged, the stability of the excavation, whether underground or on the surface, can be affected. Special controlled blasting procedures can largely eliminate such damage.

The degree of fragmentation required for any particular blast depends primarily upon the size of the equipment that will be used to load the broken rock. The size of this equipment in turn depends largely on the size of the hauling equipment, once again indicating the need for the mining engineer to take a systems approach (i.e., to consider simultaneously all aspects of the operation).

Fragmentation obtained from a blast, or size distribution of the broken rock, is determined to a significant extent by the size distribution of the rock in place, before it is blasted. Virtually all rock is intersected by weakness planes such as joints, bedding planes, or even fractures from nearby earlier blasts. These weaknesses have a pronounced effect on the size of rock fragments found after blasting. Particularly important are the spacing of weaknesses and their orientation.

C. Loading

Broken rock is loaded into haulage vehicles (e.g., trucks, railroad cars) or systems (e.g., belt conveyors) for removal out of the mine. Many types of loading equipment are used both underground and on the surface (e.g., front-end loaders and shovels). Some equipment is used underground primarily or exclusively and has been designed specifically to operate efficiently in very restricted space. Some equipment is intended virtually exclusively for surface mining, where unencumbered operations permit the use of very large units, such as dredges, draglines, and bucket–wheel excavators.

The conventional mining shovel (Fig. 11) is widely used, although alternatives such as front-end loaders and hydraulic excavators (Fig. 12) are finding increasing applications.

The dominant type of front end loader, by far, is wheel-mounted and has considerably more flexibility in operations than either shovels or hydraulic excavators. Because they are wheel-mounted, they can travel at relatively high speed, and hence they can be functional at



FIGURE 11 Large open pit mining shovel loading a haul truck. An electrically powered P & H 2800 shovel with 36 cubic yard dipper is loading a 170-ton capacity Lectra Haul Mark truck. Size of equipment can be estimated by comparison with person in left foreground. [Photograph courtesy of Harnischfeger, P & H, Milwaukee, Wisconsin.]

more than one point, not a practical mode of operation for track-mounted shovels or hydraulic excavators, whose moves are slow, and whose moves are minimized in order to minimize equipment damage. Shovels and hydraulic excavators have considerably larger breakout forces than front end loaders and hence can efficiently dig in muckpiles that may be much tighter, less easy to load, than those required for front end loaders. As a result blasting requirements may be less demanding when shooting rock to be loaded with shovels or hydraulic excavators than with front end loaders. The overall engineering problem once again is that of balancing several contradictory requirements and striving for an optimum overall solution within multiple constraints.

Similarly to other mining equipment, during the late 1990s loading equipment has grown in size and has become intensely instrumented, with the addition of extensive electronic components and computerized monitoring and communication systems. Shovel bucket



FIGURE 12 Large hydraulic excavator (Marion 3560) can be operated as either a front shovel or a backhoe (options illustrated in Fig. 28). In operation, the machine weighs about 626,000 lb (284,550 kg) and can travel up to 1.34 mph (2.16 km/h). [Photograph courtesy of Marion Power Shovel Division, Dresser Industries, Inc., Marion, Ohio.]

sizes have reached 80 cubic yards, with a gross machine weight of the order of 2,500,000 pounds, and installed hoisting power of 2000 to 3000 hp. Large shovels virtually uniformly are powered from electrical power distribution systems, whereas hydraulic excavators most commonly are driven by diesel engines, and hence are fully self-contained with regard to power. This independence from a power distribution system, combined with their somewhat higher traveling speed (up to a few miles per hour), makes them somewhat more mobile than conventional rope shovels (the latter also are sometimes referred to as electric shovels).

A major development during the late 1990s was the introduction of GPS (Global Positioning System) to guide loading operations. Reducing the need for manual surveying support lowers personnel requirements, increases safety by reducing personnel exposure, and results in more accurate and faster loading positioning. It also is an extremely useful step toward full automation of the loading function.

In most underground coal mines no distinct loading equipment is used, as loading is an integral part of the excavation machine (e.g., continuous miner or longwall shearer or plow). In mines where coal or ore is blasted, finger, disk, or gathering arm loaders (Fig. 13) are used frequently in intermediate-sized openings; overshot, overhead, or rocker loaders or machines (Fig. 14) in small excavations; and equipment similar to and adapted from surface loading equipment (e.g., front-end loaders, shovels) in large openings. In many modern mining operations, the greatest possible use is made of load–haul–dump vehicles (LHDs) (Fig. 15). These units perform both the loading and hauling functions, thereby eliminating the need for separate hauling equipment. This reduces capital investment, simplifies maintenance, reduces the number of operators, and, probably most importantly, improves flexibility in the selection of mining procedures.

A singularly important development over the past decade is that of remote-controlled underground loading equipment. The operator of the loading equipment remains some distance away from the equipment and operates it from a console connected to the equipment through a cable or a laser beam. This greatly reduces the

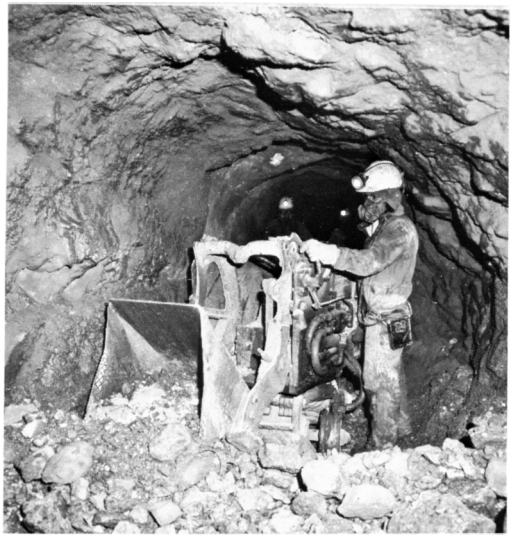


FIGURE 14 Overhead or overshot loader (Eimco Rockershovel 12B) loads broken rock ("muck") by pushing the bucket into the muckpile, then swinging the bucket overhead to dump the muck into the car behind the mucker. [Photograph courtesy of Eimco Mining Machinery International, Salt Lake City, Utah.]

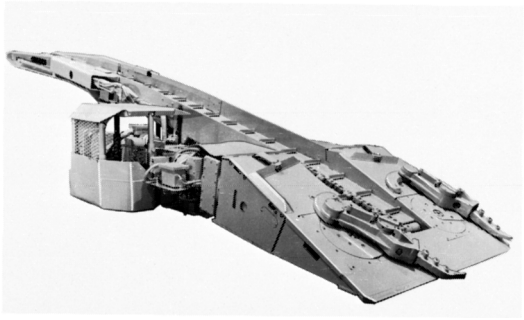


FIGURE 13 Gathering arm underground mine loader (Joy 14 BU 10-11B). The eccentrically rotating arms in front of the machine load broken rock into the chain conveyor. The conveyor moves the rock through the boom, which can be raised or lowered, and swung in different directions, thus facilitating the loading of a shuttle car (Fig. 17), which is placed with the front of its conveyor underneath the tail of the loader. This loader is rated for 12–25 tons per minute and can operate in seams as low as 38 in. (97 cm) thick. [Photograph courtesy of Joy Manufacturing Company, Pittsburgh, Pennsylvania.]



FIGURE 15 Load–haul–dump (LHD) vehicle, widely used in trackless underground mining (GHH-Fahr Lader LF-7.3). The vehicle is articulated in the center, allowing it to turn sharp corners. This model has a loading capacity of 7300 kg, a travel speed of up to 29 km/h, and is powered with a 136-kW (185-hp) diesel engine. [Photograph courtesy of M.A.N. Unternehmensbereich GHH Sterkrade, Oberhausen, Germany.]

hazards, in particular rock falls, to which the operator might be exposed. During the late 1990s a further improvement has been the introduction of remotely operated loading equipment where the operator no longer requires a direct line of sight: The operator observes the equipment loading indirectly, by means of video cameras installed strategically near the loading location. Full automation is under intense investigation by major mining companies, equipment suppliers, and research centers, and is likely to be introduced early in the 21st century.

D. HAULING (TRANSPORTATION)

The dominant haulage method in surface mining is mine trucks, typically large off-road trucks with haulage capacities of up to 360 tons (Fig. 11). Some open pits use railroad haulage or belt conveyors, the latter also frequently being used in conjunction with bucket–wheel excavators (Fig. 16), in operations where extremely large volumes of every soft rock (usually soil) are handled.

The introduction of haulage trucks in the 320- to 360-ton range during the late 1990s has been made possible by the fact that at least one tire manufacturer finally appears to have succeeded in the production of reliable tires for this class of vehicle. A typical 360-ton truck haulage truck is nearly 50 ft long, nearly 30 ft wide, over 21 ft high when loaded, and over 45 ft at the highest point when dumping. Fully loaded, the truck weighs well over 1.1 million pounds. The truck is powered by a 2750 hp diesel engine. While the introduction of trucks with a jump in capacity over the previous generation has the potential for substantial productivity improvements, it also poses serious engineering challenges. In order that full productivity gains actually be achieved, the loading equipment must match the equipment that is being loaded. Haulage roads as well as dumping points may need reengineering. Typically for the introduction of major new equipment, the impacts cascade through the entire mining cycle and mining operations, and may require pervasive reengineering in order to reoptimize production.



FIGURE 16 Bucket–wheel excavator. The wheel at the right side of the machine is equipped with buckets and is pushed into the ground while rotating. The buckets dig out the ground and unload it onto belt conveyors. The "stacker" on the left side, somewhat hidden, allows dumping at many different positions. The machine can handle 5280 m^3 (over 10,000 tons) per hour and can dig to a height of 32 m. The size of the machine can be visualized by noting the bulldozer in the forefront, and the person near the front track. [Photograph courtesy of Krupp Industrietechnik GmbH, Duisburg-Rheinhausen, Germany.]

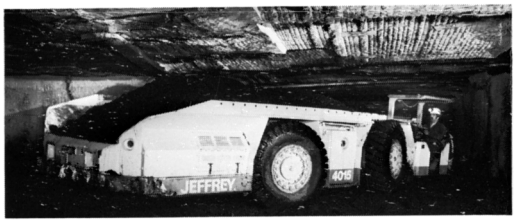


FIGURE 17 Shuttle car (Jeffrey 4015 hauls coal from the face (active coal excavation point) to the main haulage system (e.g., conveyor belts, railroad) that takes coal out of the mine. Coal is loaded onto the car from one end, e.g., from the tail boom of a loader (Fig. 13) or of a continuous miner (Fig. 21) and is moved forward by the segmented steel conveyor until the car is full. The full car shuttles to its unloading point. The conveyor unloads the coal, and the car shuttles back to the coal face. [Photograph courtesy of Jeffrey Mining Machinery Division, Dresser Industries, Inc., Columbus, Ohio.]

In underground mines, especially large ones, a distinction needs to be made between main haulage, intermediate or secondary haulage, and face haulage. Face haulage, or transportation of the broken rock away from the face, is frequently accomplished by shuttle cars (Fig. 17) or by underground mine trucks, although, as mentioned earlier, load–haul–dump vehicles are finding growing applications in this area. Intermediate and main haulage usually is on belt conveyors or by train. In deep mines, ore is usually hoisted from the mine through vertical shafts. Intensive research and development is in progress to replace these mechanical haulage systems by hydraulic or pneumatic methods, for which the broken rock is crushed into small particles so that it can be transported through pipelines. A similar objective in open-pit mining consists of breaking the ore close to the point of excavation in a portable crusher and hauling it out of the pit on a conveyor belt. This method also replaces and intermittent or discontinuous and labor-intensive haulage system (i.e., trucks) by a continuous flow.

Designing and optimizing transportation systems is a major task for the mining engineer, because transportation constitutes a major operational cost. Optimizing haulage relies heavily on computer simulation and on operations research procedures. Simulations are performed by numerically modeling the haulage cycle. Simplified, $CT = LT + TT + DT + WT + BT$, where: CT = cycle time;

LT = loading time; TT = travel time; DT = dump (unloading) time; WT = waiting time; and BT = breakdown time.

Loading time depends on the loading machine, on the haulage vehicle, on the operators of both, and on the ease of digging. Travel time depends on the quality of the road traveled, its grade, the load, and the vehicle power and deadweight. Travel time frequently is further subdivided, as a minimum into haulage time, during which the vehicle is loaded, and return time, during which the vehicle is empty. For open-pit mines, travel time is simulated based on the characteristics of individual sections of the haul road, in particular its grade, and of the haulage trucks. Dump time should be fairly constant for a well-designed dump site. The mining engineer strives to minimize the wait time. Wait time occurs primarily at the loading site and at the dump site. Multiple vehicles may approach both end points of the haulage path simultaneously. Breakdown time can be considered as a random variable and based on observational records.

All variables in the cycle time are characterized by a certain randomness, for example, as a result of weather influences and variations in equipment and in operator performance. The analysis of haulage cycles can be refined by including a mathematical description of the randomness in all or in some of the terms in the cycle time equation.

Multiple loading and dumping points, haulage roads, and haulage vehicles must be accounted for in a complete simulation of a typical mine production. The production per shift for each vehicle can be approximated by the average payload carried by the vehicle multiplied by the number of cycles completed during the shift. The number of cycles can be calculated from the cycle time. The sum of the production figures for all haulage vehicles gives the total production per shift. Simulations of haulage systems find numerous applications in mining engineering, both for premining planning and during operations. The more important ones are the selection of loading and hauling equipment and the selection and design of haulage roads. Optimization of any of these variables, and especially of the overall haulage system, typically requires the execution of a large number of simulations, given the wide range of options available for all or most of the input variables.

Centralized computerized dispatching has become standard operating practice in most large open-pit mines. With such systems the dispatcher continuously remains informed about the operating status of each piece of equipment in the fleet. The dispatcher assigns equipment duties at the beginning of each mining shift. Over the course of the shift, based on information transmitted from the equipment, reassignments can be made, for example if a shovel breaks down, or a truck. Operations research applications can be built into the system to allow optimization of the operations.

While transportation of rock is the most obvious aspect of mine haulage, the logistics involved in delivering supplies (e.g., ground support, explosives, lubricants, fuel), equipment, and personnel to all points in the mine can be equally complex and are equally important in ensuring efficient operations.

VII. AUXILIARY MINING ENGINEERING DISCIPLINES

Mining engineering is a multidisciplinary endeavor. The mining engineer calls on a variety of engineering disciplines to engineer mine operations. Several of these are used frequently and have developed into subspecialties that can be considered and have developed into subspecialties that can be considered an integral part of mining engineering practice.

A. Mining Geomechanics

Geomechanics is the study of the mechanical behavior of geological materials. The engineering aspects of these studies, or applied geomechanics, describe the response of the geological environment to the actions of mining. The mining engineer applies geomechanics to predict the response of the rock mass to mining activities. The primary geoengineering function in surface mining is to ensure stable slopes that allow maximum ore recovery with a minimum of waste haulage. In underground mining, geomechanics or rock mechanics responsibilities include the design of mine pillars and mine openings, predictions of ore caveability, support and reinforcement design, blast design, and subsidence predictions.

The responsibilities of the open-pit rock mechanics engineer change over the life of a pit. During the premining phases (e.g., exploration, mine design, mine planning), as well as during the initial mining period, the rock mechanics engineer concentrates on obtaining as much information as possible about those features of the rock mass that are likely to influence the stability of the rock slopes. This includes primarily weakness planes intersecting the rock mass, such as faults, joints, and bedding planes, as well as water pressures that act on these weaknesses. This information is used to identify likely failure modes of the slopes and to design slopes that will be sufficiently stable to allow safe and economical mining by preventing major failures as well as an excessive number of small failures. One characteristic rock-slope failure type is a wedge failure, illustrated in Fig. 18.

As mining progresses, more emphasis is placed on direct observation of slope behavior, via instruments placed on and within the slope. Instruments primarily are used

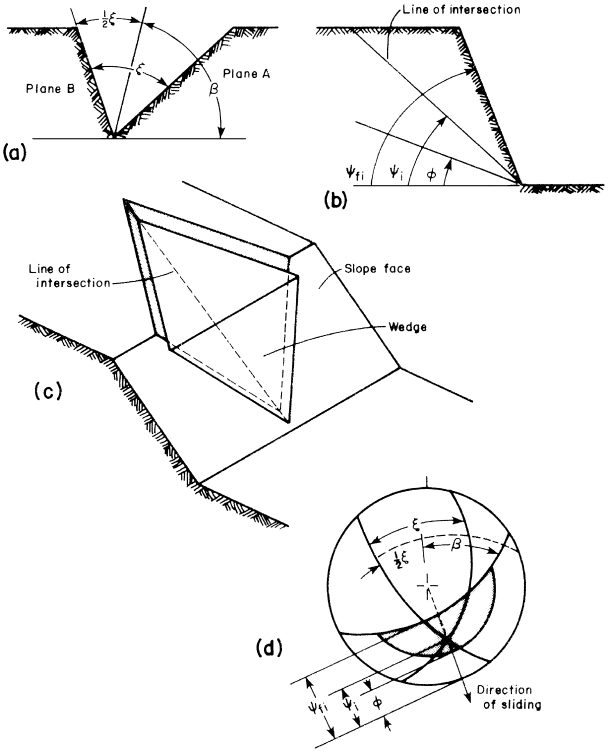


FIGURE 18 Geometry of a wedge failure in a rock slope. The wedge is formed by two intersecting weakness or discontinuity planes in the rock mass (planes A and B). Their intersection daylights in the slope face. ψ_{fi}, slope face dip angle or inclination; ψ_i, dip angle of the intersection line of planes A and B; ϕ, friction angle along the weakness planes. If the wedge is supported by friction only along planes A and B, the safety factor for the wedge can be calculated as

$$F = \frac{\sin\beta}{\sin(\xi/2)} \cdot \frac{\tan\phi}{\tan\psi_i}$$

(a) View along the line of intersection. (b) View at right angles to the line of intersection. (c) Pictorial view of wedge failure. (d) Stereographic wedge failure analysis. [From E. Hoek and J. W Bray (1977). "Rock Slope Engineering," 2nd ed., The Institution of Mining and Metallurgy, London.]

to monitor slope movements or displacements. Of particular concern are rapid slope accelerations. In order to improve the timeliness of the obtained data, instruments can be equipped with telemetric transmission of the data to a central location, e.g., engineering office, where the data is analyzed with respect to slope stability and is used as a basis to decide whether or not operations can continue safely below the slope. Mapping of geological features, especially structural ones, remains an integral part of slope engineering. The broadening information bases allow optimization of the slope design. Traditionally, toward the end of the mine life, a main objective has been to identify the steepest possible slope at which mining can be concluded. This allows maximum ore recovery with the minimum excavating, hauling, and dumping of waste rock. In order to facilitate an environmentally acceptable mine closure plan it may no longer be considered acceptable to close the pit at the steepest angle that may be physically possible.

The underground rock mechanics engineer deals with a wide variety of problems, depending on the type of mining method, the type and value of the recovered minerals, the strength of the surrounding barren rock, the importance of surface protection, and many other variables. Responsibilities might include mine pillar design (Fig. 19), support or reinforcement design for shafts and drifts, assessing the feasibility of caving an orebody, monitoring the stability of a mine by means of field instruments, and providing geomechanical input to the mine method selection, mine design, and mine planning engineers, as well as to the production managers.

Pillar loads frequently are calculated from the tributary area theory:

$$\sigma_p = \sigma_v \frac{A_t}{A_p} = \gamma z \frac{A_t}{A_t - A_m} = \frac{\gamma Z}{1 - ER}$$

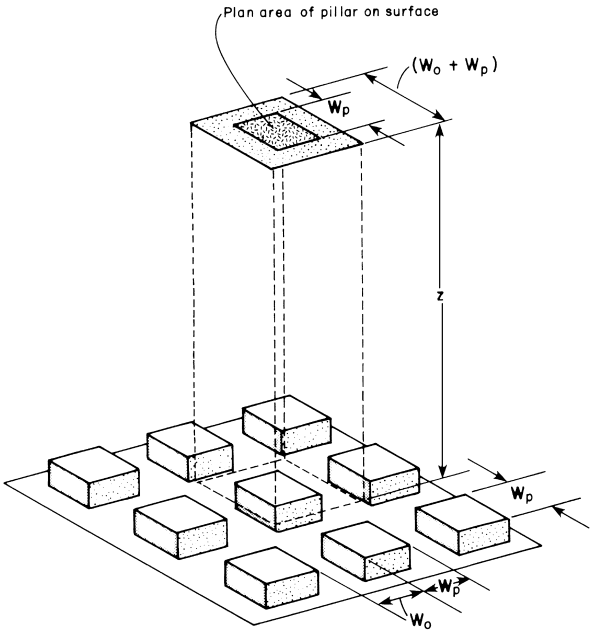


FIGURE 19 Typical room and pillar layout showing load carried by a single pillar, assuming total rock load to be uniformly distributed over all pillars. The load can then be calculated from the tributary area, giving

$$L_p = \gamma \cdot z \cdot (W_p + W_o)^2$$

where L_p is total load on the pillar, γ is average overburden unit weight, z is average pillar depth below the surface, W_p is pillar width (square pillars), and W_o is entry or drift width (assumed identical for openings in both directions). [From E. Hoek and E. T. Brown (1980). "Underground Excavations in Rock," The Institution of Mining and Metallurgy, London.]

where σ_p = average vertical pillar stress; σ_v = virgin (undisturbed) average vertical stress; γ = average unit weight of the overburden; z = depth below the surface; A_1 = tributary area; A_m = mined-out area; A_p = horizontal pillar cross-sectional area; and $ER = A_m/A_t$ = extraction ratio.

For pillar design purposes one compares pillar strength with pillar stress. In areas where mining already has been conducted, it is preferable to determine pillar strength based on *in situ* observations. When no such information is available, pillar strength can be estimated from laboratory testing of rock samples and of weakness planes (e.g., joints, bedding planes) intersecting the pillar rock.

The design of underground openings such as shafts and drifts usually is based on a combination of theoretical stress analysis and empirical rock mass classifications.

A topic of particular concern when extracting a large fraction of an underground seam or formation is the induction of surface subsidence. When the rock overlying mined-out areas is allowed to cave into the open space created by mining, a progressively upward-moving cave can develop and eventually reach the surface. This surface subsidence, if excessive, can damage structures. Particularly susceptible are bridges, tall or long buildings, and railroad tracks. The water-table level can be affected by subsidence, which in turn can damage vegetation. Surface subsidence can be predicted through a combination of field instrumentation monitoring and theoretical analysis. Its effects can be controlled by partial backfilling of the mined-out openings or by limiting extraction, that is, by leaving pillars in place.

B. Ground Excavation

Removing materials from the Earth's crust remains a major responsibility of mining engineering. The selection and design of the excavation methods is governed by the consistency or strength of the ground to be removed: For instance, soft soil or hard rock requires different excavation procedures. The selection also depends on the volumes to be excavated and on the distances over which they need to be transported.

Soils (i.e., soft, distintegrated rocks such as sands or clays) are excavated with equipment such as draglines (Figs. 4, 8, and 20), bucket–wheel excavators (Fig. 16), backhoes (Fig. 28), mechanical or hydraulic shovels (Fig. 11), scrapers (Fig. 5), front-end loaders, or hydraulic excavators (Figs. 12 and 28). Rocks are excavated either by blasting or by mechanical rock removers. The selection of the correct method for ground excavation is a critical responsibility for the mining engineer.

Mechanical excavation (i.e., excavation without blasting) is the most widely used method for mining coal. In underground coal mining, this is usually accomplished with a continuous miner. The most common type of continuous miner (Fig. 21) cuts the coal by pushing into the coal a rotating drum on which hardened steel bits are mounted. In an alternate mining method, longwall mining, which has been used for several decades in European mines and is finding growing applications in the United States, the coal is cut by a shearer (Fig. 22) or by a plow pulled along a coal face, which typically has a length of a few hundred yards.

A fairly recent development that has found rapid acceptance because of its safety and operational advantages is remote control of continuous miners. The operator runs

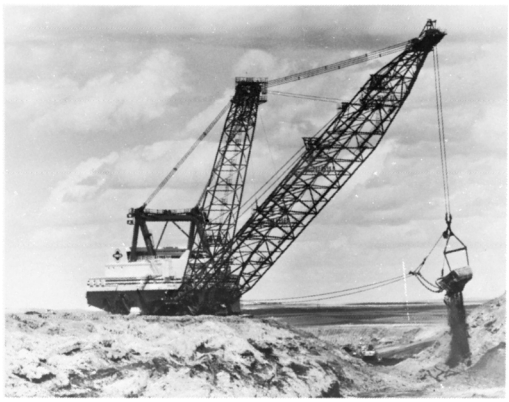


FIGURE 20 Large walking dragline (Bucyrus-Erie 2750-W 400-ft boom; 90 cu. yard bucket) uncovers a coal seam by digging out the waste rock (in this example a glacial till, left, slope directly in front of machine) on top of the coal, and dumping the waste into an area where the coal already has been removed. Pickup truck on side of dragline indicates size of machine. Coal loading is in progress in lower pit. [Photograph courtesy of Bucyrus-Erie Company, South Milwaukee; reproduced with permission from Prairie Coal Ltd., Regina, Saskatchewan.]



FIGURE 21 Continuous miner (Lee-Norse Model CM245—The Hustler) cuts coal by pushing the rotating drum, equipped with cutting bits, into the coal and by swinging the boom on which the drum is mounted up and down. The coal is gathered by the rotating arms in the lower front and conveyed through the machine to the tail where it is dumped into a shuttle car. This particular machine, designed for low-seam coal mining, can operate in seams as thin as 30 in. (76 cm), cuts a 10-ft (3-m) wide entry in one pass, weighs 56,500 lb (25,600 kg), and loads 6–10 tons of coal per minute. [Photograph reproduced with permission from Lee-Norse Company, a subsidiary of Ingersoll-Rand Co., Bristol, VA.]



FIGURE 22 Shearer loader (Eickhoff EDW-150-2L) cutting coal on a longwall face. The drum on the left rotates into the coal, while the shearer advances parallel to the coal face, to the left. The loosened coal falls onto the conveyor (on the floor), which hauls it to the end of the longwall face, from where coal is hauled out of the mine by belt conveyors. The shearer rides back and forth along the coal face, cutting in both directions. Roof supports, above machine and operator, prevent caving. Hydraulic jacks for the support can be seen on the right, behind operator. [Photograph courtesy of Eickhoff Corporation, Pittsburgh, PA.]

the miners through a portable console that allows the operator to rove away from the machine, thus improving operator safety and visibility. A topic of particular interest for longwall automation is the introduction of self-advancing support shields, in which the forward move of the support shields is initiated by the passing of the shearer or plow along the face. In-seam guidance systems that maintain the cutting action within the coal seam, while ensuring cutting as nearly as possible to roof and floor, i.e., minimizing any coal losses, is a topic of great interest for both longwall and continuous miner mining.

In metal mines in which the rock is not excessively strong, mechanical excavation frequently is accomplished by roadheader machines. Full-face tunnel boring machines (TBMs, or moles), in which, as implied by the name, the entire face is bored out in a single pass, rather than by the movement of the boom across the face characteristic for roadheaders, are less frequently used in mining. Full-face boring is used widely for sinking or raising vertical shafts.

C. Ground Control

Ground control refers to the engineering measures taken to prevent the ground around an excavated mine opening from moving into the opening in a manner that might endanger personnel or equipment, or in a manner that might preclude or excessively hamper future mining operations. The design requirements are to provide for a controlled movement of the ground.

In surface mining, ground control refers to the maintenance of stable slopes, usually the responsibility of the slope or rock mechanics engineer, and to the prevention of loose rock blocks tumbling down a slope, usually the responsibility of the operations department, followed up as necessary by safety inspections. Although technology for artificially strengthening slopes does exist — for example, by installing steel rods or cables (anchors, bolts, tendons) into the slopes — such methods are seldom used in surface mining, because they usually are not cost-effective.

In underground mining, ground control is essential to make mining feasible at all, by maintaining open the necessary access excavations. Equally important, once the overall stability of the mine is ensured, is the necessity to minimize local ground instabilities that might endanger personnel or equipment.

For underground mining one usually distinguishes between self-supported and reinforced openings. In the former case, the openings are designed and excavated such that the strength of the rock adjacent to the excavation is sufficient to maintain stability. The most widely used rock-strengthening method is the installation of rock bolts. Although many types of rock bolts exist, most commonly used are mechanical bolts and grouted bolts. Mechanical bolts consist of a mechanical anchor, a steel rod or bolt, and a bearing plate. Anchor and bolt are inserted in a hole in the rock, and tightening the anchor clamps the bearing plate against the rock surface (Figs. 23 and 24). For a

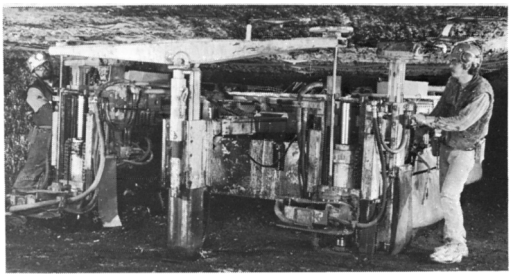


FIGURE 23 Bolt installation in the roof of an underground coal mine. A Fletcher Model DDR Dual Head Roof Bolter allows simultaneous installation of two bolts. The T-frame (TRS, temporary roof support) in front of the machine supports the roof until the bolts are installed, thus providing a safe roof condition for the operators (also called "roof bolters" or "bolters"). Features of machine shown: Both booms can swing up to 6 ft 3 in. (1.9 m) out from center and can tilt up to 90°, thus allowing a wide variety of bolt installations. Machine length, 17 ft 6 in. (5.3 m); width, 9 ft (2.7 m); height, 4 ft 4 in. (1.3 m); weight, 32,000 lb (14,515 kg); tram speed, 0–1.5 mph (0–2.4 kmph). [Photograph courtesy of J. H. Fletcher & Co., Huntington, WV.]

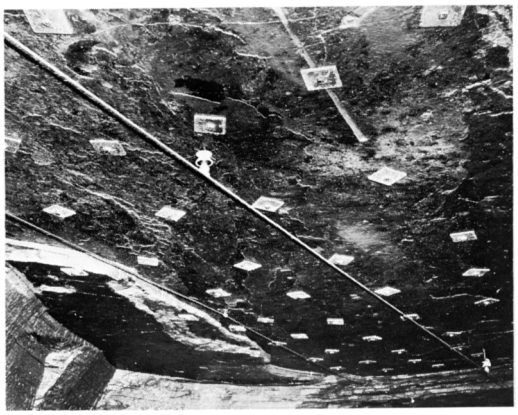


FIGURE 24 A typical rock bolt installation as seen on the roof of a mine drift or entry. Note bolt plates. [Photograph courtesy of Frazer & Jones Division of the Eastern Company, Syracuse, NY.]

grouted bolt the mechanical anchor is replaced by a grout, usually a polyester resin or a cement, which fills the annulus between the bolt and the hole. The resulting intimate contact between bolt and rock over the entire bolt length results in improved ground reinforcement, at the price of a more expensive system. Several systems, sometimes referred to as friction bolts, have been introduced that provide full-length contact reinforcement without the need for grouting. Providing even more support, a relatively flexible support that can be installed to considerable depths, are cable bolts, typically steel cables grouted in boreholes. These rather expensive reinforcement systems can tolerate large deformations without being overloaded, and hence have proven especially attractive and successful in large underground stopes under difficult ground conditions.

While bolting has improved underground safety considerably, it has frequently been a bottleneck in the production cycle, especially in underground coal mining. This problem became more obvious late in the 20th century, with the introduction of ever more powerful high-speed production equipment. In response semiautomatic, or at least highly mechanized, bolting machines have been introduced. In this equipment, the drilling, the bolt installation, or both are largely automated.

When rock strengthening by means of bolting is insufficient to maintain stability of an excavation, it becomes necessary to install supports within the excavation itself. Commonly used internal support methods include shotcrete, cast-in-place concrete, steel sets, and, rarely today, timber sets. Shotcrete is a pneumatically applied concrete, that is, concrete shot against the rock walls by means of a compressed-air-driven shotcreting gun (Fig. 25). Shotcrete usually is applied in relatively thin layers, typically about 5 cm thick (although multiple such layers can be applied).



FIGURE 25 Shotcrete application. A shotcrete gun is used to apply concrete, driven by compressed air, onto a rock surface. Welded wire mesh prevents falling of loose pieces of rock and provides additional support strength as it becomes fully embedded in shotcrete. [Photograph courtesy of Eimco Mining Machinery International, Salt Lake City, Utah.]

A major ground support development during the 1990s has been the addition of fibers, either steel or fiberglass, to shotcrete. The resulting major improvement in tensile strength and in fracture toughness of the shotcrete provides for a ground support product with significantly improved characteristics and has further broadened the applicability range of shotcrete for underground support and reinforcement. Cast-in-place concrete (concrete poured against the rock walls behind concreting forms) usually is applied in much thicker layers, rarely less than 20 cm.

Prior to the installation of any ground control system it is essential for the miner to scale, or bar down, any loose rock. During the late 1990s mechanical scaling systems have become available and have found acceptance to replace the manual scaling that had been one of the more dangerous operations in underground mining.

D. Materials Handling

A typical modern mine handles large volumes of materials, especially of ground, usually broken rock or soil. For a large surface mine this can be on the order of hundreds of thousands of tons a day, for an underground mine on the order of many tens of thousands of tons a day. The systems that handle such volumes must be designed to ensure efficiency, economy, and reliability. The selection and design of materials handling systems incorporate such factors as the volumes to be handled, distances to be traveled, topography or changes in elevation along the route, and type of material to be handled. Some aspects of materials handling are discussed in the sections on loading, hauling, ground excavation, equipment selection, and drainage.

E. Mine Ventilation

In underground mines it is necessary to provide a sufficient flow of fresh air to ensure a healthy and safe working environment. In large mines this can require delivering huge quantities of air through a complex network of airways. Engineering the mine ventilation system requires first the identification of air needs, that is, the amount and quality (temperature, humidity) of air to be supplied at various locations in the mine. These needs can be based on a variety of requirements, such as dilution of gases, temperature control, and dust control. On the basis of air requirements, air courses (e.g., a combination of shafts and mine entries) are designed, and fans (both main and auxiliary, if needed) are selected.

Dilution of gases (i.e., reduction of the gas content in the air) usually is the factor controlling minimum air requirements in coal and uranium mines. In coal mines the gas of primary concern is methane (CH_4). Methane, in concentrations between 5 and 15% in air, is highly explosive, and an air–methane mixture within this concentration is readily ignited. Coal mines therefore are subject to stringent legal requirements aimed at preventing such explosions, primarily through strict maximum concentration limits, intended to prevent the formation of explosive mixtures, and through the use of equipment designed to prevent the formation of ignition sources (e.g., sparks, excessive temperatures).

Radon gas, found commonly in uranium mines, requires dilution to a level that usually governs ventilation air requirements in uranium mines. In underground mines with a large fleet of diesel-powered vehicles, the need to reduce diesel exhaust concentrations to acceptable levels frequently governs minimum air requirements.

In deep mines, especially in locations with a high thermal gradient (i.e., a rapid increase in rock temperature with depth), the need to reduce the temperature in the workplace to an acceptable level frequently determines the minimum amount of fresh air that needs to be supplied. In deep, hot mines, it often becomes necessary to provide air conditioning, to cool and dry the air. In principle, the requirement exists to provide sufficient oxygen to sustain human life. In practice, during normal operations all underground mines provide far more fresh air than is needed for this purpose.

Once the required air quantity has been decided on, it becomes necessary to design the appropriate means for delivering the necessary air to each point where it is needed. As usual in mining engineering, it will be necessary to take an integrated and iterative design approach, taking into account the many other requirements imposed on the underground openings. Usually the fresh air is delivered to the mine working level or levels through vertical shafts. These shafts have multiple access functions: For example, they provide ingress and egress for mine personnel, deliver equipment and supplies to the working levels, hoist ore from the working levels, and supply fresh air to, as well as exhaust used air from, the workings. When it is not possible to provide for all these functions simultaneously, which is most likely for very large mines, it is necessary to sink shafts exclusively for the purpose of supplying or exhausting air.

Once the air reaches the operations level it is routed through a series of entries, drifts, etc., to the location where it is needed, and from these it courses, now contaminated with gas, dust, and water vapor, and heated up, through a series of exhaust airways to the exhaust shaft, and finally back to the atmosphere. Along its travel through this circuit, the air encounters considerable resistance, in the form of friction against the walls of the openings through which it travels, because of constrictions in narrower passageways, and because of energy losses at each sharp bend. It is therefore usually necessary to force the air through the mine. This is accomplished by means of mine fans (Fig. 26), very large fans, some of which can deliver well over 500 m^3 of air per second and are driven by motors of over 5000 hp. Table V lists the most commonly used types of main fans. Examples of actual mine ventilation systems are summarized in Table VI. Such equipment needs to be designed carefully, based on the total air volume that needs to be delivered and on the total resistance (pressure drop) that must be overcome.

The equations governing airflow through an airway can be written as

$$H = KPLQ^2/A^3 = KPLV^2/A = RQ^2$$

where H is the pressure loss, K is a friction factor expressing the resistance of an airway to airflow, P is the perimeter of the airway, L is the length over which the pressure drop occurs, Q is the volumetric flow rate, A is



FIGURE 26 Two surface exhaust mine fans (Spendrup AMF 2000-80-10 vane axial fans) operating in parallel. The fans exhaust contaminated or used air from an underground mine through the exhaust drifts or shafts, thus establishing an airflow through the mine and pulling fresh air into the mine through the intake drifts or shafts. Specifications of the fans shown: Fan diameter is 79 in. (200 cm) internal diameter; blades are adjustable (at standstill) to allow operation in most efficient position; each fan delivers 225,000 cfm (106 m^3/s), for a total of 450,000 cfm (213 m^3/s); fans induce a pressure differential across the mine air intake and outlet of 5 in. WG (5 in. water gauge, 1.24 kPa); fan speed is 1180 rpm (revolutions per minute). [Photograph courtesy of Spendrup Fan Company, Division of Spendrup and Assoc., Inc., Grand Junction, CO.]

the cross-sectional area of the airway, V is the average linear air velocity, and R is the airway resistance.

The equation involves several simplifications, notably in the assumption of a square relation and of a constant friction factor. These are considered acceptable in light of the numerous system uncertainties and frequent system changes and uncertainties in the parameters, particularly K. The most reliable method for determining K is by measurement in the mine. Tables and graphs of extensive measurements have been published and are used for planning purposes.

TABLE V Mechanical Ventilation Equipment Used in Mines[a]

Fans	Centrifugal	Radical tip
		Forward curved
		Backward curved
	Axial flow	Tube–axial
		Vane–axial
Compressors	Centrifugal	
	Axial flow	
	Positive displacement	Rotary
		Axial flow
Injectors	Cylindrical	
	Conical	
	Venturi	

[a] Reproduced with permission from Y. J. Wang (1982). Mechanical ventilation equipment. Chapter 10, pp. 255–289, "Mine Ventilation and Air Conditioning," 2nd ed. (H. L. Hartman, ed.), John Wiley & Sons, New York.

While chronologically, over the lifetime of a mine, the selection of main fans is the first task of the ventilation engineer, many other responsibilities occur during operations. Because a mine continuously changes as rock is being removed, the ventilation circuits also change continuously. The primary responsibility of the ventilation engineer during production is to ensure that all underground environmental conditions, such as minimum air volume required at each working point, maximum gas or dust concentration, and maximum temperature, can be satisfied. This requires frequent, often continuous, monitoring, and frequent adjustments to the ventilation circuits. Such adjustments are made by a variety of control methods: opening or closing of various mine openings to airflow by means of stoppings or walls across the entries, guiding the air to the face with brattice curtains, installation of auxiliary fans, use of ventilation pipe or ducts, and so on. Control procedures commonly used in room and pillar coal mines are illustrated in Fig. 6.

Analysis and design of mine ventilation circuits has much in common with many other types of engineering network analyses, such as electrical power, gas, and water distribution systems. As a result of the nonlinear relation between pressure and flow rate, usually assumed quadratic, some forms of the laws governing circuit flow are fundamentally different from the usual expressions for linear circuits. Kirchoff's second law, for example, that the sum of the pressure drops around any closed loop must equal zero, is written as

$$\Sigma H_i = \Sigma R_i \mid Q_i \mid Q_i = 0$$

where H_i is the pressure loss along loop segment i, R_i is the resistance, and Q_i the flow rate.

While the equivalent resistance for airways in series remains simply the sum of the individual resistances, the equivalent resistance R_{eq} for a number of parallel airways is given by

$$1/(R_{eq})^{1/2} = \Sigma\left(1/R_i^{1/2}\right)$$

where R_i is the resistance of each airway.

By combining the resistances of all the airways in a mine, the engineer determines the mine resistance R_m, and the mine characteristic curve, the pressure differential H_m across the entire mine required to generate a total mine flow Q_m:

$$H_m = R_m Q_m^2$$

The intersection point of the mine characteristic curve and the fan characteristic determines the operational point of a particular fan across a particular mine configuration.

TABLE VI Ventilation Data from Metal Mines[a]

General				
Company	AMAX	Magma	UN-Homestake	Sunshine
Mine	Climax CO	San Manuel, AZ	Section 23	Sunshine, ID
Commodity	Molybdenum	Copper	Uranium	Silver
Mining method	Block caving	Block caving	Room/pillar	Cut/fill
Ore production (tpd)[b]	28,000	62,500	1068	800
Number shifts per day	3	3	2	2
Employees (underground/shift)	500	490	105	N.A.
Ventilation system				
General type	Combined	Blower	Exhaust	Exhaust
Total quantity,[c] cfm $\times 10^6$ (m^3/s)	1.25 (590)	1.2 (570)	0.7 (330)	0.13 (61)
Total mine head (pressure), in. water (kPa)	14.5 (3.61)	6.9 (1.7)	6.5 (1.6)	27.0 (6.72)
Fan data—main fans				
number and diameter (in.)	1×84	1×48	1×38	6×32, 54
	1×96	3×60	$7 \times 42\frac{1}{4}$	
	2×108	5×72	1×45	
	3×120		4×60	
Type	Axial	Axial	Axial	Axial
Total hp (kW)	4800 (3600)	3000 (2200)	1300 (1000)	1060 (800)
Fan data—other fans				
Number and size (in.)	32×48	30×34	N.A.	$4 \times 32\frac{1}{2}$
	6×54			
Rated hp	20, 60, 100, 150	20	—	60
Design data				
Minimum quantity in work place, cfm[c] (m^3/s)	Variable	7000 (3.3)	8500 (4.0)	5000 (2.4)
Minimum velocity in work place, fpm[d] (m/s)	230 (1.2)	50 (0.25)	140 (0.7)	N.A.
Air conditioning				
Plant	None	Cooling	Heating	Cooling
Temperatures in work place, °F (°C)				
Wet bulb	49 (10)	78 (26)	N.A.	80 (27)
Dry bulb	51 (11)	83 (28)	N.A.	82 (28)
Ventilation cost,				
$/ton or % direct cost	N.A.	N.A.	7%	$0.75[e]

[a] Reproduced with permission from T. J. O'Neil and B. R. Johnson (1982). Metal mine ventilation systems, Chapter 14, pp. 379–420, "Mine Ventilation and Air Conditioning," 2nd ed. (H. L. Hartman, ed.), John Wiley & Sons, Inc., New York.
[b] Short tons per day.
[c] cfm, Cubic feet per minute.
[d] Feet per minute.
[e] Excluding power.

Mine ventilation analysis, especially for large multilevel mines, is quite complex, involving several driving forces (e.g., fans, natural pressure differentials due to differences in elevation and in temperature, thermal draft induced by a fire) and several causes of energy or head losses (e.g., friction losses, resulting from friction along shaft and mine entry walls, and shock losses due to sudden changes in direction or in cross-sectional area, as in haulage drifts, where a train can fill a substantial fraction of the drift). The fluid of interest, air, is compressible and changes in composition as water is picked up or condenses out. Frequently air "losses" occur through fractured rock between the air-intake and the air-return courses (or even the surface). Finally, the mine ventilation circuit continually changes as mine development and ore production create new openings (while in some mining methods mined-out areas are backfilled or allowed to close by caving). The complexity of the system requires computer simulation for efficient and realistic analysis. A simplified schematic ventilation diagram prepared to illustrate a

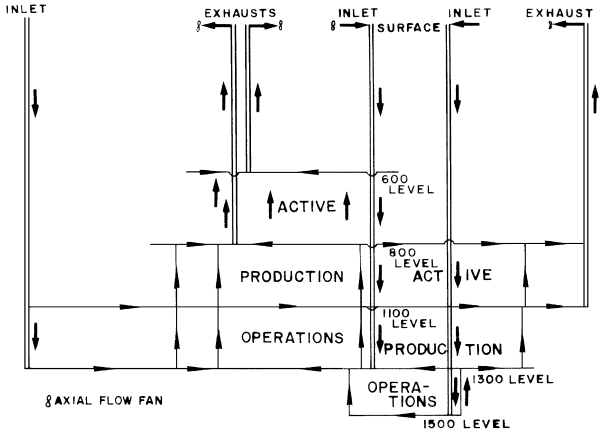


FIGURE 27 Simplified ventilation circuit for a mine with three exhaust shafts, three air intake shafts, and five working levels to which fresh air must be supplied and from which used air must be removed. Computer programs are used to analyze mine ventilation circuits, e.g., calculate airflow through each branch, pressure drop along each branch, temperatures at various points in the mine, and fan capacity requirements. [Reproduced with permission from Control Data Application Software MIVENDES—Mine Ventilation System, Control Data Corporation, Minneapolis, MN.]

computer analysis is shown in Fig. 27. Considerable engineering judgment is required for the determination of the necessary input variables for such an analysis. For analyses or simulations of operating mines, it is desirable to obtain an extensive data base of in-mine measurements.

F. Mine Surveying

Mine surveying has the same objectives as conventional or civil engineering surveying: identify various positions with an acceptable accuracy with respect to a certain reference frame, and provide the necessary data to allow preparation of maps and cross sections. In addition, the mine surveyor usually is responsible for staking out position and direction of future excavations, particularly important for example when a long haulage drift is started from both ends, and for measuring progress of the excavations.

While surface mine surveying has much in common with conventional surveying, and to a great extent uses the same methods and instruments, underground mine surveying remains quite different, because of a combination of factors that complicate underground surveying considerably. Common to all types of mining, of course, and the primary reason for the importance of mine surveying, is that by definition the geometry of a mine changes constantly, as rock is removed and as waste dumps are built.

Mine surveying, similarly to other land surveying, has seen major changes over the past few decades, and especially during the late 1990s. GPS (Global Positioning Systems) surveying has extremely rapidly become the norm for a variety of mining and mining-related applications. Locating precisely the position of exploration activities, whether physical geological mapping, geophysical surveys, geochemical sampling, exploratory drilling, or any other field activity, has become extraordinarily simple and inexpensive.

Similarly, in surface mining GPS application has become the norm. It allows for extremely fast, accurate (to the centimeter), and relatively simple determination of precise positions for many applications, e.g., blasthole locations, volume determinations for pit cuts, heap leach pads, and overburden disposal dams. One great benefit of GPS surveying is that all data can be collected in a format compatible with and directly transferable to the mine planning and the mine operations data bases. The most revolutionary development likely to result from the use of GPS applications is the development of autonomous equipment, i.e., equipment that will not be operated by people. Major surface mining equipment will be provided with GPS antennas, which will allow the equipment guidance system to position the equipment by itself. For hauling equipment such as trucks one of the main reservations slowing down implementation of such systems is safety concerns. For loading equipment such as shovels and excavators the positioning is complemented by the requirement for a robust manipulation control system, i.e., essentially a robotic mode of operation. Any such systems will have to be extremely rugged, given the exceedingly rough physical environment, in particular including frequent large dynamic load changes, within which they will have to operate with great reliability.

GPS information has proven particularly effective and efficient in large earth-moving applications. Dozer operators can be provided with a direct display at their operating console of the precise positions from which and to which they are moving dirt. Such operator assistance greatly simplifies the control requirements for the construction of major facilities such as waste embankments and heap leach pads, and is particularly helpful during the preparation of reclamation and restoration for recontouring prior to revegetation. Similarly, operators of excavation equipment such as shovels, front end loaders or hydraulic excavators can be provided with precise location of the ore/inert rock contacts in blasted muckpiles, without the need for manual surveying of such boundaries. Haulage trucks can be directed to their appropriate dumping destination based on the loader's digging position. A growing concern with the rapid increase of instrumentation of various types on the earth-moving equipment is the risk of operator information overload, and possible resulting confusion.

Laser scanning has become a most productive and efficient tool for numerous mining applications. A laser beam

scans the surrounding rock surfaces and provides a rapid and sufficiently accurate three-dimensional profile of rock surfaces. It is used for such applications as the measurement of mined-out volumes, the determination of volumes and geometry of waste dumps and heap leach pads, and the profile of underground excavations such as mine drifts and stopes. Laser profiling has considerable benefits: safety improvements, by reducing personnel exposure, increased measurement speed, and direct generation of digital record files. The combination of GPS surveying and laser scanning provides the technology for far more comprehensive surveys at reasonable cost, thus providing a more accurate definition of the continuously changing topography of the area affected by mining.

G. Equipment Selection

A modern mine is an extremely capital-intensive industrial operation, and a major fraction of the initial investment is the acquisition of the mining equipment. In addition, during the life of the mine, equipment replacements or additions usually will become necessary. Equipment selection therefore is an integral part of mining engineering.

Mine equipment selection is a complex process, continuing for nearly the entire life of a mine. It must be integrated with overall mine planning and design, and it must take into account numerous site-specific factors, such as climate, power supply, and labor supply, as well as rock and ore conditions, mine size, etc. The requirement that equipment selection must be integrated with mine planning is illustrated by Fig. 8: The feasibility of the planned mining sequence depends entirely on the availability of a dragline that can strip the coal seam to the proposed depth and spoil the overburden to the required distance and height. This figure also illustrates the next complication, the fact that several entirely different types of equipment often can be used to accomplish the same goal. In the case under consideration, it might be possible to strip with a large stripping shovel, with a bucket–wheel excavator (Fig. 16), or with a fleet of scrapers (Fig. 5). Although in many cases particular site conditions (e.g., rock hardness, depth, rainfall) will dictate or strongly favor a particular type of equipment, in many other cases only detailed analysis will provide the necessary information to allow an optimum selection.

Once mine design has been narrowed down to the fairly specific description of unit operations, several choices usually remain. At this point it becomes particularly important to match the equipment to the specific task it will be performing. One aspect of the procedure followed is illustrated for stripping and for loading equipment in Fig. 28 and in Tables VII and VIII. The figures and tables present general specifications of the equipment, describing the range of parameters over which it can perform, as well as the space requirements for the equipment. Such figures and tables are available from all manufacturers who supply

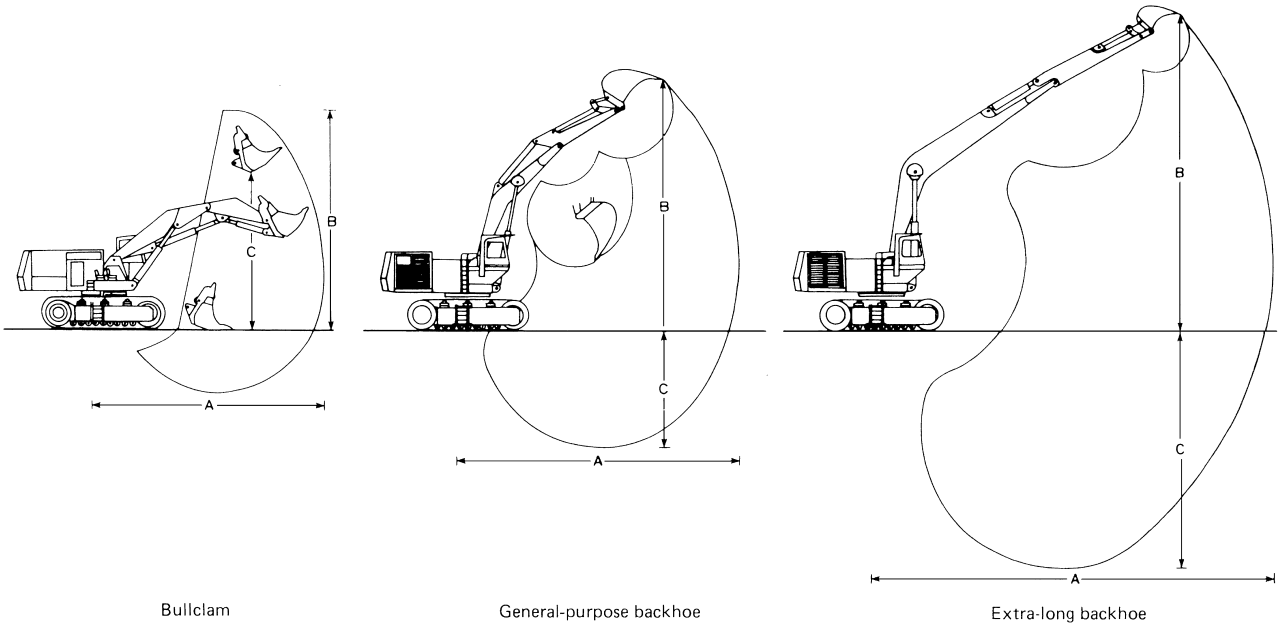


FIGURE 28 Working ranges for a large hydraulic excavator (DEMAG H241). Boom and bucket can be selected for the excavator to dig primarily at grade (left), or below grade (middle and right), over a longer range (right), but then at a reduced bucket load capacity compared to the shorter range (middle). Numerical values of the range parameters and bucket capacities are given in Tables VI and VII. [Figure courtesy of Mannesman Demag Baumaschinen, Düsseldorf, West Germany.]

TABLE VII Working-Range Options for a Large Hydraulic Excavator (DEMAG H241) (Fig. 28)[a]

	Bullclam	Standard backhoe	Long backhoe
Maximum digging radius, A, ft–in. (m)	50–6 (15.4)	60 (18.8)	78–9 (25)
Maximum digging height, B, ft–in. (m)	50–10 (15.5)	55–9 (17)	65–7 (21)
Maximum dumping height, C, ft–in. (m)	32–10 (10)		
Maximum digging depth, C, ft–in. (m)		26–3 (8)	47–7 (14.5)

[a] Reproduced with permission from Mannesmann Demag Baumaschinen, Dusseldorf, West Germany. Dimensions A, B, and C are illustrated in Fig. 28.

the type of equipment under consideration. In general, a range of individual equipment models will be able to perform the required task. Final selection usually is based on extensive computer simulations, matching various types of equipment (e.g., shovels, trucks) with each other as well as with the range of mining conditions expected.

VIII. MINE UTILITIES

An integral part of mining engineering is to provide the utility services required to maintain efficient functioning of the mine.

A. Mine Power

In most mines electrical power is used. Some underground mines still operate primarily on compressed air. In either case a comprehensive power distribution system is required for all equipment (e.g., drills, loaders, pumps, fans).

Compressed air usually is generated on the surface, in a main compressor station, and delivered through a network of pipes. Because of the extreme inefficiency of compressed-air systems, especially the delivery methods and the equipment powered by compressed air, it is probable that the use of compressed air in underground mining will continue to decrease.

TABLE VIII Bucket Capacity Options for a Large Hydraulic Excavator (DEMAG H241) (Fig. 28)[a]

	Material loaded unit weight, loose		Bullclam or standard backhoe bucket capacity	
Type	lb/cu. yd	kN/m^3	cu. yd	m^3
Heavy, e.g., ore	4700	28	12	9
Average e:g., overburden rock	3000	17.5	19	14
Light, e.g., coal	1900	11	28	21

[a] Reproduced with permission from Mannesmann Demag Baumaschinen, Düsseldorf, West Germany.

Surface mine electrical power distribution systems include a delivery system, typically a high-voltage transmission line from a utility where available, or power generated at the mine, a primary substation, one or more secondary substation levels, and a pole-mounted transmission line and cable power distribution system. A common trend both in underground and in surface mining is toward electrical power delivery to the mining equipment at higher voltages in order to increase transmission as well as operational efficiency.

Energy costs are a significant fraction of mine operating costs. The energy crisis of the 1970s, in addition to drastic cost increases and associated problems such as temporary fuel shortages, created serious uncertainties with regard to the choice of the most desirable types of energy to be used for mining. The strong renewed interest in electrically driven open-pit mine haulage systems is an example of the rethinking of mining methods resulting from energy consideration. Especially for large surface mines this has major mine operations selection implications, which remain far from resolved.

B. Mine Drainage

Most mines, especially the ones that penetrate deep into the Earth's crust, collect large volumes of water because they drain water from the surrounding surface and rock formations. In order to permit efficient mining it is necessary to keep a mine relatively dry, or at least to prevent flooding of the mine. Mine drainage requires the design and operation of a system for collecting the water entering the mine, for removing it, and for disposing of it.

Under extreme conditions, when it has not proved feasible economically or technically to prevent excessive water inflow, mines have been abandoned because of flooding. Even when conditions do not reach such an extreme, removing water from a mine can be a major operational cost. The first consideration of mine drainage requirements, therefore, must occur-during the mine feasibility evaluation. If any possibility exists that water problems are likely to be encountered, hydrological investigations should be included in the site investigation program.

Design of the mine drainage system will be based on estimated inflows, during the premining design phase, and is likely to require some adjustments once mining has started and actual inflows can be measured. The collection system usually will be designed to bring all water to one of the lower points in the mine, such as near a shaft bottom

for an underground mine, or on the pit floor for a surface mine. This can be accomplished by a series of ditches, excavated to allow gravity flow to the greatest extent possible, supplemented by auxiliary pumping through water pipelines as required. From the main collection point, near the main pumping station, the water will be pumped to the surface (Fig. 1). An important design decision is the determination of the size of the main collection basin or sump, which must be sufficiently large to prevent mine flooding during a power or main pump failure.

C. Mine Lighting

Mine lighting, or illumination, in surface mines that operate around the clock as well as in underground mines, has long been recognized as an essential factor in improving the working environment, thereby enhancing safety and productivity. The issue has largely been resolved in surface mining, through a combination of stationary light sources, usually pole-mounted powerful spot- and floodlights, together with similar light sources on all mobile equipment. Lighting for underground production operations remains a fairly difficult and somewhat controversial subject. Traditionally, fixed lights are installed in stationary and intensely used areas such as shaft loading stations, underground repair and maintenance shops, storage rooms, and pump and ventilation stations. Along the haulage and access drifts to the production sections, as well as in the production areas themselves, only minimal, if any, lighting is installed permanently. Light necessary for production is provided by equipment headlights, complemented by a few spotlights on the equipment and by the caplight worn on the hardhat by each miner. This results in a very limited illuminated area, and hence limited visibility, as well as in a highly variable light intensity.

Providing improved lighting and illumination has been the goal of considerable research and development. Achieving this goal is complicated by the need to avoid blinding miners, particularly when operating equipment, and to minimize eye stress and fatigue resulting from frequent transitions between dark and lighted areas. Further complications and potential safety hazards arise from the fact that the production areas continuously change as ore is mined. This necessitates frequent rerouting of electrical power supply systems, and moving the lights themselves. The addition of electrical circuitry in the limited and crowded operational space poses risks. Power lines, cables, and lights can be damaged easily by mining equipment or by even small rock falls.

D. Communications

Effective communications between the frequently widely separated locations of the various operations within a mine are an essential prerequisite for efficient production. In most surface mines all vehicles are equipped with two-way radios. In larger mines operations are directed from a central dispatch station. Most underground mines have extensive telephone networks, which are legally required in underground coal mines. In mines with electrical train haulage, the trolley wires that supply the electrical power to the locomotives commonly also are FM carriers for voice communications between loco drivers, dispatchers, and other points along the track. Attempts at more widespread use of radio communications in underground mines have not yet been entirely successful. Considerable research and development is in progress toward the establishment of wireless underground communication systems. It appears probable that such systems will find increasing commercial applications because of their many obvious benefits, such as rapid communications in case of accidents, fires, and equipment breakdowns requiring maintenance personnel. In addition, telemetering of various mine parameters (already widely used in surface mining), such as ventilation air flow, temperature, and equipment position and condition, will provide for improved control, engineering, and management of mine operations.

E. Maintenance

Maintenance, both preventive and remedial, is required to obtain maximum productivity from the large fleet of expensive equipment characteristic of most mining operations. Keeping all equipment running as continuously as possible is complicated by the severe operational and environmental conditions to which it is subjected. A typical open-pit truck, for example, loaded by a typical mining shovel (Fig. 11) is repeatedly subjected to an impact load of 50 tons or more, dropping from a height of 5–10 m. When fully loaded, the truck carries a payload of 100–200 tons of rock, in addition to its own weight, which is not much less than the payload, along an unpaved road, frequently up a steep grade (Fig. 3). In order to maximize production, speed will be the maximum compatible with safe operations. At the unloading point the rock load slides, rolls, and tumbles out of the truck body. Part icularly at loading and unloading points the environment is dusty, even with extensive water spraying. The temperature can be far below freezing, as for iron ore mines in Northern Minnesota, or can exceed 50°C, as in deep copper pits in Southern Arizona.

As the size of mining equipment has grown, each individual unit has become increasingly important in terms of total production, relative fractional production of the total mine output, and personnel productivity. Shutdowns of large high-production equipment carry a heavy penalty

in terms of lost production, in addition to direct repair or maintenance costs. This explains why maintenance has become an ever more important aspect of modern mining. The trend toward growing responsibilities and more sophistication for maintenance engineering is likely to continue, if not accelerate, as a result of the introduction of more automation, remote control, electronics and computerized operations and instrumentation. For most large surface mining equipment, the maintenance cost already exceeds operational, fuel, and capital depreciation costs combined. The growing importance of maintenance is reflected in an organizational structure where maintenance departments in modern mines report directly to a much higher management level than was the case until fairly recently. Similarly, maintenance includes a relatively larger fraction of total mine personnel and typically has significantly better trained personnel.

Maintenance consists of preventive and of breakdown maintenance, although these categories usually are further subdivided. The prime objective of preventive maintenance is to minimize the need for breakdown maintenance, the unscheduled repair of equipment that has become inoperative. Production losses are particularly severe when primary equipment such as loading shovels breaks down, because this results in at least a temporary shutdown of units such as haulage trucks that depend on the primary ones. The objective of preventive maintenance is to minimize breakdowns by systematic inspection and replacment of part prior to failure. This is a complicated and major task, considering that a large modern mine, whether surface or underground, have a fleet of more than 100 primary mining units and easily an equal number of support vehicles. Moreover, the equipment is quite varied. In an open pit it will include drills, shovels, front-end loaders, bulldozers, haulage trucks or trains, water trucks, personnel vans, road graders, mobile cranes, forklifts, small trucks, and automobiles. The variety of tasks involved in optimizing the availability of such a fleet is apparent. Planning, scheduling, and record-keeping are integral parts of maintenance operations and are largely computerized. This is especially important for preventive maintenance, which to a significant extent is based on frequency-of-repair histories.

On larger modern equipment extensive on-board monitoring is performed, including tire pressure and temperature, vibration frequencies and amplitudes of rotating engine components, temperatures and pressures in bearings, electric motor amperage and power consumption, etc. The data often can be transmitted remotely to a maintenance engineering tracking facility, where equipment's vital signs are monitored continuously or checked periodically in order to assess its operational condition. Alternatively, or complementary, the data may be displayed, continuously or on request, in the operator's cabin. These continuous or intermittent monitoring systems lend themselves very readily to a connection to alarm systems that inform the operator, maintenance personnel, and/or supervisory personnel of conditions that require immediate or early attention.

The extensive data bases that are being collected digitally provide an ideal tool for data mining, a systematic search of the accumulated data in order to detect patterns and trends about historical equipment performance. While past patterns may not necessarily predict future behavior and performance, they can be most helpful in identifying potential problem areas and sources.

Maintenance organizations include central facilities, the major electrical, mechanical, and machine shops, as well as decentralized specialized facilities, the latter including field repair trucks, inspection trucks, and lubrication trucks. For underground mines, all maintenance but major overhaul usually is performed underground. Most mobile surface equipment, except for very slowly traveling units such as shovels and drills, is maintained at the central shops. The combination of well-planned preventive maintenance with well-organized fast-response breakdown maintenance makes it possible to obtain an equipment availability of 80–90% for most surface mining equipment. Availability of underground mining equipment has not yet reached these levels.

An important aspect of mine maintenance, usually assigned to the mine operations rather than to the mine maintenance department, is maintenance of haulage roads. Smooth roads (or railroad track) allow higher safe speeds and reduce equipment wear and tear. The latter aspect is particularly important for equipment running on rubber tires (e.g., trucks, load–haul–dump vehicles, front-end loaders, shuttle cars), for which tire costs are by far the largest operating cost.

IX. HEALTH AND SAFETY ENGINEERING

Minning has traditionally been a high-risk occupation. It is part of the responsibility of the mining engineer to ensure that all design and operations include health and safety considerations as a prime objective. Stringent and detailed regulations, in the United States both federal and state, govern virtually all aspects of mining.

In recent decades this situation has been reversed rather dramatically, especially in developed industrial societies, to the point where mining ranks about average in terms of accident risks. In part at least this favorable trend is the result of reduced personnel exposure to hazards. As equipment has grown larger and more productive, the number of people exposed to hazards has diminished greatly. In part improved health and safety improvements are the result of

increasingly stringent regulations and mine inspections. In recent years the reduced exposure risk has decreased further as a result of the growing use of remote-controlled mining equipment. Particularly in underground mining operations the equipment operator can remain under well-controlled ground, or even on the surface, while mining equipment is operated remotely, e.g., to load and haul rock.

Health and safety remain major concerns in less developed countries, particularly in small artisanal mines, mines where typically little or no engineering is applied. International and various domestic government agencies and nongovernmental organizations are trying to address health and safety problems in such small mines. The task is daunting.

X. ENVIRONMENTAL ENGINEERING AND RECLAMATION

Mining removes mineral resources from their natural environment, and hence by definition disturbs nature. It is the function of enviromental engineering to ensure that the impact of the mining operations on the natural environment—on land, on air, and on water, or, more generally, on the biosphere—is minimized or is maintained within acceptable limits. Mining, surface mining in particular, requires a change in land use, at least for the life of the mine. Reclamation restores land use as rapidly as possible, during or after mining, or returns the land to an environmentally acceptable condition upon completion of mining.

Effective reclamation requires that the postmining condition of the land be taken into account during mine planning, that is, that the mining operations be planned and designed with the ultimate land restoration and use as an integral part of the planning objectives. (This consideration helps explain the difficulties encountered by mines when stringent reclamation requirements are imposed or changed while they are already operating, i.e., when mining methods and plans already have been selected, mining equipment has been acquired, etc.)

The environmental impact of mining can be subdivided into the impact on land, on water, and on air. The primary air pollution resulting from mining is the generation of dust, that is, particulate matter that becomes airborne. (Sulfur emissions from coal-burning power plants or metal-ore smelters are more directly associated with the user, or customer, of the mine, than with mining itself.) The federal government and most states have enacted legislation that requires the mine operator to maintain fugitive dust within acceptable limits. Typical control procedures include extensive water spraying of haul roads, by means of special large tank trucks, of transfer (loading, unloading) points, and of storage areas. Wetting agents often are added to the water in order to improve retention efficiency.

Land reclamation is a particular concern for surface coal mining, especially strip mining which tends to disturb relatively large ares in a relatively short time. (The abandonment of unreclaimed surface coal mines, especially in Appalachia, is at least partially responsible for the enactment of strict reclamation legislation. Some of the more destructive past practices, such as conventional contour mining, have now been replaced entirely by mining methods, such as haulback mining, that allow full reclamation of the mined areas.)

Reclamation practices of necessity must differ significantly depending on local soil, topographical, and climatic conditions (e.g., revegetation in the semiarid Southwest imposes significantly different seeding, watering, feeding, and contouring requirements than in the Midwest, where precipitation usually is ample). The typical first step in reclamation is the removal and temporary storage of the top-soil, prior to mining (Figs. 4 and 5). Waste rock removed during mining is spoiled in mined-out areas. Upon completion of grading or surface recontouring, the top-soil is replaced. The restoration of the land is completed with revegetation. The latter phase usually is accomplished with the assistance, or under the direction, of agronomists and plant scientists. The major environmental impacts of underground mining include subsidence of the ground surface as a result of removing rock from the subsurface and changes in the groundwater, primarily the water table level drawdown caused by pumping water from the mine. Acidic mine drainage can result where the water is released from operating or abandoned coal or metal mines in which the rock contains sulfur. Remedial action includes water treatment, mine design to minimize releases, and mine closure and abandonment procedures engineered to eliminate long-term impacts.

Reducing environmental impact has become a major preoccupation for the mining engineer. It affects all phases of engineering, from premining impact evaluations and extensive permit applications, through mine closure. Environmental protection is a major area of research and development, both by the industry and by the federal government.

XI. THE FUTURE

Mining, the basis and source of the industrial revolution, has been eminently successful in providing the material needs that have expanded greatly the material well being of modern societies. In 1998 the per capita U.S. consumption of mineral raw materials exceeded 47,000 lbs. The minerals industries have succeeded in providing for the increased minerals demand of developed societies at an

ever-decreasing price, economical, social, i.e., in terms of health and safety of the workforce, and environmental, i.e., with ever-decreasing impact on the natural environment within which mining, of necessity, must operate. To a large extent these improvements have resulted from improvements driven by mining engineering, as well as by societal pressures.

At the front end, improvements in exploration, driven by an enhanced understanding of geology and in particular by ever more sophisticated geophysical, geochemical, and geological field techniques, have resulted in more than simple reserve replacements for many decades: For virtually all minerals the newly discovered reserves have exceeded, usually greatly, the reserves that have been mined. For the foreseeable future, the mineral reserve base of the world is likely to continue to expand. Exploration improvements are only one of the reasons. At least as important are reduced production costs, in the mining and the processing cycles, which makes it feasible to extract economically minerals from deposits that even only a few years earlier would have been considered uneconomical.

The major challenge for the industry is whether or not it will be able to provide for the mineral needs of the large majority of the human population for whom so far the material basis for a reasonable standard of living remains elusive. Future mine development is challenged by major obstacles, the dominant one being ever more stringent environmental regulations and constraints.

Over the past few decades mineral production has become subject to a broad and deep set of environmental regulations. The trend toward ever more stringent environmental constraints is likely to continue. One result is that over the past few decades mine planning and design has needed to integrate a variety of disciplines that address environmental issues. Mine site investigations, design, planning, operations and closure now routinely include hydrological, geochemical, paleontological, archeological, wildlife, and atmospheric, as well as socioeconomical and political, studies. The challenge for the mining engineer is to integrate the priorities of the numerous parties involved into a workable productive mine plan. More broadly, like all engineering today, mining engineering needs to account for ecological constraints, and for the concomitant political and social constraints. Moving to the forefront in this regard is the requirement to develop a sustainable industry, however elusive a working definition of that term may be for now.

To what extent the increased costs imposed by society can be offset by enhanced technology will determine the future costs of mineral raw materials. Technological developments for the near future appear fairly clear. Equipment is likely to become ever more automated. GPS (Global Positioning System) equipment location combined with robotic operation of the equipment is likely to lead to autonomous operation of drills, shovels, hydraulic excavators, and haulage trucks. Remote control of underground mining equipment, e.g., longwall shearers and support systems, continuous miners, roadheaders and LHDs (load–haul–dump vehicles) will be implemented on a wider basis. The labor force will continue its transition from production to ever more technologically advanced control, operational, and maintenance functions.

Notwithstanding the occasional claims that advanced societies are becoming physically lighter, e.g., through the use of microelectronics and nanotechnologies, the statistics clearly indicate that over the past several decades the mineral raw materials demands are increasing, especially in the most material-intensive demands, housing and transportation: sizes of houses, urban sprawl, weights of automobiles, and associated road demands increases. Notwithstanding considerable improvements in energy efficiency, notably in transportation, the overall demand for energy continues to rise, as reflected for example in the steady if not spectacular increase in coal production for several decades.

In advanced societies access to land for mineral development is likely to continue to become more difficult. In several advanced mineral producers, notably the United States, Canada, and Australia, an increased sensivity to native lands rights reduces land access. So does the growing practice of removing exceedingly large remote areas from availability for mining. Nearer to urban areas, where most construction materials are mined, public sensitivity and opposition to mineral production is not likely to abate in the foreseeable future, if ever.

Political instability is likely to constrain mineral development in many less developed societies, especially in Africa. During the 1990s the political climate has been quite favorable for mineral development in most of Central and South America, as reflected in major increases in exploration and in mine development capital investment. Although the future remains less clear for several major Asian countries, the combination of population pressures and economic development demands suggests that further large-scale mine development is likely to continue.

Predictions of future production of major minerals are somewhat uncertain. Over relatively short time frames, e.g., up to 20 years, predictions for major commodities probably are reasonably reliable, although externalities always could cause some fluctuations, possibly sharp ones for some materials. The materials mined to the greatest extent, i.e., with the largest volume and value, are used in construction: aggregate, sand, gravel, limestone for cement. Although somewhat subject to the economic cycle, demand is fairly steady, showing a modest increase

over time, with population and wealth growth. Coal, the second largest mined product in terms of value and tonnage has shown a fairly steady U.S. and worldwide growth throughout the 1980s and 1990s. Because of increased environmental regulations and increased competition from natural gas for the generation of electric power, by far the largest use of domestic coal, it is unlikely that this growth will continue, at least in the United States. Some of the largest overseas coal producers, notably China and India, are investing heavily in coal-fired electric power plants and large ultramodern coal mines. These investments, in response to severe energy shortages, suggest increased coal production in these countries for the next few decades at least. The third major mining component, the production of metals, tends to be far more sensitive to economic cycles, e.g., copper to the construction of buildings, power lines, manufacturing, etc., or to extraneous factors, e.g., gold to the gold price fluctuations resulting from central bank sales. If the less developed countries, particularly the very large Asian societies, manage to maintain even moderate growth rates toward a living standard corresponding to the mineral consumption of the developed societies, the demand for base and ferrous metals should show a fairly steady growth. For some specialty precious metals, in particular platinum and palladium, the demand is likely to continue to rise significantly in light of their use in pollution control devices such as automobile catalytic converters. The future of gold depends largely on the whims of central bankers, highly unpredictable certainly to a nonfinance specialist. Both gold and silver are quite sensitive to the demand in Asia, and hence to the economic development on that continent.

An interesting topic that has received comparatively little attention, at least in the professional mining literature, is that of small-scale mining. It therefore may deserve pointing out that small-scale mining is a significant component, and in many less developed countries the major component, of the mining industry. Particularly with regard to employment, small-scale mining is a major factor, estimated to employ some 13 million people wordwide, and supporting economically some 80 to 100 million people. Although a modest contributor to worldwide mineral production, it is a noticeable factor in the production of gemstones and of precious metals. Small-scale mining allows a significant number of people to earn some income, but frequently at a steep cost. It tends to be markedly less safe than large-scale industrial mining, and its environmental impact is far more deleterious and far more difficult to bring under control. Given socioeconomic imperatives, there is every reason to believe that small-scale mining will continue to grow, at modest rates, for the foreseeable future, especially in Africa, Latin America, and Asia. In many of the countries with small-scale mining, export of the mineral production is a major contributor to the national economy and the balance of payments.

SEE ALSO THE FOLLOWING ARTICLES

EXPLORATION GEOPHYSICS • GEOENVIRONMENTAL ENGINEERING • GEOSTATISTICS • MINERAL PROCESSING • ORE PETROLOGY • ROCK MECHANICS • SOIL MECHANICS

BIBLIOGRAPHY

American Geological Institute (1997). "Dictionary of Mining, Mineral, and Related Terms," American Geological Institute, Alexandria, VA.

Craddock, P. T. (1995). "Early Metal Mining and Production," Smithsonian Institution Press, Washington, D.C.

Dutton, A. H. (compiler) (1993). "Handbook on Quarrying," 5th ed., Department of Mines and Energy, South Australia.

Gertsch, R. E., and Bullock, R. L. (eds.) (1998). "Techniques in Underground Mining," Society for Mining, Metallurgy, and Exploration, Inc., Littleton, CO.

Ghose, A. K. (ed.) (1994). "Small-Scale Mining: A Global Overview," A. A. Balkema, Rotterdam.

Hartman, H. L. (senior editor) (1992). "SME Mining Engineering Handbook," Society for Mining, Metallurgy, and Exploration, Inc., Littleton, CO.

Hoek, E., Kaiser, P. K., and Bawden, W. F. (1995). "Support of Underground Excavations in Hard Rock," A. A. Balkema, Rotterdam, Brookfield, VT.

Hustrulid, W., and Kuchta, M. (1995). "Open Pit Mine Planning and Design," A. A. Balkema, Rotterdam, Brookfield, VT.

Kennedy, B. A. (1990). "Surface Mining," 2nd ed., Society for Mining, Metallurgy, and Exploration, Inc., Littleton, CO.

Marcus, J. J. (ed.) (1997). "Mining Environmental Handbook," Imperial College Press, London.

The Northern Miner, Whiteway, P. (ed.) (1990). "Mining Explained," 6th ed. Northern Miner Press, Toronto.

Paşamehmetoğlu, A. G., and Özgenoğlu, A. (eds.) (1998) "Environmental Issues and Waste Management in Energy and Mineral Production," *Proceedings of the Fifth International Symposium on Environmental Issues and Waste Management in Energy and Mineral Production—SWEMP '98*, Ankara, Turkey, 18–20 May 1998, A. A. Balkema, Rotterdam, Brookfield, VT.

Ripley, E. A., Redmann, R. E., and Crowder, A. A. (1996). "Environmental Effects of Mining," St. Lucie Press, Delray Beach, FL.

Runge, I. C. (1998). "Mining Economics and Strategy," Society for Mining, Metallurgy, and Exploration, Inc., Littleton, CO.

Stack, B. (1995). "Encyclopaedia of Tunnelling, Mining and Drilling Equipment," Muden Publishing Co., Hobart, Tasmania.



Naval Architecture

Robert Latorre
University of New Orleans

I. Introduction
II. Historical Development
III. Problems in Ship Design
IV. Design Spiral

GLOSSARY

Block coefficient Coefficient of hull fullness,

$$C_b = \frac{\text{Volume underwater}}{\text{Length} \times \text{Width} \times \text{Draft}}.$$

Center of buoyancy Location of center of underwater volume.
Center of gravity Location of center of vessel weight.
Dynamic loads Transient loads on hull structure due to motion of hull in seaway.
Metacenter The point of intersection of the heeled ship centerline with a vertical line through the heeled ship center of buoyancy.
Seakeeping Denotes the seaworthiness of a vessel in waves.
Thermal loads Loads on the structure due to temperature differences.
Transverse center of gravity Transverse center of the vessel's center of underwater volume.
Trocoidal wave Standard wave that is somewhat sharp in the crest and flat in the trough.
Vertical center of gravity Vertical location of the center of weight.

THE TERM NAVAL ARCHITECTURE refers to the design of ships and other floating vessels. A naval architect is an engineer who designs floating vessels such as surface ships, submarines, and structures for offshore oil exploration. In each case, the vessel design must satisfy requirements that are often conflicting and that require exercising the art of compromise. The naval architect's goal of a balanced design requires providing in the floating vessel adequate space for the payload, propulsion machinery, and fuel, as well as a comfortable living environment for the crew, within budget limitations.

I. INTRODUCTION

A. Evolution of Modern River, Ocean, and Passenger Vessels

It is impossible to discuss naval architecture without mentioning ships and other hing vessels. The nav al architect designs a ship or floating vessel, which when built will perform a desired task safely and efficiently. In the early days this was fairly straightforward. The general design of sailing ships and riverboats (Fig. 1a) continued up to 1870 when sail power (Fig. 1c) was gradually

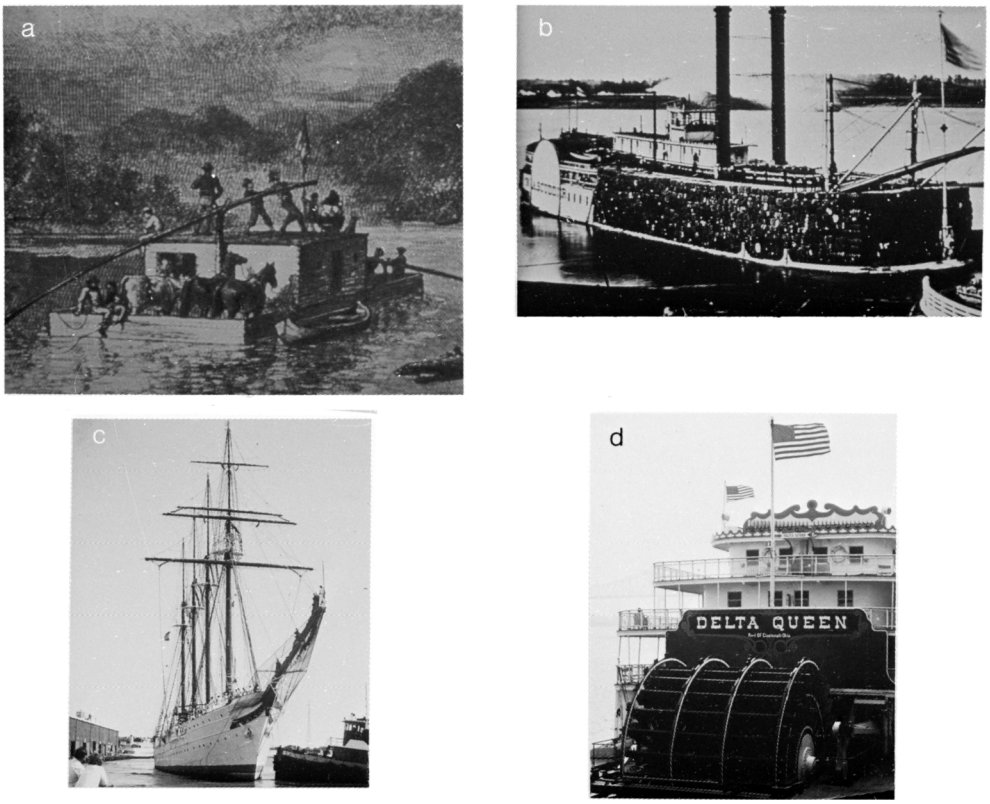


FIGURE 1 Early river and ocean vessels. [Courtesy of D. Lewane.] (a) River pole boat, 1800s. (b) Mississippi steam boat loaded with cotton, 1870s. (c) Sailing ship, 1800s to present. (d) Stern paddlewheel, 1800s to present.

augmented by a steam engine, driving either a screw propeller (Fig. 2) or a paddlewheel (Fig. 1d). With time, the vessel requirements evolved and distinctly different ship designs appeared on the rivers and seas. Beginning in the 1920s, paddlewheel steamers were gradually replaced on the rivers by diesel towboats with screw propellers recessed in the stern (Fig. 3d). Standardized locks and river-barge size resulted in large river cargo movements. Now barge tows five long by three wide being pushed by 5000–6000 hp twin-screw towboats are a common sight on the waterways.

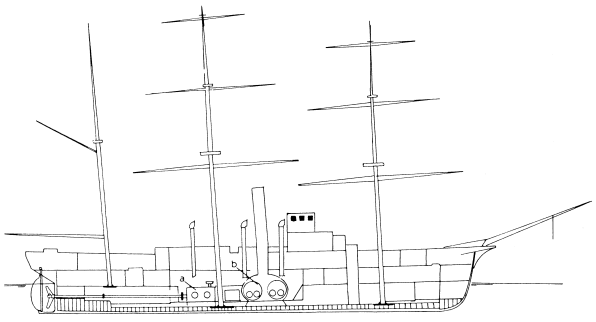


FIGURE 2 Combination sail- and screw-propelled ship, 1870s. (a) Horizontal compound steam engine. (b) Boilers. Length, 176 ft; beam, 32 ft; draft, 15 ft; displacement, 1025 tons; sail, 10,075 ft^2.

The international movement of oil and other bulk products such as grain, ore, and coal, resulted in the introduction of large ocean-going bulk ships and specialized terminals, as shown in Figure 4. In contrast, the movement of people over the water has involved a variety of vessel types. Figure 5 shows a sample of three passenger-carrying vessels. The crewboat in Figure 5a taxies the crew to and from the offshore oilrigs in the Gulf of Mexico. The crewboat is in direct competition with the helicopter, so it is designed using a high-speed planing hull and is fabricated out of lightweight aluminum. The paddlewheel excursion boat in Fig. 5b is designed to capture in an afternoon the period of Mark Twain with its Mississippi steamboat lore. The cruise ship in Fig. 5c is based on the modern concept of enjoying both the vacation trip and the vacation destination. It is designed as a floating hotel, with provisions for continued activity throughout the 1–2 week voyage.

B. The 1960s Revolution in Cargo Transfer

The most startling changes in ship design have been in the handling of boxed cargoes such as canned goods, electronic equipment, and machinery. Up to the 1960s, these cargoes were stored in the ship holds and hoisted

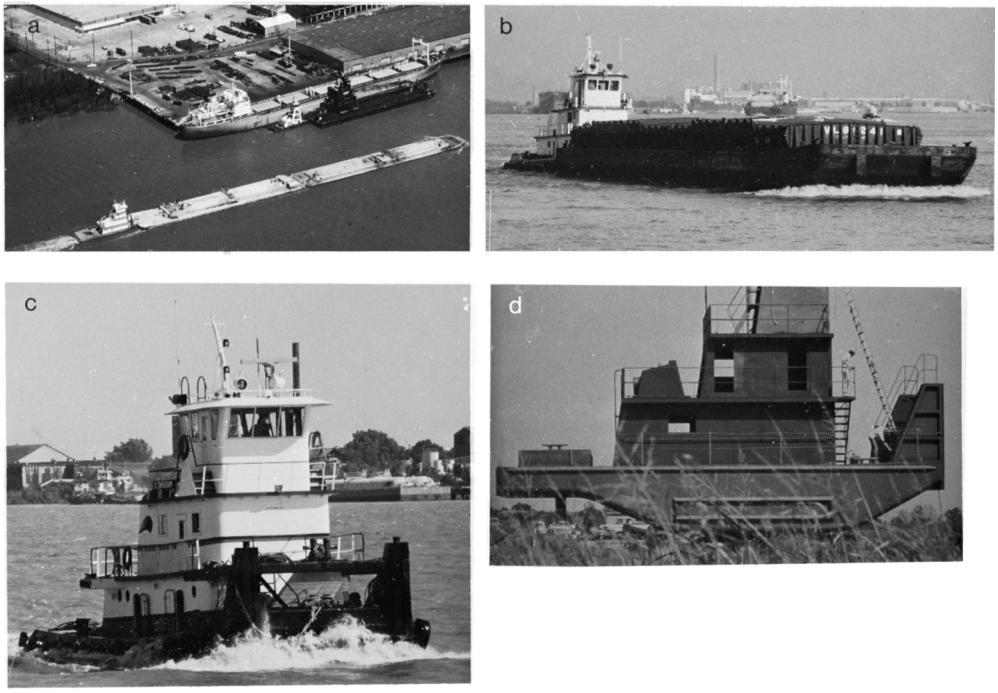


FIGURE 3 Modern river towboats pushing barges. [Courtesy of D. Lewane.] (a) Towboat pushing chemical barges. (b) Towboat pushing hopper barge. (c) Towboat. (d) Towboat under construction. Note recess in stern for propellers.

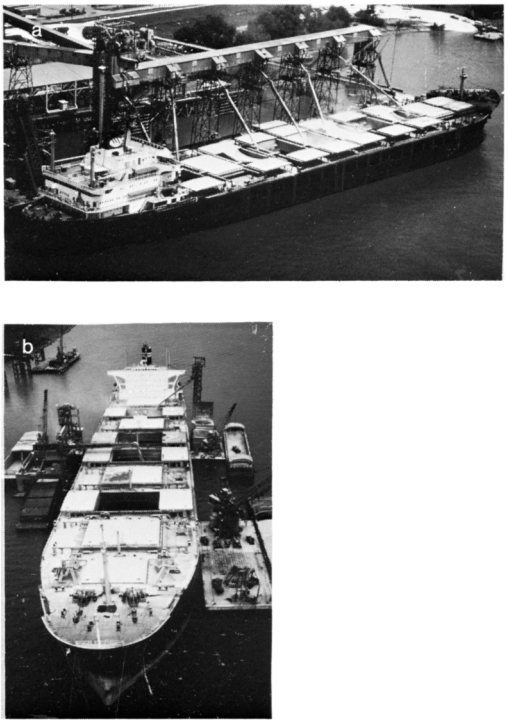


FIGURE 4 Loading grain into large ocean-going bulk carriers. [Courtesy of D. Lewane.] (a) Riverside loading facility. (b) Off-loading barges using cranes and conveyor.

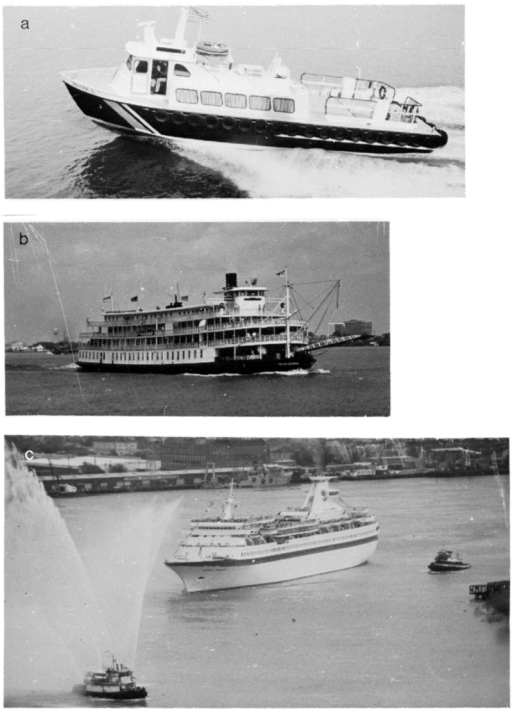


FIGURE 5 Passenger vessels. [Courtesy of D. Lewane.] (a) Aluminum crewboat for offshore oil rigs (high-speed planing hull). (b) Stern paddlewheel excursion boat (low-speed hull). (c) Cruise ship (high-speed displacement hull).



FIGURE 6 Cargo handling ship systems. [Courtesy of D. Lewane.] (a) Break-bulk ship unloading onto wharf. (b) Container ship being unloaded by container crane. (c) Roll-on/roll-off ship unloading via bow ramp. (d) SEABEE ship loading two 1000-ton barges on stern elevator. (e) LASH ship receiving 500-ton barge under stern crane. (f) LASH ship lifting 500-ton barge.

onto the wharf by the ship's cranes (Fig. 6a). These cargoes were later packed into boxcars or truck trailers and then dispatched to their final destination. In the 1960s, a number of businessmen and naval architects studied this cargo movement and concluded that the wharf activity was a "bottle neck" that lengthened the delivery time and exposed the cargo to damage and pilferage. At the same time, a number of ship operators developed the successful practice of shipping locked trailers. After the voyage, these trailers would be confined to a truck and delivered in what became known as a "door-to-door" operation. The success of this operation spurred the development of a number of systems and ship designs shown in Figs. 6 and 7.

1. Conventional break bulk (Fig. 6a). The cargo is shipped on pallets and stored in a warehouse for rail or truck shipment.
2. Container ship (Fig. 6b). The cargo is loaded onto metal containers, which are sealed and transported to the containers, which are sealed and transported to the container yard for loading onto the container ship.
3. Ro-Ro ship (Fig. 6c and Figs. 7a,b,c). When the roll-on-roll-off ship docks, its ramp is lowered and the trailers and other vehicles are rolled off and on. During the voyage they are tied down to the ship deck and the ramps are secured.
4. SEABEE (Fig. 6d). Cargo is loaded into 1000-ton barges, which float to the SEABEE ship stern. Two barges are then lifted by a submerged elevator to the proper level and moved forward where they are secured until unloading. The SEABEE carries 26 barges in three decks.
5. LASH (Figs. 6e and 6f). The cargo is loaded onto 500-ton barges, which are floated to the LASH-ship stern. The crane frame locks into fittings on the barge deck and

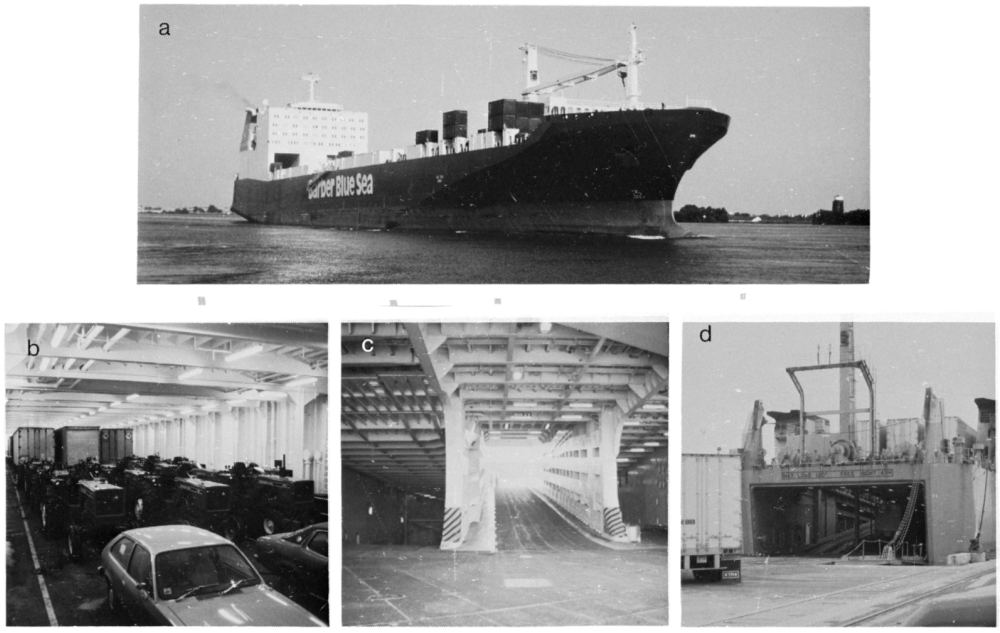


FIGURE 7 Roll-on/roll-off ship. [Courtesy of D. Lewane.] (a) Roll-on/roll-off ship entering port. (b) Wheeled cargo tied down to deck. (c) Ramps set for unloading operation. (d) Stern ramps lowered and trailers being towed off.

lifts the barges to the appropriate hold. The LASH carries up to 83 barges.

C. The 1990s Emergence of High Speed Craft

The search for higher speed in commercial and military activities has lead to the development of a number of different vessel types, supported by hydrostatic forces (displacement hulls) as well as hydrodynamic forces and aerodynamic lift. These vessels and their corresponding speed ranges are illustrated in Fig. 8.

Worldwide research and development is focused on high-speed trans-ocean commercial and military transport crafts. With the capability of carrying 5000 tons of cargo for 5000 miles at 40–70 knots these high-speed craft will create another cargo-transport revolution similar to the container ship in the 1960s and 1970s (Kennel 1997).

The high speed catamaran has grown in both size and speed. In 1998 the InCat catamarans Catalonia and Cat-Link V, built in Australia, crossed the Atlantic (2972 nautical miles—New York to Tarifa) at record speeds of 38.85 and 41.28 knots (Clifford 1998). These speed records were set by a 91.3 meter long aluminum catamaran hull, powered by quad 1000 hp waterjet units. The vessels have 800 passengers, 225 cars or a combination of cars, and four large buses (Clifford 1998). A typical 25 knot container ship takes 18 days to sail from Norfolk, Virginia to Santos, Brazil. With a high-speed 40 knot craft, the voyage time would drop to 10–11 days.

This performance breakthrough is advanced by the naval architect following the technology cross (Basin, Latorre 1997):

1. Use of strong lightweight materials.

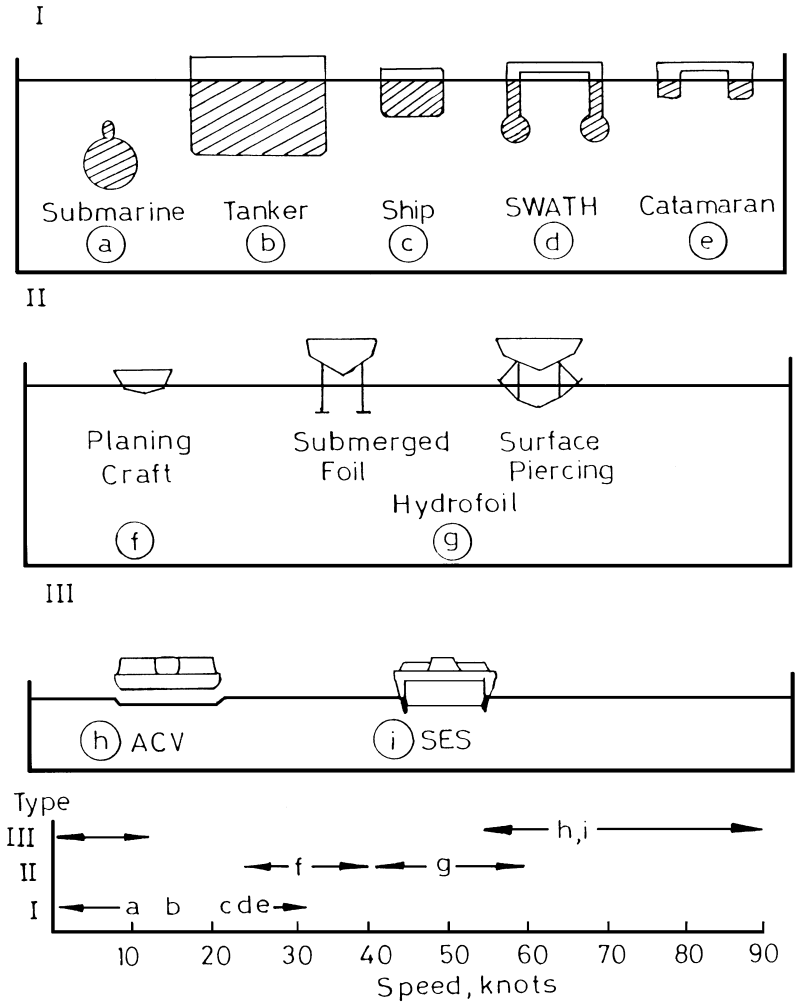


FIGURE 8 Vessel types based on support mechanism and relative speeds. (I) Displacement (hydrostatic) vessels. SWATH = small waterplane area twin hull. (II) Hydrodynamic-lift-supported vessels. (III) Aerodynamic-lift-supported vessels. ACV = air cushion vehicle. SES = surface effect ship.

2. Rational estimation of loads and design of an optimum structure.
3. Adoption of low drag, seakeeping and maneuvable form.
4. Use of efficient propulsion/maneuvering systems.

This represents a performance breakthrough that can be visualized by comparing the state-of-the-art in 1970 and 1995 high-speed craft plotted on the basis of:

Power coefficient:

$$Q = \frac{0.148 \times BHP}{\Delta V_s} \tag{1}$$

Speed coefficient:

$$K = \frac{0.538 \times V_s}{\Delta^{1/6}}. \tag{2}$$

The performance line of 1970 has been broken by the current designs, as shown in Fig. 9. The lower 1995 performance line shows a significant improvement in both power and the speed coefficient.

With so many alternatives, it is necessary to develop a measure of merit to select the "best" design. In some cases it is easy to consider the cost of each design and select the minimum cost design. In cases where the vehicles are different, such as helicopters and high-speed boats, a more general index has been used. One example is the adoption of the transport efficiency TE:

$$\text{TE} = \frac{(\text{total vessel weight}) \times \text{maximum cruise speed}}{\text{total installed power}} = \frac{WV}{P}. \tag{3}$$

This is equavalent to the transport factor, TF, introduced by Kennel (1998). Gabrielli and von Karmin developed an extended expression for the upper limit of vessel performance ULVP:

$$\text{ULVP} = \frac{\text{WV}^2}{P_{\max}} = 8800 \text{ ft/sec} = 2680 \text{ m/sec} = 5200 \text{ knots}. \tag{4}$$

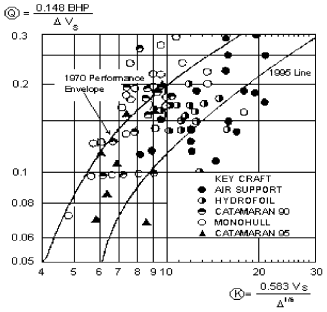


FIGURE 9 Examples of offshore oil drilling rig.

TABLE I Ranking of High-Speed Craft Based on Lowest Normalized Fuel Cost [Eq. (3)] at 0.2 Payload to Weight Ratio ($W_p = 0.2$)[a]

Rank	Craft
1	Ferry boat
2	28-Knot SWATH ship
3	Passenger jet
4	Airship
5	Catamaran vessel
6	Jet surveillance plane
7	Sidewall hovercraft
8	Diesel patrol boat
9	Hydrofoil boat
10	Planing hull boat
11	Gas turbine patrol boat
12	Helicopter

[a] Adapted from Crew's plot; see Clayton, B. R., and Bishop, R. E. D. (1983). "Mechanics of Marine Vehicles." Gulf Publishing, Houston.

Clayton and Bishop (1982) discuss this limit in their book. The limit is approached at low speeds by ships and at high speeds by aircraft.

Increasing energy costs have focused on the energy consumption of each vehicle referenced to the baseline of a typical marine diesel engine. Using this basis of the normalized fuel cost (NFC) developed by Crewe:

$$\text{NFC} = \frac{\text{fuel consumption relative to diesel}}{\text{maximum cruise speed} \times \text{payload}}. \tag{5}$$

Table I illustrates the relative NFC ranking of the vessels in Figure 8 for the range of payload to weight of 0.1–0.2. In the case where the requirement for speed is not critical, the NFC indicates the conventional ship hull has the lowest values. In contrast, the gas turbine patrol vessels and helicopter have the higher NFC values. This ranking helps explain the economic success of the planing hull crewboat (Fig. 5a) relative to helicopter in crew exchange from shore to the offshore rigs.

II. HISTORICAL DEVELOPMENT

A. Colonial Period

From the earliest records to the present, ships and other floating vessels have continued to be major components in international commerce and waging war. While archaeological explorations of sunken ships help us picture these early vessels, there is little documentation of the practice of naval architecture, or ship building, as it was often called in early records. This was mainly because the practice of naval architecture was treated as a secret and passed down as an apprenticeship to only a few in each country. This

resulted in the evolution of ship designs well matched to the local navigation conditions and available building materials. The exchange of technical information was also limited until the publication, in the 1750s, of books dealing with scientific and practical aspects of naval architecture. These early books focused on:

1. Science of ship design: These books were concerned with the estimation of the forces acting on the vessel hull, the equilibrium draft, and trim, as well as the ability of the vessel to float upright. This has evolved into the present-day areas of hydrostatics, hydrodynamics, resistance, propulsion, ship strength, seakeeping, and maneuvering. The book by P. Bouget, *Traite du Navire*, published in 1746, is often cited as laying the foundation for the scientific approach in ship design.
2. Practice of shipbuilding: These books covered the design of the hull and other equipment for manufacture in the shipyard. This has evolved in conjunction with the science of ship design to include material science, ship strength, outfitting, and ship production. The books of W. Sunderland, *Shipbuilding*, published in 1755; of M. Murray, *A Treatise on Shipbuilding*, published in 1765; and of F. H. Chapman, *Architecture Navalis Mercatoria*, published in 1768, were widely used by the naval architects of that period.

These two directions have continued in the parallel to the present. The vessels in Figures 1–9 were designed by naval architects having the benefit of an engineering education and a large technical library. Yet a number of smaller boats are built by skilled crafters working from plans or hard-earned experience. The main difference is that while the boat builder can evaluate the design after it is built, the naval architect is often able to predict the vessel behavior using engineering analysis at the design stage and correct problems before the vessel is built.

This difference was appreciated by the early colonists, who brought with them naval architects and ship builders to assist them in settling the New World. Records indicate that by 1700, four shipyards were operating in Philadelphia. These yards and the other colonial shipyards began a fruitful interchange of European practice, which gradually led to the world-famous American sailing-vessel designs of the Revolutionary War and the War of 1812, and later the clipper ships and schooners. When Congress met in 1776, they ordered 13 frigates to defend the colonies. The 697-ton frigate U.S.S. *Rayleigh* (131 ft long × 34 ft wide) was completed in 60 days and attracted so much favorable attention that after its capture it was towed and dry-docked at Portsmouth, England, where it was examined by the naval authorities in 1779. The lack of a standard U.S. naval design practice allowed several innovations, which contributed to the successful U.S.S *Rayleigh* design.

1. Proper guidance in deciding to adopt a fast-maneuverable figure instead of building the heavier ships of the line favored by the British.
2. Selection of a French hull form, which was considered faster at that time, thus adopting the best ship concept. Records indicate that Benjamin Franklin had made experiments in a model basin as early as 1764.
3. Tight weight control and the use of lightweight yellow pine planking, in contrast to the British use of heavy oak.
4. Use of standard items and design for production to minimize cost and delivery time.

Records indicate that up to the Revolutionary War naval architects were apprenticed. This changed in Philadelphia when the ShipWrights Company was organized for the study and improvement of naval architecture. In 1795, this guild established a school for the instruction of ship carpenters and established a technical library of the best books from Europe.

B. The Nineteenth Century

The transition in the nineteenth century from sailing vessels to the modern metal vessel propelled by a screw propeller is a fascinating study of government and naval policy developing under pressure from other nations, designers, entrepreneurs, and reformers. This transition began with the American appreciation of the value of time and speed, which led to major innovations in the early 1800s.

The first innovation was the introduction of vessels propelled by steam engines turning paddlewheels and/or screw propellers. These first appeared on the Hudson, James, Ohio, and Mississippi rivers, and later were used in coastal trading. The second innovation was the introduction by 1820 of packet ships, which sailed on the first and sixteenth of each month to England. With the packet ships it became profitable to load as much cargo as possible, which in some cases exceeded the design of the vessel. Thus by 1830, recommendations for the maximum loading were published in England. This was followed by the load-line marks or Plimsoll marks (1875), which now appear on cargo ships.

The third innovation was the replacement of wood by iron and later steel in shipbuilding. Beginning in 1825 with the iron steamboat, there was a gradual development of iron shipbuilding in the United States that led to a number of iron vessels being built in the 1840s and 1850s.

While the British iron ship *Great Eastern* (692 ft long 3 120 ft wide) is considered a predecessor of modern metal ships, the transition from wood to metal was hastened by the Confederate Navy's program of developing floating mines, iron submarines, and iron-clad vessels to break the Union blockade of the Southern ports. This forced the Union Navy to begin building ironclad vessels such as the U.S.S. *Cairo*, restored in Vicksburg, Mississippi.

The Confederate program lead to the building in 1861 of the iron-hulled submarine *Pioneer* in New Orleans and then the submarine C.S.S. *Hunley* in Mobile, which sank the U.S.S. *Housatonic* in Charleston harbor in 1864. The success of the iron-clad C.S.S. *Virginia* (*Merrimack*) in sinking the wooden U.S.S. *Cumberland* and U.S.S. *Congress* in Hampton Roads in 1862 lead to the introduction of armor plate in naval design. The submarine *Pioneer* was later salvaged and is on display in the Presbyter across from Jackson Square in New Orleans.

In addition to these three innovations, 1880–1900 was a period when scientific revolution was taking place. Theoretical and experimental research was used to solve engineering problems. This began the rational approach to the design of ships and floating structures, which has been evolving to the present. The industrial and scientific revolutions resulted in naval architecture being introduced as a course at major European universities, where it was studied after a rigorous introduction to mathematics, physics, and the mechanics of solids and fluids. Graduates from these courses began teaching naval architecture throughout the United States, Canada, and South America. The growth of naval architecture in the United States finally led to the formation, in the 1890s, of the Society of Naval Architects and Marine Engineers. They continue as forums for the development of ship design and provide invaluable exchange of technical information.

C. Role of the Modern Naval Architect

From the colonial period to the nineteenth century, the naval architect's activities gradually broadened to encompass a wide range of tasks in vessel design and production shown in Fig. 10. The tragic sinking of the R.M.S. *Titanic* in 1912 made ship safety an important issue. This resulted in the government regulation of passenger and commercial vessels and the development of inspection procedures to insure adequate ship strength and proper safety in machinery design. With time, the marine industry grew and naval architects began to specialize, as shown in Fig. 10.

The simple drawings used by early American shipwrights developed into an elaborate set of plans outlined in Table II. In commercial shipbuilding, these are reviewed at the early stages of vessel design by agencies such as the U.S. Coast Guard, as well as classification societies such as the American Bureau of Shipping, to certify that the design and vessel construction satisfies regulations and that the vessel is safe.

As Fig. 10 illustrates, there is a large amount of information transferred in this process. The computer has streamlined this somewhat. Nevertheless, to provide a strong background for the naval architect to successfully enter the marine industry has been the challenge faced by the educational group.

A typical naval architect graduating with a B.S.E. degree will have completed a four-year engineering course, which includes the courses in Table III. These

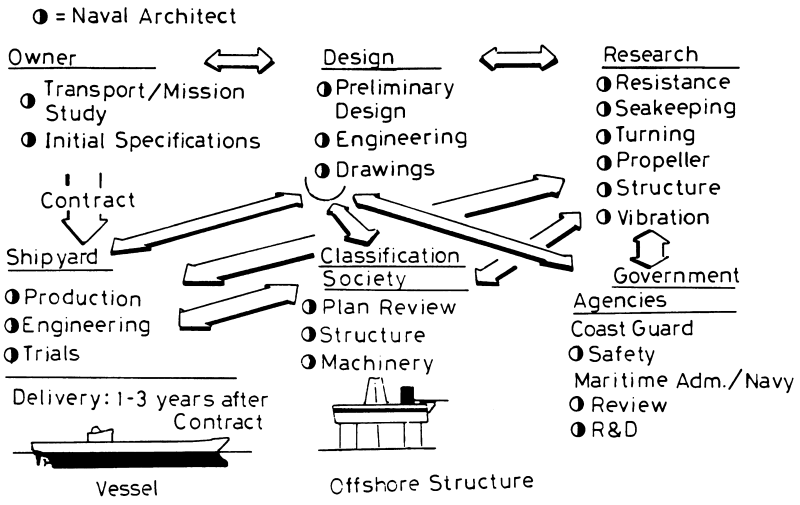


FIGURE 10 Naval architect's role in vessel design production.

TABLE II List of Plans Developed for Commercial Ship Design

Overall design
Outboard profile, general arrangement
Inboard profile, general arrangement
General arrangement of all decks and holds
Arrangement of crew quarters
Arrangement of commissary spaces
Lines
Shell Expansion
Hull strength
Midship section
Steel scantling plan
Marine engineering
Arrangement of machinery—plan views
Arrangement of machinery—elevations
Arrangement of machinery—sections
Arrangement of main shafting
Power and lighting system—one-line diagram
Fire control diagram by decks and profile
Ventilation and air conditioning diagram
Diagrammatic arrangements of all piping systems
Heat balance and steam flow diagram—normal power at normal operating conditions
Electric load analysis
Safety calculations
Capacity plan
Curves of form
Floodable length curves
Preliminary trim and stability booklet
Preliminary damage stability calculations

courses include use of computers and databases, as well as structural and hydrodynamic laboratories (Fig. 11). With the increasing number of naval architects engaged in research and development, advanced degrees at the Master's and Ph.D. levels are starting points. This has led to government and industry support of a number of long-range research projects conducted at university, private, and government laboratories, as well as shipyards in the areas listed under research in Fig. 10.

III. PROBLEMS IN SHIP DESIGN

The diversity of ships, submarines, and other floating structures, and the continual evolution of needs result in many unique design requirements. Because of the interrelationship of the vessel, its payload, propulsion system, and crew habitability, a system approach is required

TABLE III Typical Naval Architecture and Marine Engineering Courses Taken with Other Engineering Courses for a B.S.E. Degree (NA&ME)

Sophomore (second year)
Introduction to ship and offshore design
Form calculations and stability
Junior (third year)
Offshore structure and ship strength
Marine engineering I
Computers in naval architecture
Ship resistance and propulsion
Offshore structure and ship dynamics
Senior (fourth year)
Offshore structure and ship design
Offshore structure and ship design project
Maritime economics
With elective courses
Ship hydrodynamics II
Computer aided design (CAD)
Offshore structures and ship strength II
Marine engineering II
Hull surface design
Small craft design
Admiralty law for engineers
Materials for marine design
Ship production

to design the vessel. These subproblems are discussed here.

A. Weight and Buoyancy Considerations

The basic consideration of the weight of the vessel, W, and the submerged volume, V, which are related by Archimedes's law,

$$W = \gamma V = \gamma C_B 3\, L3\, B3\, T, \tag{4}$$



FIGURE 11 Resistance test of model cargo ship in NAME 3150 Student Laboratory. [Courtesy R. Latorre.]

where L is length, B is beam, and T is draft, or by the Bernoulli equation,

$$W = \int_A P_z \, dA. \tag{5}$$

where γ is the density of water, weight/volume, P_z is the vertical component of pressure acting normal to the hull, A is the surface area of submerged hull, and C_B is the block coefficient.

For illustration, the vessel weights and volumes are designated as the hull (W_h, V_h), machinery and fuel (W_m, V_m), payload (W_p, V_p), and crew/passengers (W_c, V_c).

The naval architect must ensure that

$$W = W_h + W_m + W_p + W_c, \tag{6}$$

and in the case of a submarine, the underwater volume must be adequate to support the vessel weight as well as contain the complete submarine systems.

$$V = V_h + V_m + V_p + V_c. \tag{7}$$

1. Relationship of Requirements

When a vessel mission is stated, four elements are usually given:

1. The vessel speed and endurance, which influence V_m and W_m and in some cases impact W_h.
2. The payload, which determines W_p and impacts W_h.
3. Restriction in length L, beam B, or draft T (Fig. 12), which are related to W and V as well as W_h and V_h, and directly impact W_m and V_m.
4. Price and operating cost limitations, as well as time of delivery.

This leads to two basic procedures in design. The first procedure is simply to modify an existing design to vary speed or payload or to operate in a different environment. In the past this led to a very slow evolution in design until "break-throughs," such as steam engines, diesel engines, and containers, forced the abandonment of the old design. The second procedure is to search a design base and after preliminary studies develop realistic estimates for the values in Equations (4), (5), and (6). The second approach has been adopted when the basic elements of conventional design are mixed with new concepts, such as the cargo-handling systems in Figs. 6 and 7.

2. Weight Calculations

In the design, a number of preliminary arrangements such as Figures 13 and 14 are prepared to scale. The weight

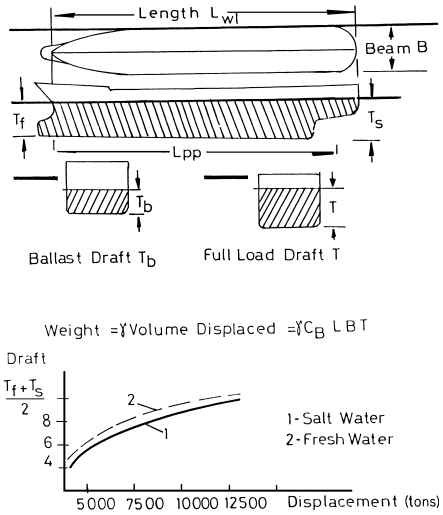


FIGURE 12 Geometry of floating ship.

[Eq. (6)] as well as the center of gravity can then be resolved for each element $w_i(x_i \; y_i \; z_i)$ in the vessel.

$$W = \sum_i w_i. \tag{8}$$

The longitudinal center of gravity is

$$LCG = \bar{X} = \frac{\sum_i x_i w_i}{W}. \tag{9}$$

The vertical center of gravity KG, G or VCG:

$$G \text{ or } VCG = \bar{Z} = \frac{\sum_i z_i w_i}{W}. \tag{10}$$

The transverse center of gravity is

$$\bar{Y} = \frac{\sum_i y_i w_i}{W}. \tag{11}$$

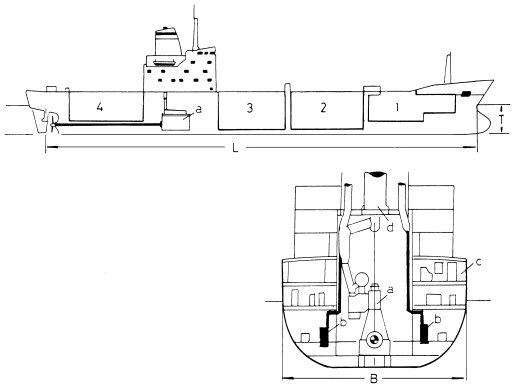


FIGURE 13 Arrangement of cargo and machinery. 1–4, Cargo/container holds. (a) Slow-speed diesel engine. (b) Diesel generator. (c) Control room. (d) Boiler heated by engine exhaust.

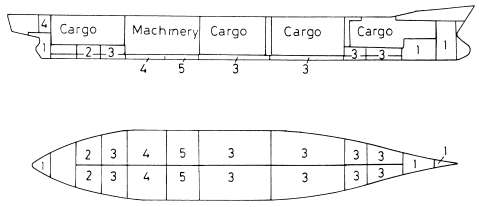


FIGURE 14 Tank arrangement. 1, Salt-water ballast. 2, Fresh water. 3, Fuel oil. 4, Lube oil. 5, Diesel oil.

Y should coincide with the vessel centerline for normal sailing.

3. Hull Form Development and Hydrostatic Calculations

The development of the hull or hull surface has been a task approached mathematically as well as artistically. The hull must form a low drag, underwater envelope that will contain the vessel payload, machinery, etc., and it must be designed for production. In some cases developable surfaces have been adopted to reduce the amount of hull plate forming. In the past, shipyards used a large room or "mold loft" where the hull would be plotted as a series of contours, which are defined as the hull lines. These lines would be adjusted until a "fair" surface would be defined in three planes. The result is a hull of a specified volume V defined by a table of offsets (x, y, z) as well as mathematical curves $y = f(x, z)$ or $y(z)_{x=const}$ shown in Fig. 15.

Returning to the considerations of a body in equilibrium, it is necessary to define the magnitude of the forces as well as their line of action. In the preceding weight calculations, these are obtained for the hull by equations (8–11). In an analogous manner, it is necessary to determine the buoyancy force (FB) as well as its longitudinal and vertical locations denoted by LCB and KB (or B), respectively. These calculations of FB, LCB, and B are made for the vessel in various degrees of heel and trim.

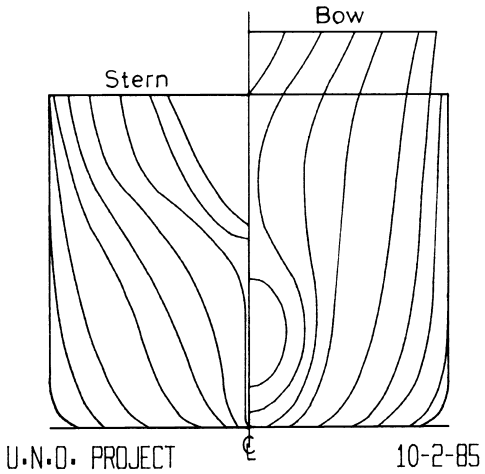


FIGURE 15 Body plan of ship.

While the details are given in the reference works cited in the Bibliography, it is possible to briefly describe the method adopted. The hull envelope is assumed symmetric along its length and sliced into n ranges between 10 and 20 stations. At each station the hull offsets or half breadths can be defined as $y(z)_{x=const}$. The sectional area of each station is given by:

$$S_n = \int_{keel}^{draft} y(z)_{x=const}\, dz. \tag{12}$$

The hull volume is then obtained:

$$V = \int_{bow}^{stern} S_n\, dx. \tag{13}$$

The longitudinal center of buoyancy is then obtained:

$$LCB = \frac{\int_{bow}^{stern} X \cdot S_n\, dx}{V}. \tag{14}$$

In a similar manner, it is possible to determine the vertical center of buoyancy VCB, KB, or B. In each of these integrations, a numerical integration procedure is adopted that is generally run on a computer.

The results of these hydrostatic calculations are prepared in graph format, as well as in numerical files for performing shipboard loading and safety-check calculations. A typical graph of the vessel weight in fresh and salt water as a function of its average draft is shown in Figure 12b.

4. Metacentric Stability

One of the concepts unique to naval architecture is that of the metacenter, denoted as M in Figure 16. As the vessel heels at an angle υ, the center of buoyancy B moves, and a couple is created between the vessel weight W and the force of buoyancy FB. The restoring arm, which is denoted as GZ, can be estimated for small angles of heel from the distance between the center of gravity KG and the metacenter KM measured from the keel:

$$\mathrm{GZ} = \mathrm{GM} \sin \upsilon, \qquad \text{for small } \upsilon, \tag{15}$$

where:

$$\mathrm{GM} = (\mathrm{KM} - \mathrm{KG}).$$

For safety assessment, the righting arm GZ is calculated for the entire range of heel angles up to 80–90°. Obviously, if GZ is zero or negative then the vessel will have no positive restoring arm and will become unstable.

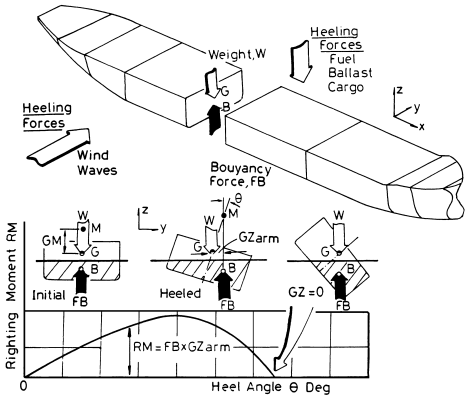


FIGURE 16 Ship intact stability.

Two cases are usually taken into account in these calculations. The first case considers the hull floating intact, and the second case considers the hull flooded in one or more compartments. The second case reflects the sinking of the vessel by flooding, as the R.M.S. *Titanic* experienced when the iceberg tore its hull. The division of the hull into "floodable lengths" to survive such accidents is called subdivision. Passenger-carrying vessels are designed both to comply with intact stability regulations, as well as meet the subdivision requirements for floodable length.

Because the safety of the passengers is at stake, the vessel is checked after building to establish its GM. This is done by moving weights transversely across the deck and recording the movement of a pendulum. An example of this "inclining experiment" is shown in Fig. 17.

B. Structure

The naval architect, in contrast to other engineers, was unable until the 1970s to estimate the loads the ship structure might encounter in various seas when in service. This was because a floating vessel is supported by an ever-changing sea surface, which creates both static and dynamic loads. This problem is also complicated because wooden or steel structures are continually degraded by corrosion and marine life.

1. Classification Society's Rules for Structure

The operation of a cargo or passenger carrying vessel exposes the cargo, passengers, and crew to some risk of loss. This requires insurance that will cover possible losses in goods and life. This insurance requires an appraisal of the vessel's soundness and seaworthiness. This is done by nonprofit organizations called classification societies. By providing plan approval before vessel construction, and periodic survey after a vessel begins construction, the classification society classes the vessel in its register. This certifies that the vessel is sound and seaworthy and conforms to recognized standards of sound construction.

In the early period of wooden ships, the classification society functioned as an agency providing periodic inspection. With the introduction of iron and steel, the naval architect converted from wood to metal ship structure and in the process reduced the size and thickness of the structural members. To provide guidance, the classification societies issued rules, which have gradually been improved. In the period when the loads on the ship were not clear, the rules were guided by the structure used of ships that had been successful in service. Full-scale measurements and model tests led to successful rational structural analysis in deriving the present classification rules. The development of structural analysis continues today, especially in the areas of applying probabilistic methods and finite-element analysis to marine structure analysis.

2. Design Loads

To establish the required structural elements or scantlings, it is necessary to determine the maximum loads expected in service:

1. Calm-water bending moments and shear forces.
2. Wave-induced bending moments and shear forces.
3. Impact and hull springing loads.

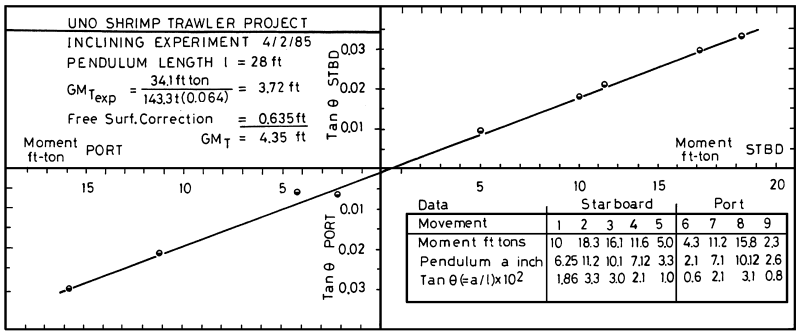


FIGURE 17 Inclining experiment results.

4. Thermal loads due to the temperature differences between the bottom structure and the upper structure.
5. Dynamic loads due to cargo movement and hydrodynamic pressure.
6. Other service loads encountered in the docking, launching, mooring, and worst-case grounding/collision scenarios.

The next section discusses forces 1–3.

3. Longitudinal Bending Moment Estimation

For longitudinal strength, the floating hull is analyzed as a built-up beam that has a distributed load $F(x)$ representing the difference between the local weight $W(x)$ acting downward ($-z$ direction) and the buoyancy force FB(x) acting upward ($+z$ direction) along the length of the structure (Fig. 18).

$$F(x) = W(x) - \mathrm{FB}(x). \tag{16}$$

The term $W(x)$ is defined for a number of conditions representing the vessel with or without its payload. The term FB(x) can be calculated using the "standing wave" approach, which assumes the vessel is supported on a trochoidal shaped wave equal to the vessel length and a crest to trough distance equal to (1/20) L. This produces two extreme cases used in design (Fig. 18):

1. Wave crest at middle ($x = \frac{L}{2}$) causes the structure to bend upward or "hog."
2. Wave crest at each end ($x = 0, L$), causing the structure to bend downward or "sag."

Using the results of classical beam theory, the corresponding shear and bending moments can be obtained from the distributed load $F(x)$ in Eq. (16).

As more knowledge of wave bending moments and shear force became available, it became possible to consider the total bending moment M_t as the sum of the maximum still-water bending moment M_{sw} and the maximum wave-induced vertical bending moment in waves, M_w:

$$M_t = M_{sw} + M_w. \tag{17}$$

The total vertical shearing force F_t is the sum of the maximum still-water shearing force F_{sw} and the maximum wave-induced shearing force F_w:

$$F_t = F_{sw} + F_w. \tag{18}$$

A series of studies has shown that wave-induced loads can be low frequency, corresponding to the ship motions, and in some cases the wave impact can excite a high-frequency hull girder vibration called springing. This phenomenon has been observed on the long slender ore carriers sailing on the Great Lakes. For this ship, the maximum springing bending moment M_{sp} at the vessel midships is added to Eq. (17).

4. Structure Design

The next level of structure design treats the hull structure elements that are illustrated in Figure 19. Here the results of beam theory are used to define the required longtime section modulus (SM) of the hull. Experience provides the designer with reasonable values of permissible bending stress σ_p for the hull materials and intended service. The SM is then related to the maximum value of the bending moment M_t:

$$\mathrm{SM} = M_t/\sigma_p. \tag{19}$$

Knowing SM, the naval architect proceeds to select the structural elements in Fig. 18 using classification society

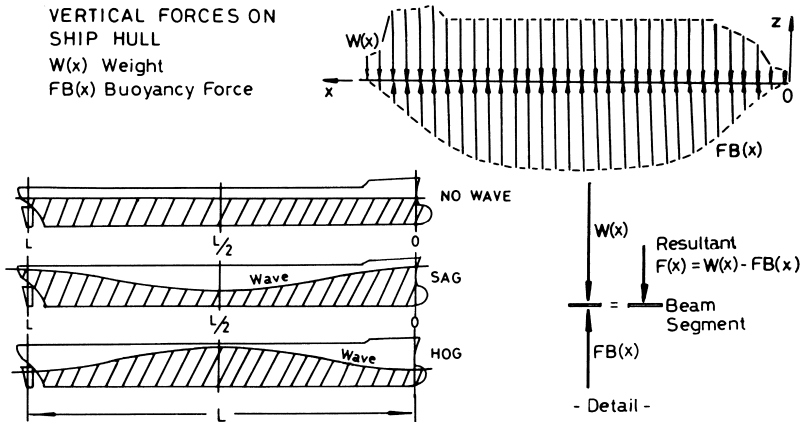


FIGURE 18 Illustration of vertical forces on ship hull and hog-sag condition caused by wave.

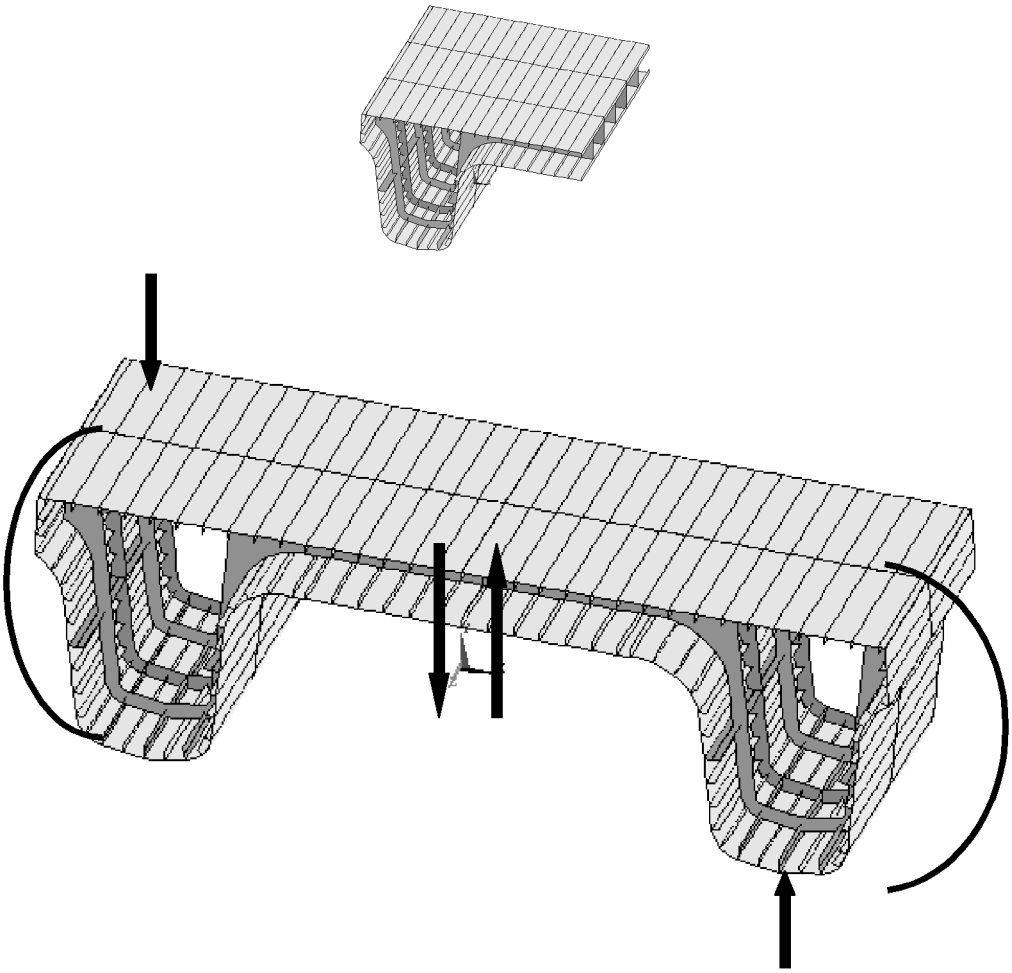


FIGURE 19 Ship structure.

rules for guidance as well as performing detailed calculations using finite-element structural analysis computer programs. The resulting structure must provide the required SM from Eq. (19) as well as have adequate strength from the viewpoint of the three components of hull stress:

1. Primary stress σ_1, due to the hull bending as a beam under longitudinal load $F(x)$ in Figure 16a.
2. Secondary stress σ_2, due to the normal loading of a stiffened panel, such as the plate between two adjacent frames.
3. Tertiary stress σ_3, due to out-of-plane deflection of the hull plates.

The selection of the structure on this basis is providing the designer with a comprehensive understanding of the floating structure for obtaining the optimum structure in terms of structural weight, structural strength, and structural cost.

C. Hydrodynamics

The three main areas of marine hydrodynamics are:

1. Powering—floating vessel resistance and propulsion.
2. Seakeeping—floating vessel motion in waves.
3. Maneuvering—ability of floating vessels to maintain and change course.

Ship performance was initially indexed of speed, and the naval architect focused on the design of the hull and propeller to minimize power for a given speed. As the capability to satisfactorily resolve the powering problems was developed, more emphasis was placed on sustaining

performance in waves and selecting designs that would safely navigate channels and harbors developed centuries ago for smaller vessels. The impact of ship performance was conclusively demonstrated at the turn of the century by Admiral Dewey, using modern ships to defeat the Spanish fleet at Manila Bay, and by the later defeat of the Russian Baltic fleet at Tsushima by the Japanese Admiral Togo, making use if the speed and maneuverability of his fleet. These victories established the need for naval ships with superior hydrodynamic performance. This began a large naval research effort that continues throughout the world to develop experimental, theoretical, and numerical tools for the hydrodynamic design of ships and other vehicles floating above and below the water surface.

1. Powering

The hydrodynamic forces on a floating body have been a classic problem that remained a puzzle until the work of the English scientists William Froude and Osborne Reynolds in the 1800s. While scale models were easily built and then towing resistance measured, extrapolation of the results to full scale was not successful. The problem was that the resistance R_t of a moving object floating on the water surface is due to three components:

1. Wave-making resistance—R_w: the energy used to create waves (Fig. 20).
2. Pressure resistance—R_f: the resistance associated with the incomplete pressure recovery at the stern.
3. Viscous resistance—R_v: the resistance associated with the formation of a viscous boundary layer on the hull surface.

$$R_t = R_w + R_f + R_v. \tag{20}$$

Unfortunately, the extrapolation of R_w, R_t, and R_v is governed by different similitude relationships, so the model test did not correctly model the full-scale vessel resistance. The work of Froude and Reynolds developed the basis for the Froude extrapolation method, where the model resistance R_{tm} is measured at scale speed V_m determined by Froude number F_n:

$$F_{n_s} = F_{n_m} = \frac{V_s}{\sqrt{gL_s}} = \frac{V_m}{\sqrt{gL_m}} = const, \tag{21}$$

when g is the gravity constant so

$$V_m = V_s/\sqrt{\lambda}, \tag{22}$$

where γ is the scale factor L_s/L_m.

The results of detailed measurements are then used to correct the viscous resistance R_v for the actual model and later for the ship Reynolds number

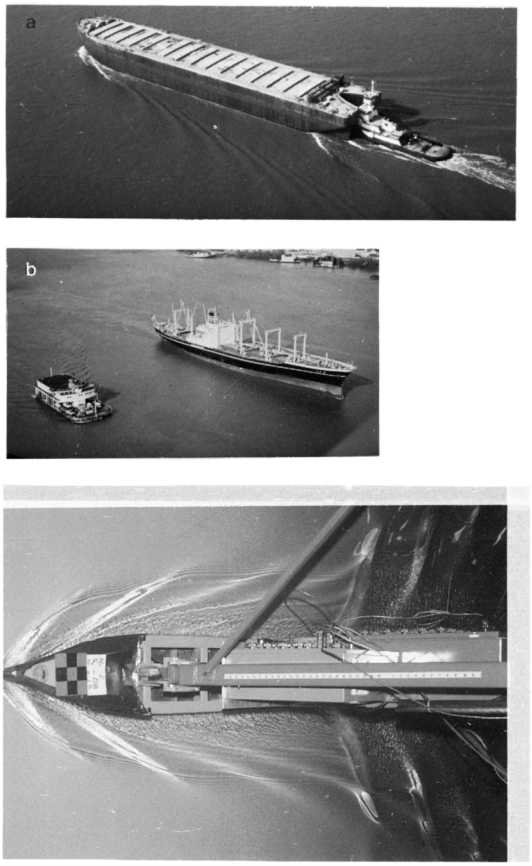


FIGURE 20 Wave pattern from moving ships. [Courtesy of D. Lewane.] (a) Barge being pushed by tug. (b) Ferry boat and cargo lines. (c) Model ship in shallow water. [Courtesy of D. Lewane.]

$$R_{n_s} = \frac{V_s L_s}{\nu_s} \neq R_{n_m} = \frac{V_m L_m}{\nu_m}, \tag{23}$$

where ν is the kinematic viscosity of the water. This method allows the accurate prediction of the ship's resistance from the model tests.

To ensure accuracy, models with a scale ratio as large as $\gamma = 20$ are tested in large experimental towing basins such as the towing tank at the David W. Taylor Naval Ship Research and Development Center outside Washington, DC. (Fig. 21).

In a parallel effort, theoretical models have been developed to calculate the viscous resistance R_v, as well as the wave-making resistance R_w and pressure resistance R_f. This work, in conjunction with the model experiments, is providing a detailed understanding of the basic mechanisms of resistance and the methods to reduce resistance.

The design of the propulsor is the second task in powering. There are a number of alternative systems, including a conventional propeller (Fig. 22), a propeller with movable blades called a controllable-pitch propeller, a propeller in a

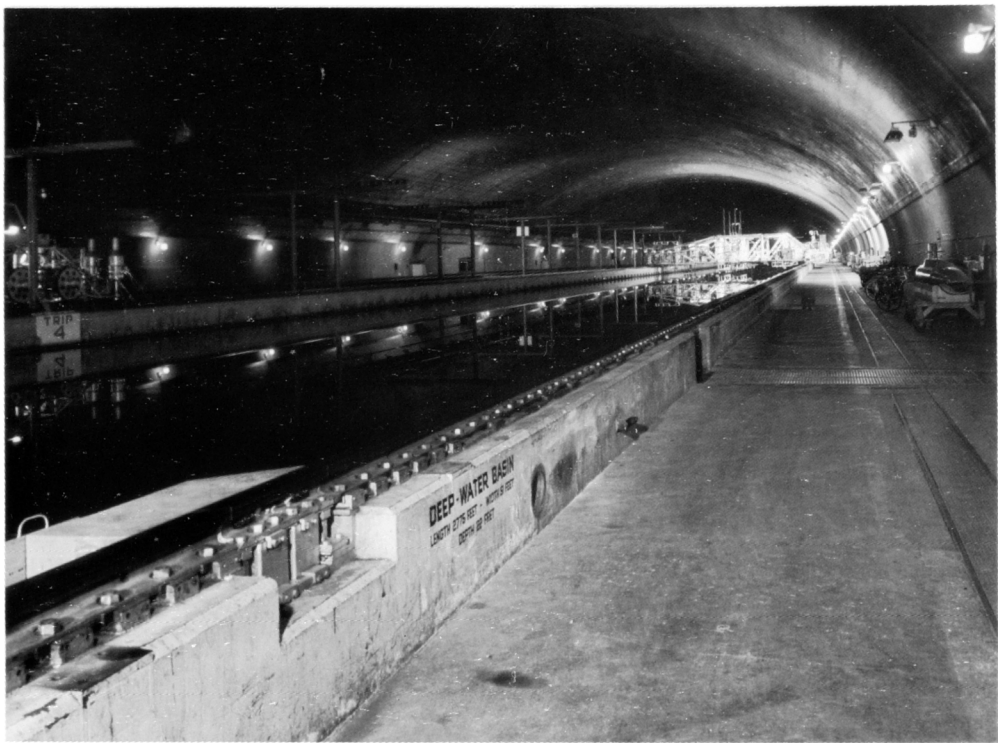


FIGURE 21 Towing tank at David W. Taylor Naval Ship Research & Development Center. [Courtesy of U.S. Navy.]

nozzle (Fig. 23), as well as counterrotating propeller-pump jets and wind-powered system. When the propulsors operate in the wake behind the ship, there is a need to account for the unfavorable effect of the hull and the favorable effect of the propeller in accelerating the flow around the vessel stem.

The design of the propeller is also a critical element in the vibration of the ship, which is due to the propeller blades passing through a nonuniform inflow behind the hull and transmitting periodic forces into the shafts and the ship hull. Because the propeller develops lift along the blade surface, there are conditions when the pressure on the suction side falls below the vapor pressure, and vapor cavities form and collapse on the blade surface. In addition to reducing the propeller performance, the cavity collapses and creates noise that enables the detection of the ship by underwater acoustic sensors.

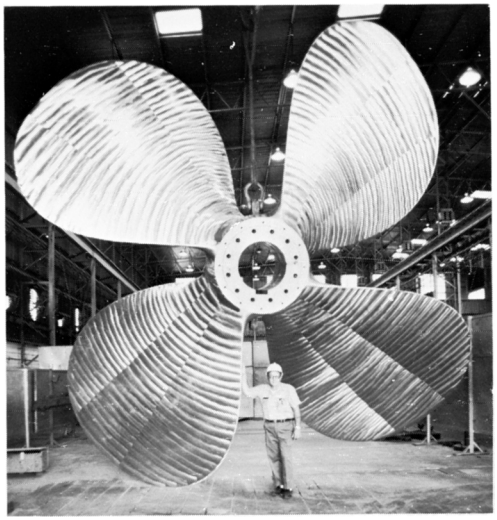


FIGURE 22 Large marine propeller 23 ft in diameter weighing 84,000 lb. [Courtesy of G. Harney, Avondale Shipyards Inc.]

2. Seakeeping and Maneuvering

The development of seakeeping and maneuvering has followed a similar pattern. After the introduction of the appropriate similitude relationships, scale-model tests in towing tanks fitted with scale wave-makers were used to study the vessel seakeeping. Large maneuvering test tanks were built to obtain the coefficients used in the differential equations describing ship maneuvering. At first the equations were linear, which restricted the numerical work to small motions. Then, as the techniques became more refined, the nonlinear coefficients were measured. This formed the basis for the realistic performance of video games and other personal computer software.

The tests in waves have progressed from testing in regular waves with constant wave height and length to testing in an irregular wave train, which is generated to duplicate

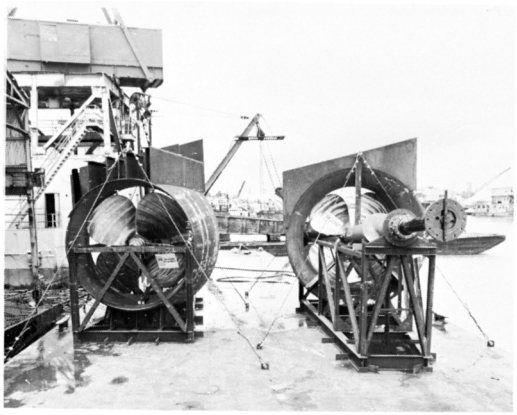


FIGURE 23 Ducted propeller set (10 ft diameter). [Courtesy of G. Harney, Avondale Shipyards Inc.]

a given sea spectrum. As with powering and maneuvering, there is an ongoing development of computer models to describe the motion and loads of the floating structure.

IV. DESIGN SPIRAL

The synthesis of all the activities in Figure 10 is an ongoing task of the naval architect and is now being accomplished by exchange of data via data banks, which are managed by the owner, design group, shipyard, and regulatory and classification societies. The emphasis on achieving economies in production has, in some cases, resulted in a system that at a certain stage makes estimates of the vessel cost. This approach is spearheading the move to "design for production," which has been cited as one of the reasons for the success of overseas shipyards in the rapid delivery of standard vessels built recently.

Having touched on the role of the naval architect, it is possible to consider the whole picture of the design of

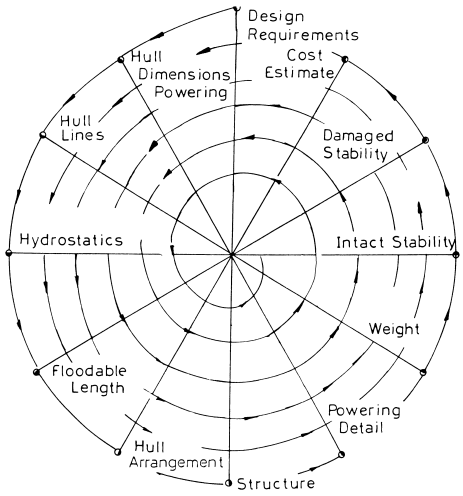


FIGURE 24 Ship design spiral.

TABLE IV Breakdown of Design Work for a Large Merchant Ship

Phase	Work
Concept	20–30 person days
Preliminary design	250–350 person days
Concept design	5000–6000 person days
Detail design	50,000–60,000 person days

floating vessels. This is illustrated in Fig. 24, first used by Professor Evans, as a spiral that converges on the desired vessel after successive interactions. For a large merchant ship, this can involve between 50,000 and 70,000 person days, which are broken down in Table IV.

The design spiral begins with the vessel mission requirements that are discussed in the Sections I and III.A.1. The next step is the selection of the proportions and an estimate of the powering requirements discussed in Sections III.A.1 and III.C.1. Then the lines and body plan are developed and the hydrostatic calculations are performed as outlined in Section III.A.2. This is followed by a calculation of vessel floodable length and arrangements. Then the structural calculations are made, following the procedure in Section III.B. At this point, the vessel is defined and the powering refined, as well as the weights and volumes. This is followed by a more accurate stability calculation (Section III.A.4). The damaged stability calculation is made to ensure the subdivision is adequate, and then the spiral ends with a cost estimate. Because this is an iterative process, the design will evolve to a detail design level, which for a large merchant vessel requires 5000–6000 person days of effort. The generation of the building plans and the other drawings for a large merchant ship required by the classification society and regulatory bodies will occupy the remaining 45,000–55,000 person days.

SEE ALSO THE FOLLOWING ARTICLES

AIRCRAFT PERFORMANCE AND DESIGN • FLUID DYNAMICS • MARINE ENGINES • OFFSHORE STRUCTURES (MARINE ENGINEERING) • SPACECRAFT SYSTEMS DESIGN AND ENGINEERING

BIBLIOGRAPHY

Allmendinger, E. E. (ed.) (1990). "Submersible Vehicle Systems Design," Society of Naval Architects and Marine Engineers.

Baker, W. A. (1976). Commercial Shipping and Shipbuilding in the Delaware Valley. *Proc. 1976 Spring Mtg. Society of Naval Architects and Marine Engineers, New York*.

Basin, M., and Latorre, R., "Development of High-Speed Craft with Aero-Hydrodynamic Support," Proceedings from the International Conference on FAST '97 Transportation, Vol I, Baird Publications, Sydney Australia, July 1997, pp. 85–89.

Bhattacharyya, R. (1978). "Dynamics of Marine Vehicles," Wiley, New York.

Chapman, F. H. (1975). "Architectura Navalis Mercatoria" (reprint). Allan Coles, London.

Claton, B. R., and Bishop, R. E. D. (1983). "Mechanics of Marine Vehicles," Gulf Publishing, Houston.

Clifford, R. "INCAT-THE FIRST 40 YEARS," Baird Publications, Australia, 1998, pp. 148–152.

Davis, C. G. (1983). "American Sailing Ships, Their Plans and History," Dover Publications, New York.

Gutelle, P. (1979). "Architecture du Voilier," Tome 1. Theorie, Editions Maritimes et Outre-Mer, France.

Hatch, G. N. (1971). "Creative Naval Architecture," Thomas Reed, London.

Hughes, O. (1983). "Ship Structural Design: A Rationally Based, Computer-Aided, Optimization Approach," Society of Naval Architects and Marine Engineers, New Jersey.

Kennel, C., Lavis, D., and Templeman, M. "High-Speed Sealift Technology," SNAME-Marine Technology, Vol. 35, No. 3, July 1998, pp. 135–150.

Latorre, R., and Herrington, P. D. "Design of a 33 Knot Aluminum Catamaran Ferry," SNAME-Marine Technology, Vol. 37, No. 2, Spring 2000, pp. 88–99.

Newman, J. N. (1977). "Marine Hydrodynamics," MIT Press, Cambridge.

Perry, M. (1965). "Infernal Machines, The Story of Confederate Submarine and Mine Warfare," Louisiana State University Press, Baton Rouge, LA.

Rawson, K. J., and Tupper, E. C. (1983). "Basic Ship Theory," Vols. I and II. Longman, New York.

Schonknect, R., (ed) (1980). "Ships and Shipping of Tomorrow," Cornell Maritime Press, Carterville, Md.

Taggert, R., (ed.) (1980). "Ship Design and Construction." Society of Naval Architects and Marine Engineers, New York.



Offshore Structures (Marine Engineering)

James F. Wilson
Duke University

I. Historical Perspective
II. Platforms
III. Moorings
IV. Pipelines
V. Challenges

GLOSSARY

Anchor legs Mooring line array for a buoy, especially for catenary anchor leg mooring (CALM) systems.
Catenary Shape of a flexible cable hanging freely between two fixed points of support.
Compliant tower Truss structure anchored by piles and held upright by mooring cables.
Gravity platform Offshore concrete structure that is kept upright by self-weight only.
Jacket structure Truss structure with steel tubular legs through which piles are driven into the seabed; also a jacket-template structure.
Jackup rig Fixed-leg structure with a platform with an adjustable elevation.
Jetting Process of making a hole or trench in the seabed with a stream of compressed air.
Laybarge Barge for assembling and deploying pipelines to the seabed.
Reel barge Barge for deploying pipelines from an onboard spool wound with the pipe string.
Riser Vertical pipeline between the wellhead and ship or platform; may contain drilling equipment or electric power lines, or may transport oil or gas.
Semisubmersible Platform structure that floats on pontoons positioned below the level of waves.
Spread mooring Configuration in which perimeter cables extend outward from the ship to fixed buoys or anchors.
Stingers Pontoons placed under deployed pipelines at launching for the purpose of avoiding excessive pipe bending stress.
Tension leg platform (TLP) Floating platform structure moored with taut, vertical cables to seabed piles.

OFFSHORE STRUCTURES are constructed facilities designed for use on or above the continental shelves and on the adjacent continental slopes. These structures take many forms and serve a multitude of purposes: towers for microwave transmission, installations for power generation, portable pipeline systems for mining the ocean floor,

and a few platforms and floating islands that serve as resort hotels. Most structures offshore, however, have been built to support the activities of petroleum industries—activities that include the exploration, drilling, production, storage, and transportation of oil. Exploratory drilling is done from mobile platforms or carefully positioned ships; production and storage operations involve more permanent structures; and pipelines, buoys, and mooring systems for floating structures and ships support all oil acquisition activities. The engineer involved in designing these marine structures must rely on the knowledge and experience of meteorologists, oceanographers, naval architects, geologists, and material scientists. The marine engineer's goal is to conceive and design lasting structures that not only can withstand the adverse conditions of high winds and waves, earthquakes, and ice, but also will remain in harmony with the natural environment.

I. HISTORICAL PERSPECTIVE

The earliest offshore structure for oil drilling was built about 1887 off the coast of southern California near Santa Barbara. This was simply a wooden wharf outfitted with a rig for drilling vertical wells into the sea floor. More elaborate platforms supported by timber piers were then built for oil drilling, including installations for the mile-deep well in Caddo Lake, Louisiana (1911) and the platform in Lake Maracaibo, Venezuela (1927). Soon after these early pier systems were built, it became apparent that the lifetime of timber structures erected in lakes or oceans is severely limited because of attacks by marine organisms. For this reason, reinforced concrete replaced timber as the supporting structure for many offshore platforms up to the late 1940s. Over the next 50 years about 12,000 platform structures were built offshore, usually of steel but more recently of precast concrete. Figure 1 shows some of these structures, of which the steel truss is the most common. The chief functions and engineering features of these structures, together with their supporting components (mooring systems and pipelines), are discussed in this article.

Offshore mooring systems have a variety of configurations. All have anchors or groups of piles in the seabed with flexible cables (ropes or chains) leading from them to buoys, ships, or platform structures. The function of a mooring system is to keep the buoy, ship, or platform structure at a relatively fixed location during engineering operations. Engineering efforts in mooring systems have focused in recent years on the development of new anchor configurations with higher pullout loads, larger capacity and lower cost of installation for deeper water applications.

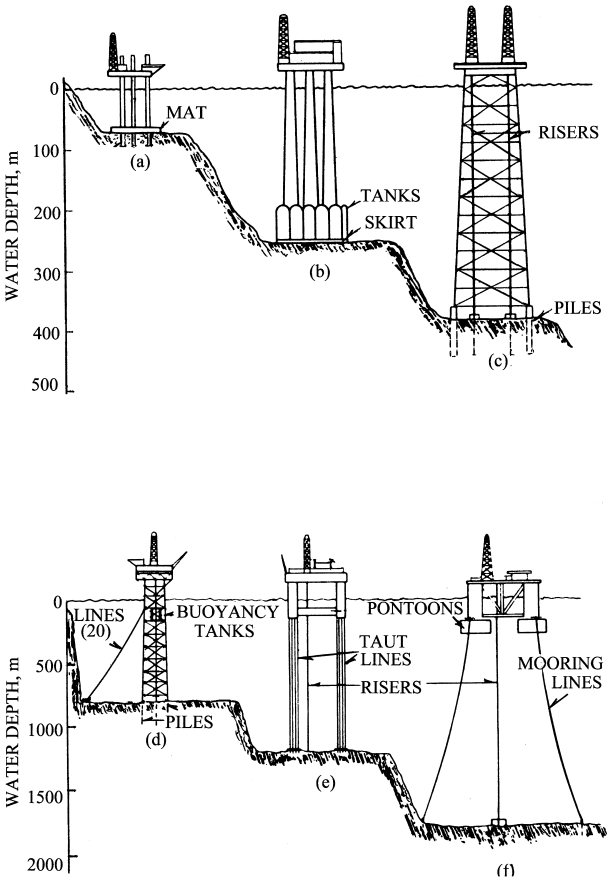


FIGURE 1 Water depth of six typical offshore platforms. (a) Jackup rig; (b) gravity platform; (c) jacket structure; (d) compliant tower; (e) tension leg platform; and (f) semisubmersible.

When pipelines were first laid offshore, no extraordinary analyses or deployment techniques were needed since they were in shallow water and were of small diameter, somewhat flexible, and made of relatively ductile steel. As platforms were built in deeper and deeper water with multiple well slots, larger diameter pipelines of higher strength were required. During the 1960s, engineers met this challenge with new designs with refined methods of analysis and deployment. Pipeline systems evolved into two main types: seafloor and vertical configurations. Both are used to transport gas and oil, but the vertical systems also include risers to carry drilling tools, electric power lines, dredge pipes for deep sea mining, and cold water pipes (CWP) for ocean thermal energy conversion (OTEC).

Throughout the world there are at present about 80,000 km of marine pipelines. Since 1986, the rate of building new marine pipelines has been about 1000 km per year. Individual pipelines on the sea floor vary in length from 1 to 1000 km and in diameter from 7 to 152 cm. For instance, a Norwegian project features a 1000 km line extending from the Troll field to Belgium, which

was completed in 1992. Kuwait has the loading line of largest diameter, 152 cm. The pipelines of smaller diameter are used to transport oil and gas from wellheads. Those of larger diameter are used to load and unload oil from tankers moored at offshore terminals. The deepest sea floor pipelines at present are the 46 cm diameter gas lines in the Gulf of Mexico, for which the maximum depth is 1400 m. Sea-floor pipelines are often anchored to the seabed or buried in trenches for protection from erosion and the undermining effects of currents. Some seabed pipelines have a coating of concrete to add protection and to reduce buoyancy.

II. PLATFORMS

Six typical offshore platforms are depicted in Fig. 1. Examples of the most common offshore platforms and their chief design characteristics are now reviewed.

A. Fixed-Bottom Structures

A mobile structure often used for exploratory oil-drilling operations is the self-elevating platform commonly called a jackup or mat-supported rig [see Fig. 1(a)]. Typically, a platform is supported by three to six legs that are attached to a steel mat resting on the sea floor. In soft soils, the legs pass through the mat and may penetrate the soil to depths of up to 70 m. To the bottom of each leg is attached a steel saucer or "spud can" to help stabilize the structure and minimize leg penetration into the soil. The height of the platform above the seafloor can be adjusted by using motor drives attached to each leg. Jackup rigs operate at water depths of up to 100 m.

A platform designed to be used in a fixed location as a production unit is shown in Fig. 1(b). Such a unit, called a gravity platform, consists of a cluster of concrete oil-storage tanks surrounding hollow, tapered concrete legs that extend above the water line to support a steel deck. A typical unit, of which there were 28 operating in the North Sea in 1999, has one to four legs and rests directly on a concrete mat on the sea floor. With ballast consisting of sand in the bottom of the tanks and seawater in the legs, these structures depend on self-weight alone to maintain an upright position when subjected to the highest waves that are expected to occur in a 100-year period. A realistic 100-year wave that may occur in the northern North Sea is 27.8 m. At present, the largest concrete gravity platform is the Troll structure, and one of moderate size is the Statfjord-A Condeep structure, both located in the North Sea. The latter structure is 250 m high and has three legs. Located off the coast of Norway, the Statford-A Condeep unit has slots for 42 oil wells that reach to depths of 2800 m. When in operation, it accommodates a crew of 200 people who live and work on this structure.

Found more frequently among the permanent, fixed-bottom structures, however, is the steel truss or jacket template structure shown in Fig. 1(c). As for the gravity platform, each steel jacket unit is designed for a fixed location and a fixed water depth. The first such structure was operational in 1955 in water 30 m deep. By 1996 there were approximately 6500 jacket structures, the tallest of which was the Bullwinkle unit located in the Gulf of Mexico. The common characteristics of these jacket structures are their tubular legs with tubular braces in K or X patterns: legs that are somewhat inclined to the vertical. Piles driven through these legs into the sea floor and clusters of piles around some of the legs maintain structural stability in adverse weather. One of the largest jacket structures is the 380 m high Cognac unit, which has 10 legs with 24 piles extending 140 m into the soft clay soil of the Gulf of Mexico.

B. Compliant Towers

An alternative class of offshore structures meant for depths from 300 to 800 m is the compliant tower such as that shown in Fig. 1(d). Such a tower may or may not have mooring lines. It is a pile-supported steel truss structure designed to comply or flex with the waves and has considerably less structural material per unit height when compared with a common jacket template tower.

The first compliant tower was Lena, which was installed in the early 1980s in the Gulf of Mexico. Including its three-level drilling and production deck and its drilling rigs, this tower reaches a total height of 400 m. Each of the 20 stabilizing cables, attached 25 m below the water line and arranged symmetrically about the structure, extends a horizontal distance of about 1000 m to a line of clumped weights that rest on the sea floor, to an anchor cable and an anchor pile. Under normal weather or small storm conditions, the cables acts as hard springs, but with severe storms or hurricanes, the cable restraints become softer or compliant. That is, the amplitude of tower rotation increases at a rate greater than that of the loading because, the clumped weights lift off the sea floor to accommodate the increased storm loads on the tower. When storms or hurricane conditions are anticipated, operations on compliant towers cease and the crew is evacuated. Installation of the Lena cables was more difficult and costly than anticipated. Since Lena, compliant towers without cables have been designed by Exxon. Two of these were installed during 1999 in the Gulf of Mexico.

Figure 2 shows that the lowest or natural frequency of oscillation in sway for a typical compliant tower during storm conditions is about 0.2 rad/s, whereas the tallest

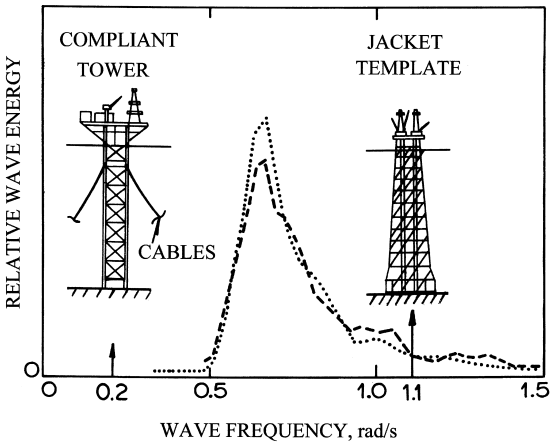


FIGURE 2 Two storm wave height energy spectra measured in the Gulf of Mexico and the fundamental sway frequencies for two offshore platforms located there.

jacket structures have natural frequencies near 1.1 rad/s. The important point is that offshore structures are designed to have frequencies well removed from the frequency range of the highest energy waves for storm seas (0.6–0.9 rad/s), as illustrated in Fig. 2. If the fundamental sway frequency of a structure coincides with that of the highest energy waves, structural resonance, high amplitude motion, and subsequent structural failure may occur. It is important to note that the storm wave height energy spectra shown in Fig. 2 are highly site-dependent.

C. Buoyant Structures

The tension leg platform (TLP) can be economically competitive with the compliant tower for water depths between 300 and 1200 m. In this structure [Fig. 1(e)], the buoyant force of the submerged pontoons exceeds the structure's total gravity or deadweight loading. Taut, vertical tethers extending from the columns and moored to the foundation templates on the ocean floor keep the structure in position during all weather conditions. The heave, pitch, and roll motion are well restrained by the tethers; but the motions in the horizontal plane, or surge, sway, and yaw, are quite compliant with the motion of the waves. There are at present a total of three tension leg platforms. The first production TLP was built 150 km off the coast of Scotland in the mid-1980s. Conoco installed the Julliet in 1989, and Saga Petroleum installed the Snorre near Norway in 1991. The tethers for the Snorre are 137 cm in diameter. By the late 1990s, a total of eleven TLPs were installed, three in the North Sea and eight in the Gulf of Mexico.

Another popular buoyant structure is the floating production system. Such structures are practical for water depths greater than 1200 m. but also at lesser depths where the field life of the structure is relatively short. An example of a buoyant structure is the semisubmersible with fully submerged hulls [Fig. 1(f)]. Other examples include ships converted to floating production systems. In the late 1960s, companies initiated research and design for these semisubmersible, multihull tubular structures and ships that would remain relatively stable in rough seas. In the late 1990s, the first three deep draft caisson vessels, or Spars, were installed for use in 180 m water depths. Spars are floating vertical cylinders that support production decks above storm waves. These structures are controlled to remain essentially still in stormy seas. A few of these structures are now operational. Some need to be towed from place to place; others are self-propelled. During drilling and production operations, these structures are kept in place with mooring lines and thrusters. The computer-controlled thrusters monitor the mooring line forces and accurately position the structure over the wellhead. One of the first semisubmersible structures was the Sedco 709 with a water depth rating of 1800 m. The most recent semisubmersibles using dynamic positioning are rated for 3000 m water depths.

A subsea production system provides an excellent alternative to a floating or a fixed surface facility. Much of a subsea system rests on the ocean floor, and its production of oil and gas is controlled by computer from a ship or other buoyant structure above the subsea unit. The buoyant structure and the subsea unit are often connected by a marine riser, which is discussed presently.

III. MOORINGS

A. Temporary Anchor Moorings

A classical example of temporary offshore mooring is the spread mooring configuration for a ship in relatively shallow water. Six to eight cables of wire rope or chain are unreeled from onboard winches symmetrically placed around the perimeter of the ship. Each cable hangs as a catenary curve and is attached either directly to a drag embedment anchor in the seabed or to a buoy that is anchored. Tug boats aid in spread mooring installations.

In a typical spread mooring operation for a semisubmersible in deep water, a work vessel transports each anchor while pulling out its cable attached to the semisubmersible. The vessel lowers the anchor and installs a locating surface buoy just above it.

At present, temporary systems for semisubmersible drilling rigs and construction barges are used in water depths of up to 2000 m. An example of an early and successful drilling rig is the Ocean Victory installed at a water depth of 450 m. This rig employed 12 anchors, each with a

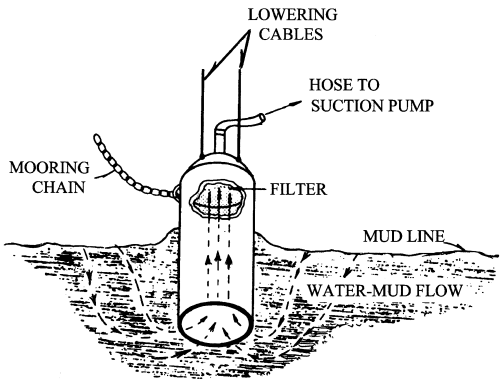


FIGURE 3 Typical suction anchor showing water flow during embedment.

holding power of 200,000 newton (N) or about 45,000 lb. Each catenary line was about 2500 m long and consisted of two equal segments: one of 8.9 cm diameter steel wire cable and the other of 8.3 cm diameter chain. The new generation rigs designed for deep water have ten times the anchor holding capacity of the Ocean Victory.

New anchor configurations to give more holding power are continually being developed. One of the more innovative designs is based on suction (Fig. 3). Versions of this anchor design have been successfully applied to temporary mooring operations in the North Sea.

B. Permanent Platform Pile Moorings

For installing piles for platform moorings, specially fitted derrick barges may use hydraulic hammers, drilling equipment, or possibly a jetting system. In jetting, seawater is forced around the base of a pile, blasting away the soil to make way for pile embedment. When installing mooring piles for tension leg platforms, template structures are carefully positioned at the site, and piles are hammered through the template that serves as a pile guide.

Suction piles are employed in deep water where the use of hydraulic hammers is impractical or too costly. Suction piles use hydrostatic pressure to push the piles to full penetration.

C. Single-Point Moorings

Single-point mooring (SPM) systems are designed to accommodate deep-draft tankers while they transfer crude oil and fuel oil to and from shore. Three typical designs are shown in Fig. 4: a single anchor leg mooring (SALM), a catenary anchor leg mooring (CALM), and an articulated column system. At present there are about 50 SALM and 150 CALM systems in operation throughout the world. The common features of these two systems are the rotating

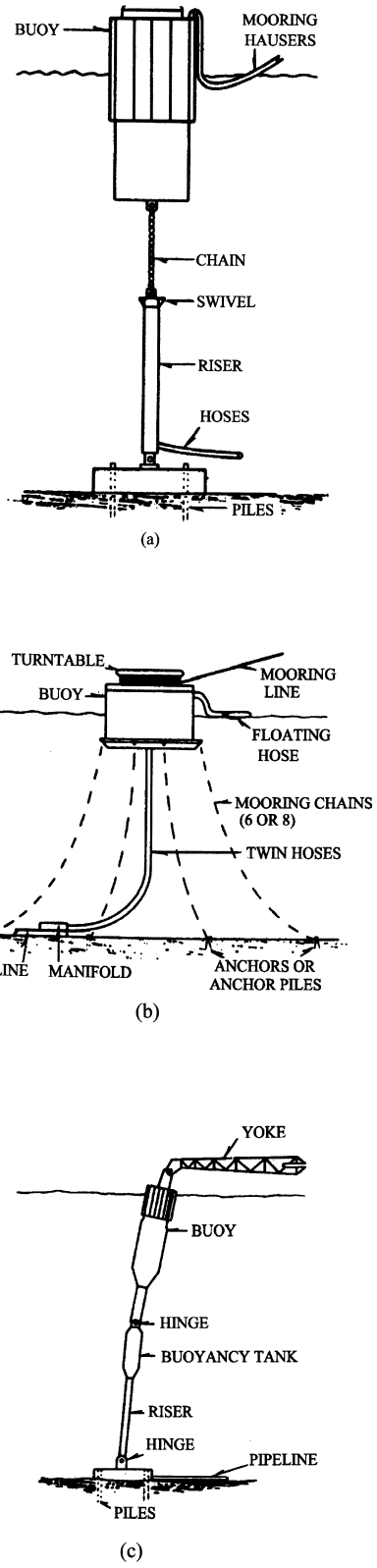


FIGURE 4 Three typical single-point mooring (SPM) designs. (a) Single anchor leg mooring (SALM). (b) Catenary anchor leg mooring (CALM). (c) Articulated single point mooring.

head on the buoy and the vertical chain that anchors the buoy to the seafloor. While a few SALM systems may have taut mooring lines for added buoy stability, all CALM systems have multiple catenary lines (or anchor legs), as the name implies. Several of the third type of SPMs, the articulated columns, have been designed and laboratory-tested but have yet to be installed offshore.

Successful SPMs are found at the world's largest terminals. One of these is Saudi Arabia's Ju'aymah exporting terminal, which has two SALM buoys and four CALM systems. For this SALM system, crude oil and fuel oil are loaded simultaneously to a moored tanker through the swivel assembly on the seabed. A second example is the CALM system for service vessels associated with the Cognac platform in the Gulf of Mexico. This SPM buoy has 12 catenary lines, each anchored to 0.76 m diameter piles. The water depth here is 275 m.

IV. PIPELINES

A. Sea-Floor Pipelines

Most of the 80,000 km of offshore pipelines have been installed by one of the following three methods. In these methods, the deployed pipeline forms an S-shape between the vessel and the sea floor.

1. *Laybarge method.* Pipe sections, which sometimes have been coated previously with concrete for protection, are welded together on the deck of a barge and deployed on rollers over the stern. Near the stern, the pipeline passes over pontoons called stingers that relieve excess bending of the pipe as it is deployed to the seabed (Fig. 5).
2. *Reel barge method.* Small-to medium-diameter pipe sections (up to 41 cm in diameter) are prewelded and coiled onto a reel mounted to the deployment vessel. As the pipeline is unreeled at sea, it passes through straightening rollers and then deployed as in the laybarge method.
3. *Bottom pull method.* Pipe sections are assembled on shore, and the pipe string is towed into the sea by a barge. During the launch to its place on the seabed, pontoons are often used under the string to avoid excess bending.

The following two methods of pipeline deployment have been studied extensively but have yet to be used.

1. *J-lay method.* A dynamically positioned vessel such as a drill ship, a converted pipelaying vessel, or a semisubmersible may be used in this operation. On board the vessel, a derrick is used to hold the pipeline vertically as it is lowered, and the pipeline forms a J-shape between the vessel and the sea floor. An efficient single station pipe welding procedure has to be developed before this method can be adapted to common practice.
2. *Floating string method.* According to this concept, the pipeline is floated on pontoons at the water surface. Then the pontoons are released successively so that the pipe string gradually sinks to the sea floor.

B. Marine Risers

A marine riser is a vertical offshore pipeline. Although marine risers make up a fraction of the 80,000 km network, they are nonetheless key offshore structures that serve a variety of functions. An example is a large vertical pipeline that contains a bundle of smaller lines connecting a wellhead to its platform: or a pipeline that transports oil from a platform to a sea-floor pipeline. A riser may act as a drilling sleeve, or it may contain electric power lines for operating sea-floor mining vehicles or other subsea facilities. A typical drilling riser, with a ball joint at the bottom and a telescoping joint at the top, is maintained under tension to ensure stability and is kept to within 8 degrees of the vertical by computer-controlled positioning of its parent drilling vessel. In the mid-1980s, the longest riser was operating off the east coast of the United States, where depths of 2100 m were reached. At present, risers for depths approaching 3000 m are becoming a reality.

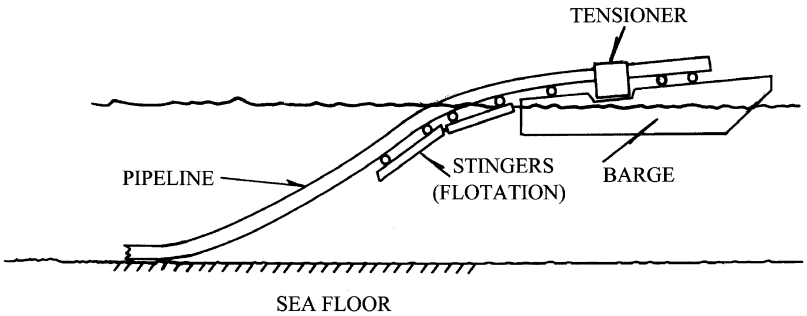


FIGURE 5 The laybarge method of pipline construction offshore.

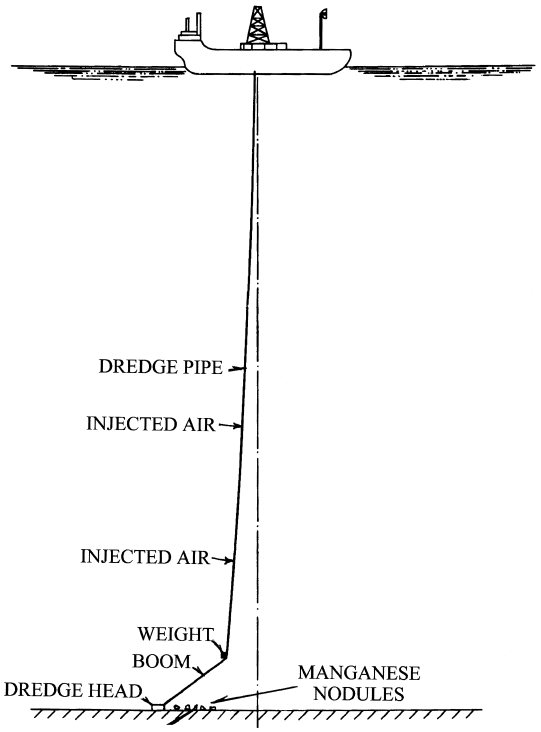


FIGURE 6 A dredge pipe "vacuum cleaner" for mining manganese nodules from the ocean floor.

Vertical pipelines are used in deep-sea dredging operations. A most challenging engineering problem, which began to receive serious attention in the late 1960s, was that of designing dredge pipes (Fig. 6) to suck up and transport manganese nodules from depths of 3000 to 5500 m to the ship. These nodules, about the size of a man's first and found in a monolayer on the sea floor in many parts of the world, contain other minerals as well, including cobalt, copper, and nickel with traces of molybdenum, vanadium, and titanium. By the mid-1980s, remote-controlled mining of the nodules was achieved with a 5500 m long dredge pipe. In this design, air bubbles are pumped into the string at various points along its length to aid in pumping the nodules to the surface. The dredge pipe's position is computer-controlled, synchronized to the movement of both the self-propelled mining scoop at one end and the ship at the other end.

Vertical cold water pipe (CWP) systems for ocean thermal energy conversion (OTEC) have been studied extensively. Calculations show that practical full-scale CWP designs are 10–20 m in diameter, are about 1000 m long, and may be used to raise the cooler water in the deep ocean to the warmer surface water. With the differential temperature between the ends of the CWP (10–20°C), it is possible to produce net power through heat exchange.

In 1979, a small-scale OTEC power plant operated successfully on a barge 2.3 km offshore of Kona, Hawaii. This CWP was 61 cm in diameter, 660 m long, and made of polyurethane. The CWP was supported by a surface buoy at the barge stern and was tension-moored with a smaller diameter polyurethane pipeline, a sea-floor wire rope to shore, and concrete blocks on the sloping sea floor. This was the world's first and so far the only OTEC system at sea to generate net useful power.

V. CHALLENGES

To meet the need for new sources of energy and minerals, marine engineers must work at the frontiers of known technology. Their main challenges are to design, deploy, and operate facilities and equipment in environments where none have before existed—in deeper and deeper water, on the slopes of the continental shelves, and in the hostile Arctic seas. To meet the challenge of innovative design, research and development in the following four broad and necessarily overlapping areas are presently underway.

A. Environmental Forces

Every offshore facility is subjected to several types of environmental loads during its lifetime. Thus, site-dependent databases are being developed to characterize the time-varying fluid-induced loads of winds, currents, and waves. Such loads occur both on a short-time scale (periodic vortex shedding, wind gusts, and wave slamming) and over periods of hours, days, and even years (steady waves, tides, and hurricanes). At some sites, data are also needed on subsea earthquake intensity, on the scouring of sea-floor foundations by currents, and on the total and differential settlements of the sea floor due to the withdrawal of hydrocarbons during the lifetime of the structure. At present, there is the particular challenge of minimizing differential settlements for gravity platforms. For the Ekofisk structural complex in the North Sea, however, this challenge appears to have been met by employing water injection. In Arctic regions, data are needed on the rates of ice accretion and on the velocity, yield strength, and mass of floating ice. Both deterministic and statistical methods are being used to measure and interpret time-dependent field data suitable for predicting structural loadings.

B. Structural Materials

An offshore structure should have a high ratio of strength to self-weight. For instance, for each added unit of deck weight for a tension leg platform, an additional 1.3 units of hull weight is required for buoyancy support, and an additional 0.65 units of mooring pretension force is needed. In this case, high-strength steel with high-fracture

toughness may soon be used to reduce the hull weight requirement, and hollow cylindrical steel link chains or synthetic mooring line materials, such as Kevlar with an abrasion-resistant polyethylene cover, may be used to increase the mooring capacity. Design data are currently being generated to determine the suitability of new, high-strength steels and composite materials in the offshore environment. These data involve measures of corrosion fatigue, fracture toughness, stress corrosion cracking, and weldability for high-strength steels and the reliability of several types of high-strength, light-weight synthetic rope mooring lines.

C. Modeling and Analysis

Once the site, the environmental load conditions, and the preliminary structural design are determined, a mathematical model of the structure is formulated, and a computer-aided analysis is performed to determine the overall motion, the critical stresses, and the reliability of the design. For tubular structures, fatigue failures have occurred at the joints. It is thus particularly important to design for joint fatigue and to include fatigue in the structural reliability analysis. Generally, the design criteria are not met after the first analysis, so a redesign is called for and the analysis repeated. Computer graphics and finite element methods of structural analysis that aid in this design process are continually being improved. Structural optimization, least-weight criteria, nonlinear dynamics, fluid–structure–soil dynamic interactions, and statistical methods are being added and refined in the mathematical models. The aims are to achieve more accurate representations for structural behavior while the structures are being deployed and in operation. Also, improvements in the predictions of structural lifetimes are sought.

D. Experimental Evaluations

Before a structure is installed offshore, extensive tests are made on its component materials and also on its overall dynamic behavior using scaled-down laboratory models with simulated environmental load conditions. Present laboratory-scale efforts are focused on two issues: fluid-structural load interactions, especially the effects of fluid vortices on structural motion; and the mechanics of structural–soil foundation interactions.

Once installed in the field, the structure may be instrumented for a period of time to determine whether its behavior under a variety of natural forces has been accurately predicted by the mathematical models and analyses used in its design. Such data are useful for future similar designs. In addition, for platforms, buoys, and pipeline systems nearing the end of their lifetimes, frequent inspection and maintenance are required to assure continuing performance. High-capacity microprocessors and computer systems are used for the retrieval, storage, and analysis of large amounts of data from both laboratory scale and field tests of offshore structural systems.

SEE ALSO THE FOLLOWING ARTICLES

CORROSION • DAMS, DIKES, AND LEVEES • FOUNDATIONS • GEOENVIRONMENTAL ENGINEERING • MECHANICS OF STRUCTURES • NAVAL ARCHITECTURE • PETROLEUM GEOLOGY

BIBLIOGRAPHY

Adrezin, R., Bar-Avi, P., and Benaroya, H. (1996). *J. Aerospace Engine*. October, pp. 114–131.

Gaythwaite, J. (1981). "The Marine Environment and Structural Design," Van Nostrand-Reinhold. New York.

Gerwick, B. C. (1999). "Construction of Marine and Offshore Structures," 2nd ed., CRC Press, Boca Raton, FL.

Graff, W. J. (1981). "Introduction to Offshore Structures," Gulf Publishing. Houston, TX.

Proceedings of the Offshore Technology Conference. Houston, Texas, 1970 to present.

Sarpkaya, T., and Isaacson, M. (1981). "Mechanics of Wave Forces on Offshore Structures," Van Nostrand-Reinhold, New York.

Wilson, J. F. (ed.) (1984). "Dynamics of Offshore Structures," Wiley (Interscience), New York.

Wilson, J. F., and Orgill, G. (1984). *ASME Journal of Dynamic Systems, Measurement, and Control* **106,** December, pp. 311–318.



Pneumatic Transport

George E. Klinzing
University of Pittsburgh

I. Introduction
II. Basic Analytic Formulation
III. Phase Classifications
IV. Flow Measurements
V. Electrostatic Effects

GLOSSARY

Choking Unstable regime experienced in vertical transport that occurs when the transport gas velocity is insufficient to overcome the gravitational forces of the particles.

Dense phase Pneumatic transport flow conditions having high solids loading with considerable solid–solid and solid–wall interactions. Generally, the voidage values are lower than 0.95 for dense phase flow.

Dilute phase Conditions of flow where the solids concentration is low, giving a voidage close to 1.0. The main energy loss in such a system is still dominated by the transport gas velocity. Oftentimes the loading ratios are in the range of 1.0.

Particle velocity Steady-state velocity achieved by the particles when transported by the gas. The particle density and diameter affects the final particle velocity.

Phase diagram Variety of diagrams that present the various flow conditions possible in pneumatic transport systems. Operability and unstable regimes are generally designated in the phase diagrams. One possible diagram relates the pressure drop to the transport gas velocity at constant solids flow rates.

Saltation Condition in a horizontal transfer line that permits the solids to deposit on the bottom of the pipe.

Slip velocity Difference between the transport gas velocity and the particle velocity.

Solid loading Ratio of solids flow rate to transport gas flow rates. This parameter is sometimes used to give a rough estimate of the flow conditions.

Solids pressure drop Pressure drop in a pneumatic conveying process that can be attributed to the solids interactions with themselves and the wall.

Transport gas velocity Gas velocity used to convey the solid material through the pipeline.

Voidage Void volume occupied by the gas in a gas–solid system. The voidage determines the regime of flow of the pneumatic transport system.

IN MANY PROCESSES the movement of solids between units is essential. Pneumatic transport of these solids by a gas stream in an enclosed conduit is an efficient, economical method to deliver these solids to the right place at the right time. Pneumatic transport involves three basic steps: (1) the delivery of the particles to be conveyed for

contacting with the transport gas, (2) the conveying of the solids in the pipeline, and (3) the collection of the particles at the end of the transport section. Fine particles of less than 1 μm to large particles in the range of 0.1 m can be conveyed pneumatically and these transported distances can vary from a few meters to a few kilometers.

I. INTRODUCTION

Pneumatic transport can be traced back to ancient times; however, the first large-scale use of this gas–solid system has been the transport of materials for cement over a distance of 2 miles in a 12-in. line for the construction of Grand Coolie Dam. Since that time, a multitude of materials has been moved pneumatically. Generally, the pipelines vary in size from 1 to 14 in. with conveyed distances of 10–2000 ft being commonplace. Diameter sizes of particles transported range from micrometers to a few inches.

Some advantages of pneumatic conveying are its complete enclosure of the material transported, adaptation to a wide variety of materials, lower initial cost, ease in automation, less maintenance, and savings on bulk shipments.

On the negative side, however, power costs have been cited as a detriment to the use of pneumatic conveying. These costs are usually associated with the air-moving equipment. New advances in this area have reduced these costs significantly. Besides the high power cost, to operate a pneumatic conveying system, higher technology is required.

To carry out pneumatic transport, a gas supply is necessary as well as a solid feeder and collection devices at the end of transfer. Gas supply can be provided by a fan, a blower, or a compressor. The solids feeders are many and varied in nature. Some of these are the rotary feeder, the screw feeder, the pressure tank, venturi feeder, and fluidized bed feeder.

Pneumatic transport can be classified as either vacuum delivery or pressure delivery, or by the phase of transport, such as dilute or dense. A vacuum system picks up solids at several points and delivers them to one site. For pressure conveying, pickup is made at a single point and delivered to multiple points.

Combinations of the two systems are also possible, giving multiple pickups and multiple deliveries. For dilute and dense designations, these correspond roughly to the amounts of solids delivered per volume of gas. No precise definition exists for these phase designators.

Transport lines can be arranged in many geometric configurations. Horizontal, vertical, and inclined lines are the usual designations.

II. BASIC ANALYTIC FORMULATION

To analyze pneumatic transport in an analytic manner, the forces acting on the system must be described in detail. For unsteady operation, these forces are balanced by the acceleration terms of the system. The basic forces acting on the gas–solid flow system are the drag forces, the pressure force, gravity forces, frictional forces, and electrostatic forces, if present. One can write this force balance as (Klinzing *et al.*, 1997):

$$\Delta m\, dU/dt = F_{\text{drag}} - pressure \cdot area - F_{\text{gravity}} - F_{\text{frictional}} - F_{\text{electrostatics}}, \tag{1}$$

where Δm is differential mass, U velocity, and t time. This expression applies to both the gas phase and solid phase, with a different formulation for each phase.

For the solid phase the expression is written as:

$$(1-\varepsilon)\rho_{\text{p}}\frac{dU_{\text{p}}}{dt} = \frac{3}{4}C_{\text{DM}}\rho_{\text{f}}\frac{(U_{\text{f}}-U_{\text{p}})^2}{(\rho_{\text{p}}-\rho_{\text{f}})D_{\text{p}}}(1-\varepsilon)\rho_{\text{p}} - \frac{dP}{dx}(1-\varepsilon) - g(1-\varepsilon)\rho_{\text{p}} - 2f_{\text{s}}\frac{U_{\text{p}}^2(1-\varepsilon)\rho_{\text{p}}}{D} \tag{2}$$

and for the fluid phase as:

$$\varepsilon\rho_{\text{f}}\frac{dU_{\text{f}}}{dt} = -\frac{3}{4}C_{\text{DM}}\frac{\rho_{\text{f}}(U_{\text{f}}-U_{\text{p}})^2}{(\rho_{\text{p}}-\rho_{\text{f}})D_{\text{p}}}(1-\varepsilon)\rho_{\text{p}} - \frac{dP}{dx}\varepsilon - g\varepsilon\rho_{\text{f}} - \frac{2f_{\text{g}}U_{\text{f}}^2\varepsilon\rho_{\text{f}}}{D}. \tag{3}$$

In these equations ε is voidage, ρ_{p} density of the solid, U_{p} particle velocity. C_{DM} drag coefficient of a multiple particle system, ρ_{f} density of the fluid, U_{f} fluid velocity, D_{p} diameter of the particle, D diameter of the pipe, P pressure, X axial distance, g gravitational constant, f_{s} solid frictional term, and f_{g} gas frictional term. Note that the drag term in each phase is the same value but of opposite sign. For steady state the left-hand side of the equations are equal to zero. Combining the two expressions, the drag terms cancel each other producing:

$$-\frac{dP}{dx} = g(1-\varepsilon)\rho_{\text{p}} + g\varepsilon\rho_{\text{f}} + \frac{2f_{\text{s}}U_{\text{p}}^2(1-\varepsilon)\rho_{\text{p}}}{D} + \frac{2f_{\text{g}}U_{\text{f}}^2\varepsilon\rho_{\text{f}}}{D}. \tag{4}$$

This is the functional equation for the design of pneumatic transport systems. The expression can be written in shorthand notation as:

$$\Delta P_{\text{total}} = \Delta P_{\text{gravity}} + \Delta P_{\text{frictional}}. \tag{5}$$

Some specific observations are essential in order to evaluate Eqs. (2)–(5). In most cases the gas gravitational contribution is small in comparison to the solid gravitational force. Thus this term can often be neglected. The drag coefficient in the above expressions is based on a multiple-particle system and can be expressed by:

$$C_{\mathrm{DM}} = \varepsilon^{-4.7} C_{\mathrm{D}}. \tag{6}$$

The drag coefficient of the single particle system, C_{D}, can be a variety of regimes depending on the slip-velocity-based Reynolds number:

$$\mathrm{Re}_{\mathrm{p}} = D_{\mathrm{p}}[(U_{\mathrm{f}} - U_{\mathrm{p}})/\mu_{\mathrm{f}}]\rho_{\mathrm{f}}, \tag{7}$$

where μ_{f} is fluid viscosity. The voidage term, ε, arises also in these basic equations. The relationship to other system parameters is given as:

$$\varepsilon = 1 - (W_{\mathrm{s}}/A\rho_{\mathrm{p}}U_{\mathrm{p}}), \tag{8}$$

where W_{s} is solid mass flow rate.

For the frictional representations the gas and solid friction factor must be defined (i.e., f_{g} and f_{s}). For the gas friction factor one can use a variety of expressions; one such convenient form is that given by Koo:

$$f_{\mathrm{g}} = 0.0014 + 0.125\mathrm{Re}^{-0.32}, \tag{9}$$

where the Reynolds number is that based on the tube dimension. To assess the solid friction factor, one finds oneself in an evolving field. The detailed gas–solid flow structure should be known in order to use the appropriate solid frictional term. The regions of dilute and dense surface have significant effects in determining the solid frictional representation. A few expressions for this f_{s} term are seen in Table I. In regard to the solid frictional term, note that the particle velocity squared is the dynamic term for this expression. Recent considerations have suggested that slip velocity $(U_{\mathrm{f}} - U_{\mathrm{p}})^2$ is the appropriate term to more accurately describe the frictional action of the system.

The effect of different pipe wall materials has been probed by Rizk (1973a) and others. This effect, however, appears to be minor and even of little consequence when a fine powder coats the conveying line wall. For inclined flow Klinzing *et al.* (1989) has given a frictional representation that covers all kinds of inclined flows from horizontal to vertical.

$$f_{\mathrm{s}} = \frac{D}{2u_{\mathrm{p}}^2}\left[\frac{3}{4}\left(C_{\mathrm{D}}\varepsilon^{-4.7}\rho_{\mathrm{g}}\{u_{\mathrm{f}} - u_{\mathrm{p}}\}^2\right)/(\rho_{\mathrm{p}} - \rho_{\mathrm{g}})d\right] - g\sin\theta, \tag{10}$$

where θ is the angle of inclination.

TABLE I Solids Friction Factors

Stemerding
$f_{\mathrm{s}} = 0.003$

Reddy and Pei
$f_{\mathrm{s}} = 0.046\mu_{\mathrm{p}}^{-1}$

Van Swaaij, Buurman, and van Breugel
$f_{\mathrm{s}} = 0.08\mu_{\mathrm{p}}^{-1}$

Capes and Nakamura
$f_{\mathrm{s}} = 0.048\mu_{\mathrm{p}}^{-1}$

Konno and Saito
$f_{\mathrm{s}} = 0.0285\,(gD)^{0.5}u_{\mathrm{p}}^{-1}$

Yang, vertical flow
$f_{\mathrm{s}} = 0.00315\dfrac{1-\varepsilon}{\varepsilon^3}\left(\dfrac{(1-\varepsilon)u_t}{u_{\mathrm{f}} - u_{\mathrm{p}}}\right)^{-0.979}$

Yang, horizontal flow
$f_{\mathrm{s}} = 0.0293\dfrac{1-\varepsilon}{\varepsilon^3}\left(\dfrac{(1-\varepsilon)u_{\mathrm{f}}}{(gD)^{0.5}}\right)^{-1.15}$

Stegmaier
$f_{\mathrm{s}} = 0.52\mu^{-0.3}\mathrm{Fr}^{-1}\mathrm{Fr}^{*0.25}(d/D)^{-0.1}$

Michaelides
$f_{\mathrm{s}} = \frac{K}{4}\frac{\sqrt{gD}}{u_{\mathrm{p}}}$
where K depends on the type of particle conveyed being equal to 0.058 for coal.

One difficulty in applying this theory is determining the particle velocity, since rarely do systems have the facility to measure this value. Some empirical expressions can be used to address this issue. Table II has a listing of particle velocity expressions that can be used in the design of these systems.

The condition of unsteady state was first addressed in the basic dynamic equations governing gas–solid transport. This condition of unsteadiness arises more often than one would suspect in the design of a transfer line. The initial injection of particles into the system is, of course, an unsteady process for accelerating the particles to a steady-flow condition. Every bend or change in flow direction of a system causes accelerations and decelerations. For long pipelines a pressure drop in the system causes the gas velocity to increase and thus continuous changing or unsteadiness of these variables. Equation (2) can be considered for the unsteady particle behavior. One can substitute for $d(t)$ the expression (dL/U_{p}) and integrate Eq. (2) to obtain:

TABLE II Particle Velocities

$D_{\mathrm{p}} < 40\ \mu\mathrm{m}$	$U_{\mathrm{p}} = U_{\mathrm{g}} - U_{\mathrm{t}}$
Hinkle modified *IGT*	$U_{\mathrm{p}} = U_{\mathrm{g}}\left(1 - 0.68D_{\mathrm{p}}^{0.92}\rho_{\mathrm{p}}^{0.5}\rho_{\mathrm{f}}^{-0.2}D_{\mathrm{t}}^{-0.54}\right)$
Yang	$U_{\mathrm{p}} = U_{\mathrm{f}} - U_{\mathrm{t}}\sqrt{\left(1 + \frac{2f_{\mathrm{s}}U_{\mathrm{p}}^2}{gD}\right)\varepsilon^{-4.7}}$

$$L = \int_{U_{p1}}^{U_{p2}} U_p dU_p \Bigg/ \left[\frac{3}{4} C_{DM} \frac{\rho_f (U_f - U_p)^2}{(\rho_p - \rho_f) D_p} - \frac{dP}{dx} \frac{1}{\rho_p} - g - \frac{2 f_s U_p^2}{D} \right]. \quad (11)$$

Equation (11) can give the acceleration length necessary to recover the steadiness of the particulate system. The pressure loss over this transient length L_a can be given as:

$$\Delta P_{accel} = \int_0^{L_a} \rho_p (1-\varepsilon) g \, dL + \int_0^{L_a} \frac{2 f_g \rho_f U_f^2 \, dL}{D} + \int_0^{L_a} \frac{2 f_s \rho_p (1-\varepsilon) U_p^2}{D} dL + \rho_p (1-\varepsilon) U_p^2 \quad \text{at } L_a. \quad (12)$$

Some concern exists as to the appropriateness of the frictional representation for the unsteady regimes. This, however, has not been fully addressed to date.

A combination of the basic Newton analyses and empiricism has proven useful in developing the acceleration length given as:

$$L_{accel} = 27.66 (d/D)^{0.953} (\mu)^{-0.0912} (\rho/\rho_p)^{-0.924}. \quad (13)$$

Limits exist in pneumatic conveying for both the horizontal and vertical transfer legs that must be addressed before actual design is carried out. There is a minimum velocity for both geometries that will permit the solids to flow. The saltation velocity for horizontal flow is when the particles begin to drop out of the horizontal flow and settle on the pipe bottom. If this condition continues to exist, the line will plug.

In practice, the term *pickup velocity* is often used in the design of horizontal pipelines. This is defined as the velocity to pick up the solids and convey them. There is some discrepancy in the literature as to the difference between the terms *saltation velocity* and *pickup velocity*.

Saltation is a phenomenon that occurs in horizontal flow in having particles fall out of the suspension, which is conveying the material, and deposit on the bottom surface of the pipe. Many studies have been explored to determine the saltation condition since it is a parameter essential in designing conveying systems. The most commonly used correlation is that of Rizk (1973b) as analyzed by Jones and Leung, which was developed for plastic pellets and does not include any effect for particle density or pipe size.

$$\mu = \frac{1}{10^{\delta}} \left(\frac{u_g}{\sqrt{gd}} \right)^{\chi}, \quad (14)$$

where $\delta = 1.44\,d + 1.96$ and $\chi = 1.1d + 2.5$. Matsumoto *et al.* (1975) have a useful correlation for saltation that breaks the analysis into two parts, one for larger particles and one for smaller ones. Cabrejos and Klinzing (1994) have determined the saltation velocity as:

$$\frac{u_s}{\sqrt{gD}} = \frac{u_{s0}}{\sqrt{gD}} + 0.0022 \left[\frac{\rho_s}{\rho} \right]^{1.25} \mu^{0.5}. \quad (15)$$

The saltation velocity has been used often as the design parameter for the velocity at the feed point for conveying. The pickup velocity has also been employed for this same purpose.

The pickup velocity in pneumatic conveying has two different definitions. One is the velocity required to pick up particles from the bottom of the pipe where they are at rest because of a flow condition or blockage. The other definition is the velocity at the feed point of the solids. Mostly in this discussion attention will be paid to the former, although the latter will be mentioned sometimes. A rule of thumb is that the relationship indicates the saltation velocity is roughly half the pickup velocity. Thus, designing on the pickup velocity for the conveying velocity is more conservative. In studying the pickup velocity of particles, one finds that as the particle size decreases from values of 1000 μm, the pickup velocity will decrease until a minimum point is reached where the pickup velocity begins to rise with further size decrease. The smaller particle has larger surface and interparticle forces acting on them. For larger particles, recent work by Cabrejos and Klinzing gives what appears to provide reliable pickup velocities.

$$\frac{u_{gpu}}{\sqrt{gD}} = (0.9428 \mathrm{Re}_p) \left(\frac{D}{d} \right)^{0.25} \left(\frac{\rho}{\rho_s} \right)^{0.75}, \quad (16)$$

$$\mathrm{Re}_p = d u_g \rho / \eta,$$

$$25 < \mathrm{Re}_p < 5000,$$

$$8 < D/d < 1340,$$

$$700 < \rho/\rho_s < 4240.$$

A phase diagram showing the interconnectiveness of these two parameters was also developed and is shown in Fig. 1.

For the vertical transport, the choking velocity has a similar connotation as the saltation velocity does in horizontal flow. The choking velocity condition is a very unstable condition in the flow where the pressure drop fluctuates widely. Yang (1985) has considered the data present in the literature and has suggested the following format for the choking velocity. Generally, designers like to stay away from the choking velocity and will grossly overdesign a system toward higher velocities that are much less efficient in transporting solids.

Bends were mentioned as regions of deceleration and acceleration in the pipeline. The previous equations could be applied to these regions. The effect of solid particles on bend losses has been studied by few investigators. One

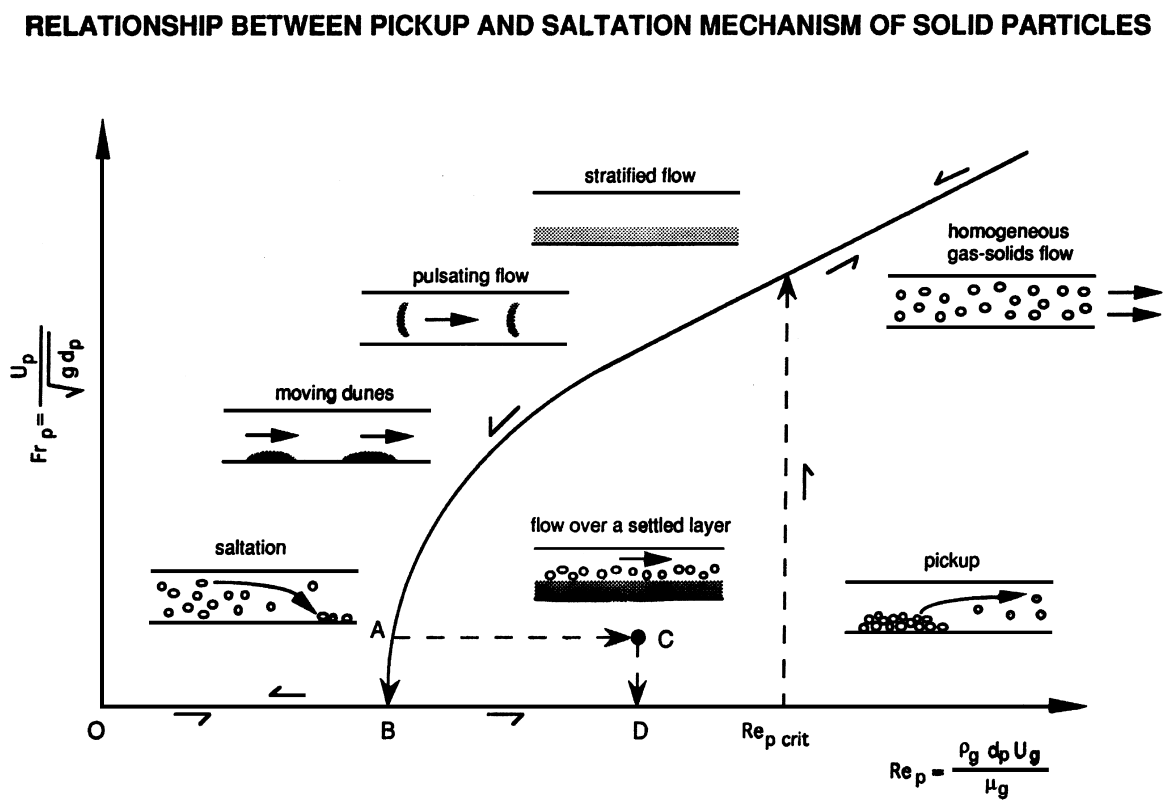


FIGURE 1 Relationship between pickup and saltation mechanism of solid particles.

such correlation has been presented by Schuchart. The pressure loss due to the solids in the bend is given as a ratio to that for a straight pipe section.

$$\left(\frac{\Delta P_{\text{bend}}}{\Delta P_{\text{straighten}}}\right)_{\text{solids}} = 210\left(\frac{2R_{\text{B}}}{D}\right)^{-1.15}. \quad (17)$$

This expression is for a variety of bend configurations where R_{B} is the bend radius. To calculate the total pressure loss in a bend the gas contribution must be added to the solid contribution. It has developed a reliable pressure loss expression for single-phase pressure losses around bends. This equation for the single-phase loss is:

$$\Delta P_{\text{single fluid phase}} = \frac{0.029 + 0.304\left[\text{Re}(r_0/R_{\text{B}})^2\right]^{-0.25}}{(R_{\text{B}}r_0)^{1/2}} \frac{L\rho_{\text{f}}\bar{U}^2}{2D}, \quad (18)$$

where r_0 is the radius of the pipe.

The above equations for bend pressure losses are usually applied to long radius bends. More recent data have shown that the simple T-bend is much preferred from an erosion standpoint and also from pressure losses. These T-bends have sudden direction changes and do not prolong the acceleration and deceleration regions. They have the ability to pack the flowing solid in the blinded T section, and any wear takes place by particles against particles instead of particles against the wall. In some cases it would be desirable to use crosses instead of T-bends, since cleaning out the system with a cross is easily done in both directions. In general, should one employ more than three T-bends in proximity in series, the pneumatic transport flow will be jeopardized because this involves too many decelerations and accelerations for the particles and the lines will most likely plug. One should make sure that there is a reasonable distance (5 ft or more) between bends. The fewer the number of bends, the more reliable the pneumatic transport system will operate.

When one is dealing with pulse-piston dense phase flow, the use of T-bends is not recommended for cohesive materials. In fact, for very cohesive materials with high moisture contents the use of T-bends is not advised.

A. Dense Phase Flows

Some basic plug flow studies have been carried out by Legel and Schwedes and Lippert. Klinzing and coworkers have explored several coal plug transport situations employing the Konrad analysis for their works. For vertical flow Gu and Klinzing have shown that the maximum plug length decreases as the pipe diameter increases. For

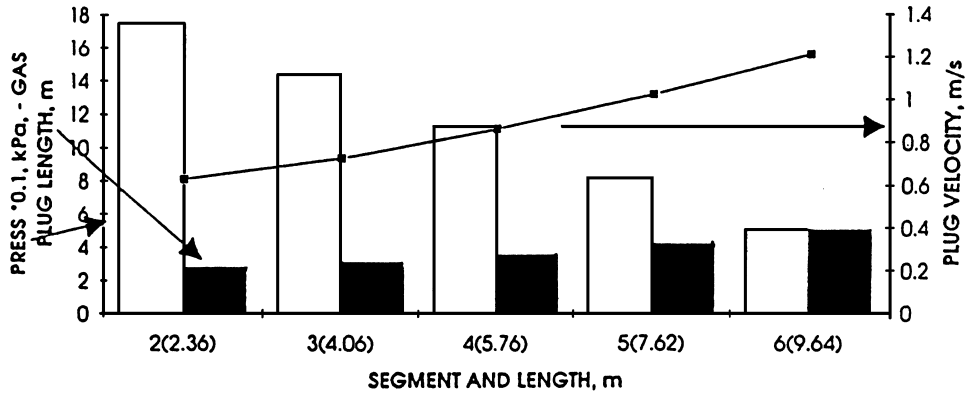


FIGURE 2 Gas and solid plug behavior: pressure, lengths, and velocities.

horizontal flow Aziz and Klinzing showed also that a linear variation of pressure drop with plug length existed for plugs up to 0.3 m in length. The minimum plug length was also determined indicating that the faster moving plugs require a longer minimum plug length before breakthrough of the plug occurs. It is essential to have a plug of length greater than minimum in order to have efficient plug flow to occur.

An analysis of the plug breakthrough and the behavior of a system containing several plugs has been carried out by Dhodapkar, Plasynski, and Klinzing. Considering the following parameters the behavior of a plug flow system is depicted in Fig. 2. These conditions apply:

$D = 0.05$ m
$\rho_B -$ bulk density $= 1000$ kg/m^3
Pressure drop across a single plug $= 23$ Pa
Internal cohesion coefficient $= 29.5$ Pa
Wall cohesion coefficient $= 24.6$ Pa
Angle of internal friction $= 38^\circ$
Angle of wall friction $= 30^\circ$
Wall shear stress $= 59$ Pa
Shear stress $= 36.9$ Pa
Plug velocity $= 1$ m/sec
Injection sequence $= 10$ sec

One notes the increase in plug velocity and the increase in the air gap length as the plug traverses the system. A recent thesis by Mi showed similar behavior for plastic pellets and wheat when an entire plug flow system was studied. Along the same lines Destoop has developed a sizing technique for discontinuous dense phase conveying.

Molerus has considered dense phase flow, which is a two-phase condition where a dilute phase transport flows over a dense, slower moving solids flow. The analysis is able to distinguish between the dilute and the heavy load flow conditions.

B. Pressure Fluctuations

Pressure fluctuations are natural in all pneumatic conveying systems. The very nature of the system with the large difference in densities of the components and the turbulent character of the flow produces pressure oscillations. Fluctuations can also be generated from the air supply, feeder, and collector systems. Over and above these equipment-induced fluctuations, the inherent fluctuations of the flow are a source of great information in analyzing the condition of the flow. Dhodapkar and Klinzing (1993a) and Cabrejos and Klinzing (1994) have performed a wide variety of analyses on these pressure fluctuations and their ability to recognize different flow regimes. Each flow pattern has a unique flow pattern and, as such, can be analyzed and stored as a library of information. In an analysis done by Cabrejos and Klinzing, a rescaling technique was employed on the pressure fluctuations. This technique develops the dimensionless ratio of the range of pressure fluctuations divided by the standard deviation of the fluctuations to form:

$$\frac{\text{Range}}{\text{Std. Deviation}} = (a\tau)^{\text{H}}, \tag{19}$$

where τ is the period of the record. In this expression the exponent varies with the condition of the flow.

C. Computer Applications

To design a dilute phase pneumatic conveying system, one needs to have a mechanism to select and calculate a number of factors and components. The first question is what type of system would be appropriate for the conveying operation? Should one use a vacuum or pressure system? Is dilute phase the mode most appropriate for transport? Using an expert system approach, our laboratory has developed a program entitled NUSELECT (Pneumatic Conveying Consultants) to help answer these types of questions. After selecting the system, one needs to choose the

appropriate feeder, injecting the material into the pipeline for transport. Again, we have addressed this question by developing an expert system program entitled FEEDER for this task. Calculations presented above should provide the basis for predicting the energy needed for transport to help complete the design. Most often a cyclone or reverse jet filter is used for collecting the product after transport. When a system is in operation, more than likely there will be occasions when the system is not functioning properly and troubleshooting the operation is needed. The expert system PANACEA provides the basis to begin troubleshooting before an expert in the field is called in for analysis (Dhodapkar and Klinzing, 1993b).

III. PHASE CLASSIFICATIONS

In Section I the phase terms "dilute" and "dense" were used as flow classifications. More detailed phase diagram behavior is necessary in order to understand pneumatic transport. Before addressing the phase diagrams, consideration of the character of the solids and their classification is beneficial to the practitioner of pneumatic transport. For this purpose the Geldart diagram generally used in fluidization will be presented. Figure 3 shows the four regions of particle classification according to their size, density, and ability to be fluidized. Group A is termed aeratable powder, which is easily fluidized; group B is a sandlike material, which also can be fluidized but has a larger particle size than group A. Group C powder is a fine cohesive powder that has considerable interparticle forces present and needs extraordinary means to cause fluidizations. Group D consists of larger sized particles that form a spoutable classification that has not been in one-to-one correspondence with pneumatic transport. Lohrmann has attempted to investigate a wide variety of materials and associate their group properties to blow-vessel behavior, while Canning and Thompson considered the classification in relation to

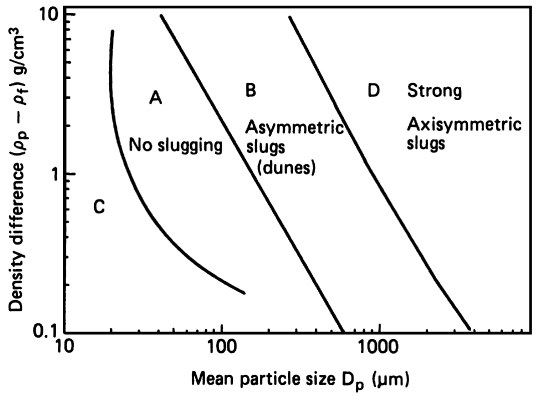


FIGURE 3 Geldart diagram for flow classification.

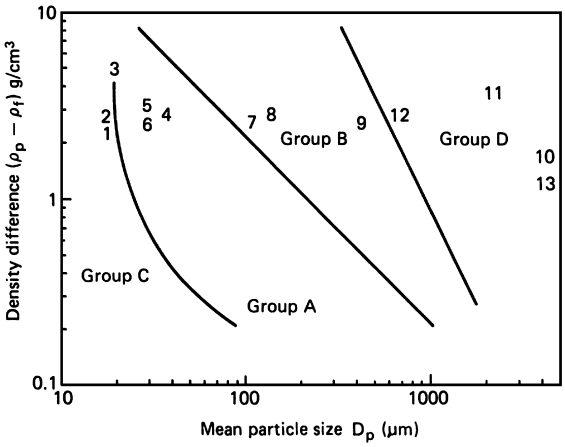


FIGURE 4 Lohrmann's study of different particles on a Geldart diagram. Code: (1) pulverized fly ash, (2) ground calcine mineral, (3) manganous oxide, (4) unslaked lime, (5) cement, (6) ferrosilicon baghouse dust, (7) sodium sulfate, (8) electric arc furnace dust, (9) foundry sand, (10) maize grits, (11) magnesite, (12) milled asbestos waste, and (13) ice.

dense-phase flow. Figure 4 is a diagram of the materials studied by Lohrmann and their relationship to the Geldart diagram. Group A powders are the best candidates for dense-phase conveying. Group B powder was found to be suitable for pulsed phase conveying, while not suitable for dense-phase conveying. Group D powders with a low density form stable plugs, while denser materials and large particles cause vibrations. These powders can be conveyed in a dense phase provided that there is a narrow size distribution. Group C powder can be troublesome in pneumatic conveying, just as they are in fluidization; generally dense phase is not recommended for this powder.

Dense flow has a number of different definitions. When we consider the high loading of solids, the term *dense phase* is often applied. In many pulse-piston flow systems, dense phase is also employed. The pulse-piston flow generally extrudes a plug of material from a feeder and then transports it with a slug of air from the air supply. This is operated in an alternating fashion at relatively low velocities (1–3 m/sec). In this fashion it is possible to convey very friable materials and cut down on the erosion seen at high-velocity conveying. It is possible to convey many materials with the pulse-piston arrangement except for sandlike materials that do not form plugs that have integrity. In the pulse-piston mode of operation, the energy loss is only due to the pressure loss across the individual plugs of solids because the gas plug energy loss is insignificant.

Zenz has proposed a convenient phase diagram behavior with plots of pressure losses versus the gas velocity at a diagram of this type. The minimum point in the pressure loss is noteworthy. Ideally this would be a convenient place to design for a system. To the left of the minimum is

a sharply increasing segment of the diagram. The region is unstable and oftentimes designers want to stay far from this region and suggest operations in the region far to the right of the minimum point. The crossover between dense and dilute phase operations is sometimes designated by the minimum point. The unstable region to the left of the minimum point can represent the region of choking and slugging operation. If the energy is sufficient a more continuous densely packed flow can take over. If the energy available is not sufficient, the flow will plug and stop.

In the field of phase diagrams Matsen has suggested another avenue of pursuit. This diagram has an analogy in thermodynamics in the vapor-liquid equilibrium behavior. Figure 5 shows one of these diagrams with the dilute-dense phases noted. The transition between the curves is termed the choking locus. This diagram is more specific in its designation of the terms dilute and dense phase. It is essential to have a knowledge of the voidage term in order to construct these phase diagrams.

Tuba, Mathur, and Klinzing have used the basic premise of Matsen and generalized and reduced the phase diagram in terms of a flux format. The reduced fluxes are given as:

$$j_g^* = j_g\sqrt{\frac{\rho_f}{gD(\rho_p-\rho_f)}},$$
$$j_p^* = j_p\sqrt{\frac{\rho_p}{gD(\rho_p-\rho_f)}}, \tag{20}$$

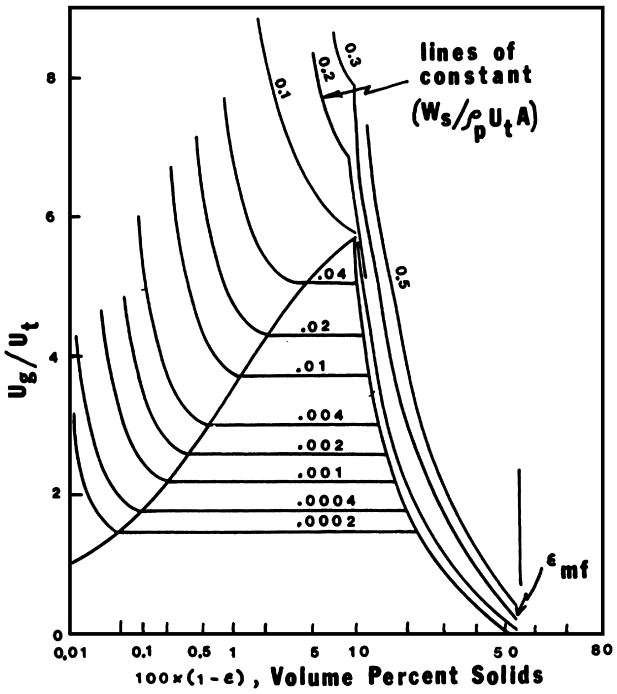


FIGURE 5 Matsen-type phase diagram.

where j_g is volumetric gas flow per area and j_p volumetric solid flow per area. On analysis of the phase relationship and similarity with thermodynamics an equation of state similar to the van der Waals equation has been suggested. Thus,

$$j_p^* = \frac{R'}{(1-B\phi)}j_g^* - A\phi^2, \tag{21}$$

where $\phi = (1-\varepsilon)$. Recently experimental data in our laboratory permitted the horizontal flow to be studied where the top of the pipe indicated dilute phase or vapor-like flow, with the lower section represented by dense phase or liquid-like flow. By a flow splitter and visual observations we obtained the phase diagram shown in Fig. 6, which relates rather well with the proposed Matsen diagram given in Fig. 5.

Work in the modeling area has suggested that a combination of the Matsen-type diagram with the Zenz diagram is able to show the phase behavior and pressure loss at the same time. Figure 7 shows the initial attempt at this synthesis.

IV. FLOW MEASUREMENTS

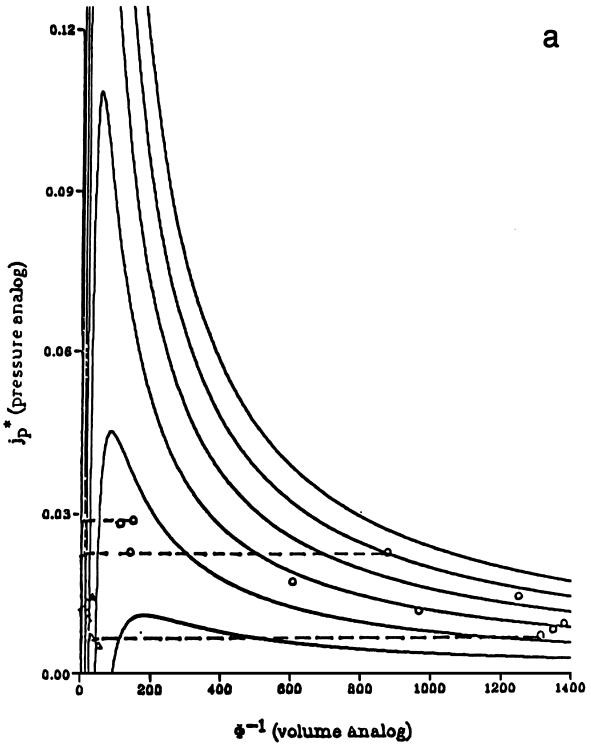
One of the most challenging tasks in pneumatic transport is the ability to be able to measure the flow of solids at any particular time. These relatively instant measurements will permit the use of automatic control in order to regulate solid–gas flow in much the same manner as we control single-phase, gas- and liquid-flow systems.

The simplest forms of flow meters follow the same instruments used in normal flow metering, namely, the orifice and venturi meters. A combination of the two in series has been utilized to measure solid flow. The orifice responds to the gas flow and the venturi responds to the solids. Work on a single venturi meter as a solid-flow measuring device has been renitiated by Crowe and Davies. These meters have been tested only on systems of low values of solid loading. Erosion by the solid particles presents itself as a concern in the orifice and venturi meters.

The detection means defines the basis of several other flow measuring devices. Light, sonic, electrical charge, and dielectric variations are the basis of a number of flow meters.

The principle of Doppler shift due to the solids present in a gas flow has been applied both on a sonic and light basis. These techniques are able to give the particle velocities very accurately. For this Doppler shift principle only dilute phase systems are applicable.

A microphone's ability to pick up the vibrations generated by solids flowing in a pipe has been used by Argonne



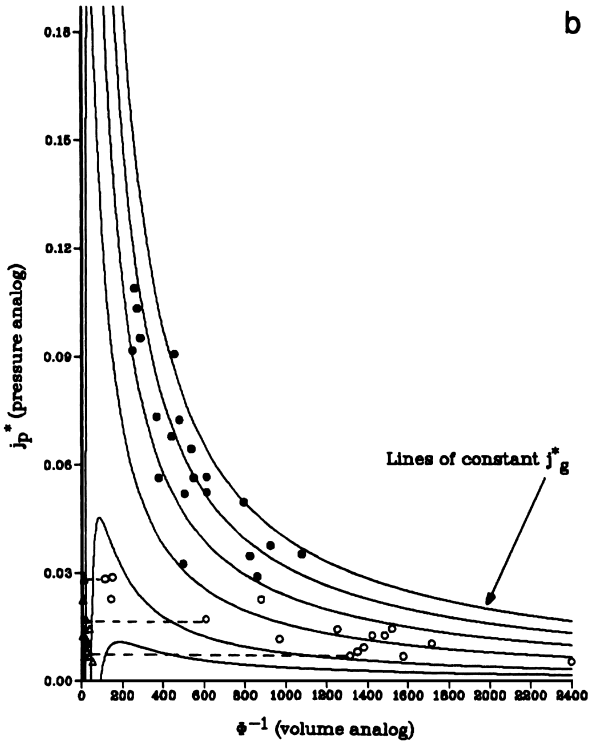


FIGURE 6 j_p^* vs. ϕ^{-1} for 450-μm glass beads in a 0.0504-m-diameter pipe showing the two phases. (a) ○ Flow above strand (vapor-like); △ strand flow (liquid-like). (b) ● Homogeneous (gaslike) phase; ○ flow above strand (vapor-like); △ strand flow (liquid-like).

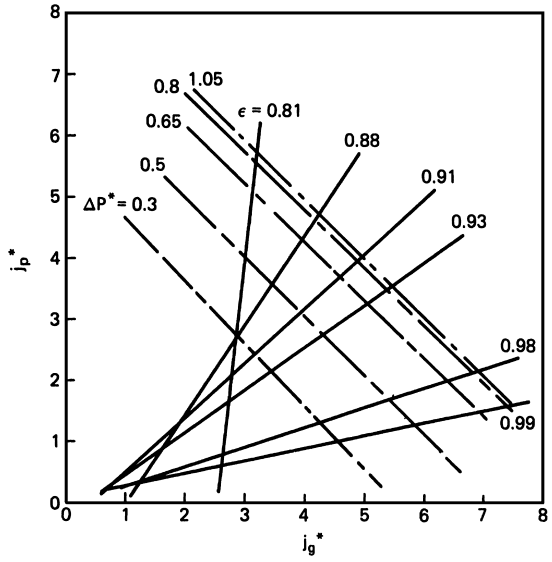


FIGURE 7 Phase diagram with pressure drop representation.

National Laboratories to determine the flow of solids in a line.

A light absorption meter using a halogen lamp source and a coupled photodiode op/amp was used by Klinzing to follow solid loading in a coal-gas transport system. The extinction of light through the solids flow follows Beer's logarithmic absorption law.

Soo and his colleagues have proposed a number of solid-flow meters based on the electrostatic and glow discharge principles. There is a variety of geometric configurations for these meters. Those inserted into the flow stream will also suffer from erosion limitations. King also has explored the analysis of electrostatics in gas–solid flows. He also used the generated charges in a cross-correlation manner to determine the particle velocities. More recently Smith and Klinzing have used modern electronic and computer technology to automate the cross-correlation technique.

The dielectric property can be conveniently measured by the Auburn international meter. This unit has the same diameter as the transport pipe and, as such, suffers little erosion as well as not affecting the flow pattern of the solids. Mathur and Klinzing have used this as a direct-flow meter and a cross-correlation technique for the particle velocity.

The Granucor flow-metering device (Endress and Hauser) is a recent commercial unit operating on the conductivity principle. Both particle velocities and flow rates can be determined with this unit.

A novel use of physics in a solid-flow meter has been utilized by Micromotion based on a Coriolis force generated in flow around a 180° bend. Mathur has tested this

unit on coal-gas flow and found it to be adequate for solids flow at moderate and high loadings. Erosion in such a unit has not been addressed but could be sizable in use with abrasive solids.

A differential pressure measuring device has been developed by Cabrejos and Klinzing (1992) to measure the solids mass flow rate. The specific pressure drop defined as the ratio of the differential pressure drop across a straight section of pipe for dilute phase pneumatic conveying divided by the pressure drop due to the flowing gas alone is found to be linearly related to the solids loading, mass of solids/mass of gas flowing. One can express this relationship as:

$$\alpha = \frac{\Delta P_{\text{total}}}{\Delta P_{\text{gas}}} = 1 + K\mu = 1 + \frac{\lambda_z}{\lambda_L}\frac{c}{V}\mu. \qquad (22)$$

From this expression knowing the pressure drops gives one the loading and from this the solids flow rate at a known gas flow rate. This unit has been developed into an automatic measuring device.

New advances in the area of solid meters are most welcome and needed. We seem to be limited only by our imagination in this field.

V. ELECTROSTATIC EFFECTS

Should one be transferring solid material pneumatically under conditions at which the relative humidity of the transport gas is high and the conveying lines are metals that are well grounded, the electrostatic effects in pneumatic transport may never occur. In ungrounded systems with low relative humidities for the transport gas, electrostatic forces can be generated in a pneumatic transport system that can affect the overall energy requirement of the system. As the particle size decreases, the size of the charging increases. Above about 400 μm, little electrostatic charging is seen. For electrostatics to be generated, rapid contact and breakage of contact between dissimilar materials is essential in triboelectrification. Charging on particles can increase until the breakdown voltage of air is exceeded. If a discharge occurs and the solid concentration and oxygen content of the gas are sufficient, a powder or dust explosion can occur. The discharge potential for air is 3000 kV/m.

The charge per unit mass on particles can be given by:

$$Q/m_p < 9 \times 10^6 \varepsilon_0 / r_p \rho_p, \qquad (23)$$

where ε_0 is permittivity of free space and r_p radius of the particle. Once particles charge they can retain this charge for a relatively long period of time, depending on the resistivity of the powder. The time constant for decay of charge is given as:

$$T = \bar{\varepsilon}\varepsilon_0 R. \qquad (24)$$

In pneumatic transport Masuda *et al.* have developed a charge expression to account for contact times of the particles with the containing wall and showed that:

$$Q = C\Delta V[1 - \exp(\Delta t/\tau)], \qquad (25)$$

where C is capacitance, ΔV potential difference, τ relaxation time, and Δt contact time. Charge transferred on contact of particles with a wall is most conveniently measured by current. Knowing the rate of collisions of the particle and the wall one can write:

$$i = QN_1 \qquad (26)$$

where i is current and N_1 collision rate. The collision rate is proportional to the solid-flow rate.

Soo was first to realize the interrelationship between particle mechanics, heat transfer, and electrostatic charge transfer. Masuda and coworkers expanded this approach, considering both elastic and inelastic collisions between particles and particles with the confining wall. The main purpose of these exploits was to determine the contact area at collision. For the current generated on an insulated pipe wall of length L the contact area is given as:

$$I = W_s\frac{\Delta Q}{m_p} = -\frac{6W_s\varepsilon_0 V_c A\Delta t}{\pi D_p^3 \rho_p Z_{0\tau}}\frac{\Delta n}{\Delta x}L, \qquad (27)$$

where ΔQ is charge increment on particle by impaction, V_c contact potential, A contact area, Z_0 gap distance, $\Delta n/\Delta x$ change in number of collision with length, and L length of pipe.

Masuda experimentally tested Eq. (27) by measuring the current generated on an isolated section of metal pipe. Figure 8 shows the agreement he achieved in his analysis. The current shows a quadratic behavior with mass flow rate. It is noteworthy that the sign of the current changes at a certain solid-flow rate.

In a comprehensive study of current generated by electrostatics in pneumatic conveying Masuda *et al.* found the current to be proportional to the diameter of the tube and inversely proportional to the particle diameter, while the velocity had an almost square dependency. A triboelectric series was suggested on the materials they studied as (listed in decreasing order):

Flour
Morundum
Glass
PVC
Steel

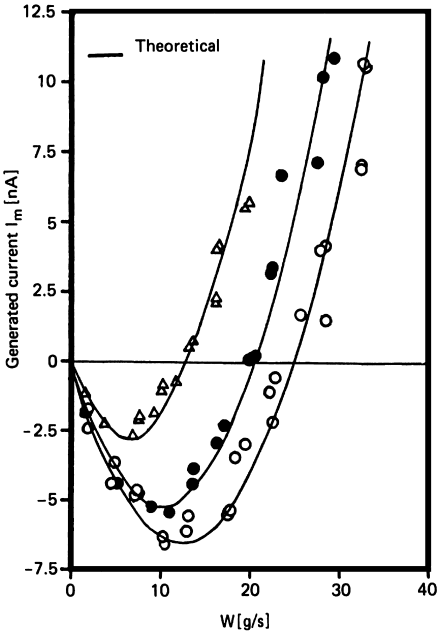


FIGURE 8 Current from pipe bend to ground; (△) 10 m/sec, (●) 20 m/sec, and (○) 30 m/sec.

Quartz sand
Teflon.

Smeltzer, Weaver, and Klinzing have proposed a film-penetration model borrowed from heat and mass transfer to explain charge transfer between particles and the wall. The basic expression for the current is given as:

$$i = h_e A \Delta V. \tag{28}$$

The term h_e is called an electrostatic transfer coefficient and can be developed for various contact time distributions. For the gamma distribution one obtains:

$$h_e = \frac{\sigma \beta^{1/2} \Gamma(\alpha + 1/2)}{(\sigma m_p / c\rho_p)^{1/2} \pi^{1/2} \Gamma(\alpha + 1)}, \tag{29}$$

where σ is conductivity, α and β coefficients for the gamma distribution, c capacitance, and m_p mass of particle.

An additional pressure drop loss can be added to the normal pressure loss when the electrostatic effects are present. Basically this can be expressed as:

$$\Delta P_{\text{electrostatic}} = E_x (Q/m_p)(1 - \varepsilon), \tag{30}$$

where E_x is the potential gradient. The potential gradient is difficult to assess in most situations; however, the charge on the particle can be found from capture in a Faraday cage.

The particle velocity is also affected by the static charges on the particles. Using the force balance at steady state, the particle velocity can be related to the electric field and charging as:

$$U_p = U_f - U_t \left[\frac{2 f_s U_p^2}{gD} + \frac{E_x Q}{g m_p} \right]^{1/2} \varepsilon^{2.35} \tag{31}$$

where U_t is terminal velocity of the particle. Ally has developed the pressure loss due to electrostatics from the use of a force balance and the application of Gauss's law. His development gives:

$$\Delta P_{\text{electro}} = \frac{45}{16} \frac{(1 - \varepsilon)^2 Q^2 D}{\pi^2 \varepsilon_0 D_p^6}. \tag{32}$$

In addition he found that the electrostatic pressure drop falls off in a logarithmic fashion with the moisture content of the system.

When one begins to review the way in which charging occurs and recalls the earlier concepts of electron transfer between different materials, one sees that charging is unique to almost each situation considered. The charge is related to the chemistry of the particle, the wall confining the flow (the work function), the diameter of the particle, the length of the transport line, the moisture content of the system, and the number of particles flowing in the system. Thus each system has a unique work function, depending on whether the system is a glass particle—plastic tube, metal particle, plastic tube, etc. The exact expressions for the dependence of each parameter on the charge are not entirely known. From recent data some estimates can be made for a number of the factors. The precise expression for the charge dependence on the particle size is not known but it appears that the smaller particles generate a larger charge in a flowing system. This result is shown from the number density considerations of the flow for two sizes of particles at the same loading conditions. In addition, one needs to consider that the maximum charge on a particle is proportional to the diameter squared of the particle. Thus one has, for example,

$$\frac{Q_{75\,\mu\text{m}}}{Q_{150\,\mu\text{m}}} \alpha \left(\frac{D_{\text{p150}\,\mu\text{m}}}{D_{\text{p75}\,\mu\text{m}}} \right)^3_{\text{number density}} \left(\frac{D_{\text{p75}\,\mu\text{m}}}{D_{\text{p150}\,\mu\text{m}}} \right)^2_{\text{maximum charge}} \alpha \left(\frac{D_{\text{p150}\,\mu\text{m}}}{D_{\text{p75}\,\mu\text{m}}} \right). \tag{33}$$

In analysis of data for the development of charge, it appears that the charge achieves an equilibrium value in the first few feet of travel length. Experimental evaluation of the effect of moisture on the generated charge indicates a rather profound dependency. It appears that this effect above all seems to dominate the particle charging. Exploring this further from the pressure drop generated due to the electrostatics effects one also finds a similar effect of the moisture on the pressure drop. It

has been noted that when the moisture ratio of kilograms water to kilograms solids exceeds 0.1, no electrostatic contributions are seen.

Each individual system will have a unique behavior of charging and resultant pressure drop increases. This effect, of course, has its basics in the work function of the system.

When one thinks of electrostatics, one often has an adverse connotation. Discharges into powders have been cited as the cause of many violent explosions. With the use of more electronic components in a variety of technologies electrostatic shielding and grounding is imperative. The use of electrostatics as a means to measure flow rates has been cited before. New developments in this area are imminent.

The use of pneumatic transport in long-distance conveying (greater than 1 km in distance) has developed in the mineral and energy industries. The most crucial item in such a system is the question of when one would step the pipe size to a larger value to reduce the gas velocity and thus the pressure loss and the wear and attrition on the particles. A number of investigators have considered this point. The ideal situation would be to have a pipeline in the shape of a horn, increasing in size as the length increases. This optimum shape has been considered, and the available pipe sizes have been fit into this shape to determine the stepping points using the principle of minimization of area between the ideal shape and stepped pipe. Employing this sequence, one finds for rather dilute phase systems that an equal distance stepping arrangement results. For the more dense phase systems, one encounters more nonlinearities in the analysis and thus unequal distance stepping is likely to result.

SEE ALSO THE FOLLOWING ARTICLES

DIELECTRIC GASES • ELECTRON TRANSFER REACTIONS • ELECTROSTATIC POWDER COATING • FLOW VISUALIZATION • PARTICLE SIZE ANALYSIS • THERMODYNAMICS

BIBLIOGRAPHY

Aziz, Z. B., and Klinzing, G. E. (1988). *Powder Tech.* **55,** 97–105.

Cabrejos, F. J., and Klinzing, G. E. (1992). *ASME 92-WA/F-3,* Anaheim, CA.

Cabrejos, F. J., and Klinzing, G. E. (1994). *Powder Tech.* **79,** 173–186.

Cabrejos, F. J., and Klinzing, G. E. (1995). *Powder Tech.* **84,** 157–163.

Canning, D. A., and Thompson, A. I. (1982). *Century II Conference, ASME Meeting,* San Francisco, CA, August.

Crowe, C. T. (1982). "Final Report on Contract DOE DE-FGF22-80PC30212."

Davies, G. S. (1981). *Proc. Powtech.* **81**.

Destoop, T. (1993). *Powder Handling and Processing,* **5**(2), 139–144.

Dhodapkar, S. V., and Klinzing, G. E. (1993a). *Powder Tech.* **74,** 179–195.

Dhodapkar, S. V., and Klinzing, G. E. (1993b). Chap. 21 *in* "Expert Systems in Solids Handling," (M. Roco, ed.), pp. 743–777, Butterworth-Heineman, Boston.

Dhodapkar, S. V., Plasynski, S. I., and Klinzing, G. E. (1994). *Powder Tech.* **81,** 3–7.

Gu, H., and Klinzing, G. E. (1989). *Powder Tech.* **57,** 59–67.

Institute of Gas Technology. (1978). Department of Energy, Contract FE 2286-32, October.

Jones, P. J., and Leung, L. S. (1978). *Proc. Pneumotransport* **4,** BHRA Fluid Engineering, Cranfield, paper D4.

Klinzing, G. E. (1981). "Gas–Solid Transport," McGraw-Hill, New York.

Klinzing, G. E., Rohatgi, N. D., Myler, C. A., *et al.* (1989). *Can. J. Chem. Eng.* **67,** 237–244.

Klinzing, G. E., Marcus, R. D., Rizk, F., *et al.* (1997). "Pneumatic Conveying of Solids," Chapman and Hall, London.

Konno, H., and Saito, S. J. (1969). *Chem. Eng. Japan* **2,** 211.

Legel, D., and Schwedes, J. (1984). *Bulk Solids Process* **4,** 53.

Lippert, A. (1965). Dissertation, University of Karlsruhe.

Lohrmann, P. M. M.Sc. thesis, University of Witwatersrand, Johannesburg, South Africa.

Masuda, H., Komatsu, T., Mitsui, N., and Iinoya, K. (1976/1977). *J. Electrost.* **2,** 341.

Mathur, M. P. (1982). *Proc. Symp. on Instrumentation and Control for Fossil Energy Processes,* Houston, TX, June.

Mathur, M. P., Klinzing, G. E., and Tuba, S. T. (1982). *Pneumatech* **I,** Stratford, England (May 1982).

Mathur, M. P., and Klinzing, G. E. (1983). *Fine Particle Soc. Meeting,* Honolulu, HI, August.

Matsen, J. M. (1981). *Proc. AIChE Meet., New Orleans* (preprint).

Matsumoto, S., Harada, S., Saito, S., *et al*. (1975). *Chem. Eng. Japan* **7,** 425.

Mi, B. (1994). Dissertation, University of Wollongong, New South Wales, Australia.

Michalides, E. (1987). *Trans. ASME* **109,** 171–178.

Molerus, O. (1981). *Chem. Eng. Sci.* **36,** 1977–1984.

Plasynski, S. I., Dhodapkar, S. V., Klinzing, G. E., *et al*. (1991). *AIChE Symp. Series* **87,** 78–90.

Rizk, F. (1973a). Dissertation, University of Karlsruhe.

Rizk, F., BHRA, Fluid Engr. (1973b). *Proc. Pneumotransport* **3,** BHRA Fluid Engineering, Cranfield, paper D4.

Smith, R., and Klinzing, G. E. (1989). *AIChE J.* **32,** 313.

Soo, S. L. (1967). "Fluid Dynamics of Multiphase Systems," Blaudell, Waltham, Massachusetts.

Yang, W. C. (1985). *Pneumatech 2 Proceedings*, Powder Advisory Center, London, England.



Railway Engineering

William W. Hay
University of Illinois

I. Key Considerations
II. Location
III. Motive Power
IV. Support Systems
V. Rail
VI. Turnouts and Crossings
VII. Track–Train Dynamics
VIII. Bridges
IX. Yards and Terminals
X. Signals and Communications

GLOSSARY

Ballast Coarse, granular material placed between the track and the subgrade to provide support, drainage, and bedding for the track structure.
Consist Makeup of a train, cars and contents, or of the locomotives pulling the train.
Container Large boxes to hold freight for movement over and transfer between rail, highway, and water.
Crossties Wood, concrete, or steel members transverse to the rails and providing attachment and support.
Gauge Distance between the rails measured $\frac{5}{8}$ in. below the gauge corner of the rails. In the United States the standard gauge is $56\frac{1}{2}$ in.
Gross tons Measure of train weight that includes the weight of cars and contents.
Heavy haul Movement of bulk commodities in high-capacity (usually 100 tons) cars and usually in unit trains.
Locomotive Self-propelled vehicle used for hauling cars and trains.
Motive power Means of moving cars or trains; locomotives.
Out of face Maintenance work performed over a continuous segment of track without interruption, in contrast to spot work where tasks are performed in separate locations.
Rail, continuous welded Rails in 39- or 78-ft or more lengths welded into one long, continuous length.
Rolling stock Railroad cars.
Tangent Straight track.
Ton mile One ton hauled one mile.
Tractive effort Pulling force exerted by a locomotive.
Truck Wheels and axles, bearings, side frames, and transverse bolster with center plate assembly by which each end of a car is supported.
Unit train Train of uniform consist dedicated to hauling a single commodity between specified points in

turnaround service; usually associated with heavy haul service.

RAILWAY ENGINEERING encompasses the location, design, construction, operation, and maintenance of railroad routes and facilities. The railroad infrastructure includes the right-of-way, track (rails and crossties), track support system (subgrade and ballast), yards, and terminal facilities. Railroad engineering is also concerned with motive power and rolling stock as these are related to route design and operation and, for electric railroads, power supply. Signals assist in the safe and expeditious movement of trains and the utilization of track capacity. Communications tie these together to make possible efficient management and administration of the entire enterprise. The application to railroads is only one of many aspects of bridge, structural, tunnel, and electrical and electronics engineering. These are disciplines in their own right and will be touched on only briefly in this article.

I. KEY CONSIDERATIONS

Railroad engineering, as presented here, is concerned with route location as defined by length of line, curvature, grades, and traffic sources and their relation to motive power and operating costs. It also concerns track and roadway and their relation to equipment design and train operation. The principles that govern location also determine how the route should be operated.

II. LOCATION

A. Economic Location

A railroad location is defined by its horizontal and vertical position. The economics of location derive from the relations between distance (length of line), curvature, grades, and motive power characteristics.

Any of several economic tests may be used to evaluate these relations, but the rate-of-return method has the advantages of traditional use and of indicating whether an attractive rate of return can be anticipated. Thus, the basic location formula is $r = (R - E)/C$, where R is the expected revenue, E is the operating cost, C is the construction cost (investment), and r is the rate of return. The rate of return must be sufficient to cover investment interest and to yield a profit in the 20–30% range.

For projects with a known and limited life, the annual cost of the discounted present worth tests can be used and, for public projects, the benefit–cost ratio. Whatever test is used, the revenues, operating expenses, and construction costs of alternatives must be determined.

B. Traffic (Revenue) Sources

A projected new line should be supported by a traffic and economic survey. Much construction has been for access to coal and iron deposits. The output of those mines establishes the traffic volume and revenue, subject to market fluctuations. For more general traffic, the number and character of traffic sources on line will be a factor. Adding circuitry to a route to reach a traffic source adds to distance and to the operating costs as discussed below. For an existing line the volume and revenue are known, but they can be affected by contemplated changes in service or routing, mergers, traffic agreements, rate structures, or changes in industrial patterns and geography.

C. Effects of Distance

Distance is the length of line. In the mid-1980s a mile of track costs $100,000 to $200,000 or more. Right-of-way (land) and grading costs vary with the terrain. The number of sidings, bridges, tunnels, grade crossings, yards, and buildings vary somewhat with distance and introduce costs that can only be established for a particular location.

The accepted unit of measure for distance is the cost per 1000 gross ton miles (1000 GTM). On tangent level track, the operating cost arises from overcoming train resistance, that is, the resistance encountered by the locomotive to the forward movement of the train and its contents. Distance costs also include fuel, wages of train and enginemen, and a proportionate cost of maintaining locomotives, rolling stock, track, signals, communications, and other facilities for traffic movement.

A cost per 1000 GTM adequate for the purpose is obtained by dividing total operating expenses by total 1000 GTMs. These data can be obtained from an existing line's accounts or borrowed from a similarly placed carrier. A total for the United States can be calculated with published data from the Federal Railroad Administration (FRA), the Interstate Commerce Commission, or the Association of American Railroads (AAR). The cost per 1000 GTM for the United States in 1978 was $10.40.

The gross costs of distance will be (gross tons/1000) $\times$ distance hauled $\times$ cost per 1000 GTM. The longer the length of line, the greater will be the effects of distance. The effects of curvature and of rise and fall can be included in distance by incorporating equivalent curve and rise-and-fall miles in the computation. This is explained later.

In comparing costs for alternative lines of different lengths, it must be noted that not all costs are attributed to the additional miles of one line over the other. The total cost for one line is computed as a base, but only the costs of the additional (or lesser) miles in the other alternative(s) are considered. Not all costs that entered into

the base line calculation will be incurred by the additional miles of the alternative. Yards, sidings, stations, bridges, and other structures occur only occasionally and, for only a mile or so of distance, may not occur at all. Thus, only a percentage of the base cost per mile would apply. That percentage is 28.9 for differentials of less than 1 mile, 34.8 for distances of 1–10 miles, and 35.0 for distances over 10 miles. If the differential distance affects the wages of train and enginemen, those percentages become, respectively, 34.1, 43.9, and 50.5.

D. Train Resistance

Train resistance is the resistance to forward movement of a train traveling at constant speed on tangent level track in still air. The locomotive must exert a tractive force or effort sufficient to overcome train and other resistances (defined later). The greater the train resistance, the more energy must be expended to move the train, giving a direct relation between train resistance and operating costs.

The industry has generally placed reliance on an equation developed by D. W. Davis, Jr., which, with appropriate coefficients derived from field tests, relates flexibility in speed, car weight, and car dimensions. Based primarily on earlier experiments by Dr. Edward C. Schmidt, it has had an accuracy suitable for most railroad purposes up to 50 miles per hour (mph).

The Davis equation has the form

$$R_u = 1.3 + 29/w + bV + CAV^2/wn,$$

where R_u is the unit resistance in pounds per ton of train or car weight; n is the number of car axles, usually four; w is the weight in tons per axle; V is the speed in miles per hour; A is the cross-sectional area of the car or locomotive in square feet (80–90 for cars, 80–120 for locomotives); b is a coefficient reflecting flange friction and shock, sway, and concussion, 0.045 for freight cars, 0.03 for passenger cars and locomotives; and C is the drag coefficient, which varies from 0.0005 for freight cars to 0.0024 for locomotives (0.0017 if streamlined) and 0.00034 for passenger cars. The value of C depends on the shape of the front end and the length and smoothness of the vehicle's exterior; A and C are sometimes combined in one coefficient.

Total car and train resistance is the product of the unit resistance and the weight of the car or train. It is significant that unit resistance decreases with increased car weight. The Davis equation is included in the American Railway Engineering Association's "Manual for Railway Engineering (Fixed Properties)."

To account for the characteristics of more recent car types and designs, the AREA manual also shows the equation

$$R_u = 0.6 + 20/W + 0.01V + KV^2/wn,$$

where R_u, n, w, and V are as before; W is the total weight in tons; and K is the air resistance or drag coefficient, which is 0.07 for conventional equipment, 0.0935 for container cars, and 0.16 for trailers on flatcars.

An attempt to make the Davis equation even broader in application is the adjusted Davis equation found in Hay's "Railroad Engineering" in the form $R_{adj} = K'R_D$, where R_D is the regular Davis value and K' is an adjustment factor, which is 1.00 for pre-1950 equipment, 0.085 for conventional post-1950 equipment, 0.95 for loaded container cars, 1.05 for trailers on flatcars, 1.20 for empty covered automobile rack cars, 1.30 for loaded auto rack cars, and 1.90 for empty auto rack cars.

An average value for train resistance is 4–10 lb/ton; 6 lb/ton is a frequently used value. Eight pounds per ton has been used in Section II.F.

For streamlined high-speed equipment, primarily passenger cars, a revision of the air resistance term in the Davis equation has been developed by Dr. A. I. Totten. Values based on front-end configuration and car length and smoothness are given in Hay's "Railroad Engineering."

Studies are being conducted by the AAR, the University of Illinois, individual railroads, and others to update the Davis equation in terms of current track and equipment. Track stiffness is found to decrease resistance, but rail surface irregularities, such as corrugations, cause an increase. Much importance is being attached to air resistance.

E. Other Resistance

Resistance is also encountered from several other sources.

1. Locomotive Resistance

In addition to overcoming the resistance of a train, a locomotive must overcome its own train resistance. An electric or diesel-electric locomotive can be considered as one or more additional cars and the Davis equation applied with appropriate coefficients for size, weight, and axle arrangement. For steam locomotives, the total resistance in pounds is usually taken as 20 times the weight on drivers in tons.

2. Starting Resistance

Additional tractive force is needed at starting to overcome inertia and warm the car bearings. For cars with roller bearings, the resistance is roughly equivalent to a Davis value where $V = 0$, but in cold weather or with friction bearings a higher value is in order, usually 18 lb/ton. When a train has been standing for a long period in subzero temperatures, a resistance of 30–50 lb/ton can be encountered. On such occasions, train tonnages may be reduced.

3. Wind Resistance

The Davis equation applies to still-air situations. A head wind velocity can be added to the velocity in the air resistance term or subtracted for a following or tail wind. In actual operation, both train and wind change direction. Side winds tend to force the wheel flanges away from the wind direction against the opposite rail. The worst condition is an opposing quartering wind, which exerts both a retarding and a flange pressure force.

The AREA presents a state-of-the-art formula in which $L = 8.25a \sin\beta\, KAV^2$, where L is the wind resistance in pounds; β is the angle between the train and wind direction; A is the projected frontal area in square feet; V is the train velocity in miles per hour; and $a = W_w / W_L$, with W_w being wind velocity and W_L being train velocity in miles per hour. The usual practice in areas of prevailing winds is to reduce the train tonnage.

Curvature, grades, and acceleration have enough significance to warrant separate comment.

F. Curves and Curve Resistance

Trains encounter additional resistance in going around curves. This is due primarily to flange friction against the gauge corner of the outside rail by the outside lead wheel of each truck and to slippage across the head of the rail by all wheels. The inner wheels may also press against the inside rail head. Values of 0.50–1.00 lb/ton per degree of curve have been used. The AREA recommends 0.80 lb. Thus, an 8000-ton train on a 2° curve would experience $8000 \times 0.80 \times 2$ or 12,800 lb of curve resistance in addition to the train resistance previously described. The effects of curve lubrication are discussed in Section V.E.1.

For a given divergence between two tangents (see curve geometry in Section VII.J), the cost varies inversely as the length of curve and directly with the central angle between the tangents. Curve resistance acts on the entire train and, like train resistance, is evaluated in cost per 1000 GTM. An average value for train resistance of 8 lb/ton is equivalent to the resistance of a 10° curve ($10 \times 0.8 = 8.0$). The equivalent curve-mile would be a 10° curve 1 mile long having a central angle of 528°. The summation of the degrees of central angle over a line segment divided by 528 gives the equivalent curve-miles for that segment.

But only a portion of costs that are specifically due to curvature should be considered. The costs of yard operation, interlocking plants, or traffic solicitation have little to do with curvature. By evaluating curve-related costs alone, it is found that only 27.24% of total costs per 1000 GTM can be attributed to curve resistance. Curve costs for a route with 2600° of central angle would be $(2600/528) \times$ cost per 1000 GTM $\times 0.2724$.

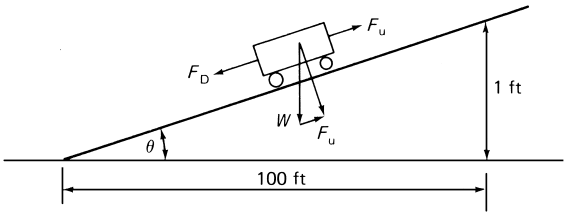


FIGURE 1 Derivation of grade resistance: $W = 2000$ lb; $\cos\theta \simeq 1$; $F/W = 1/100$; $F = 2000/100 = 20$ lb/ton.

As shown in the next section, curve resistance can be expressed in terms of the equivalent gradient and combined with grade resistance in one quantity.

G. Grade Resistance

By using only the component of height that tends to move a car down an incline, grade resistance is found to have a value of 20 lb/ton per percent of grade (Fig. 1). An 8000-ton train on a 0.40% grade would experience a total grade resistance of $8000 \times 20 \times 0.04$ or 64,000 lb. This must be added to all other resistance when determining the total amount to be overcome by the locomotive. The effect on train size and operation will be discussed in Section III.F.

H. Rise and Fall

The ascents and descents in a rail line are sources of resistance and cost. If the average value for train resistance of 10 lb/ton is assumed, it becomes the same as the resistance offered by a 0.50% grade. Such a grade rises 26.4 ft/mile, from which 24.6 ft of rise and corresponding fall represent one rise-and-fall mile. Rise and fall, as indicated in Fig. 2, can be taken as half the sum of rise and fall in both directions. The additional rise-and-fall equivalent miles can be added to the distance in determining total 1000 GTM.

But not all 1000 GTM costs are attributable to rise and fall. The percentage depends on the severity of the grades, which are customarily grouped into three categories. Class A grades require no change in locomotive throttle setting or brake application and can be ignored as a cost item. Class B grades require some slight throttle increase in going over a crest but require no brake application. The costs of such grades are 2.63% of total 1000 GTM costs.

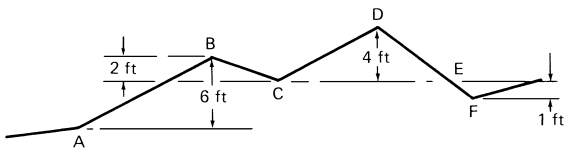


FIGURE 2 Rise and fall: A–B, 6 ft of rise; B–C, 2 ft of fall; C–D, 4 ft of rise; D–E, 4 ft of fall; E–F, 1 ft of fall.

Class C grades, requiring the use of throttle and brakes, involve 8.10% of 1000 GTM costs plus a 3.00% factor if the rise and fall is on a ruling or a helper grade (Section III.F).

I. Undulating Profiles

Rise and fall may appear in a profile as many closely spaced sags and crests. Such an undulating profile poses problems in train operation. When a long train is stretched over a series of sags and crests, part of the train is moving downgrade and part is moving upgrade. In a sag, the rear portion tends to run into the head portion. A sudden change in throttle position or a quick brake application can cause a break in two or even a derailment. The locating engineer should strive to avoid such situations. Operating personnel must learn to adapt to those that are in the profile.

J. Compensation for Curvature

When a curve coincides with a grade, the combined resistance can be sufficient to reduce train speed or cause the train to stall. The resistive effects of curvature can be combined with grade resistance by developing an equivalent grade for curve resistance. If the 0.80 lb/ton per degree is divided by the resistance of a 1% grade, that is, 20 lb/ton, the equivalent resistive grade of a 1° curve is 0.04%.

The combined resistance of a 1.20% grade on a 2° curve would be $2 \times 0.04 + 1.20$ or 1.28%. To maintain an effective 1.20% grade throughout, the grade can be compensated for curvature; that is, the actual profile grade is reduced by 0.08% to a 1.12% grade throughout the length of the curve, then continues at the initial 1.20% rate. There will be some increase in length of the grade to reach a given summit. The AREA has recommended a full 0.04% compensation, but some railroads have reduced the gradient by only 0.035 or even 0.030% when the curve is so short that the entire train is not on the curve at the same time.

III. MOTIVE POWER

A. Horsepower and Tractive Effort

The resistances discussed in the preceding paragraphs—train, curve, grade, and others—must be overcome by a pulling force exerted by a locomotive, usually a diesel-electric or electric locomotive. Steam locomotives, still used in some countries, have very limited use in the United States.

The ability to exert a pulling force at a certain speed is measured by the locomotive horsepower. Horsepower, that is, the rate of doing work, is obtained from the equation $\text{hp} = \text{TE} \times V/375$, where TE is the tractive effort or pulling force in pounds and V is the speed in miles per hour.

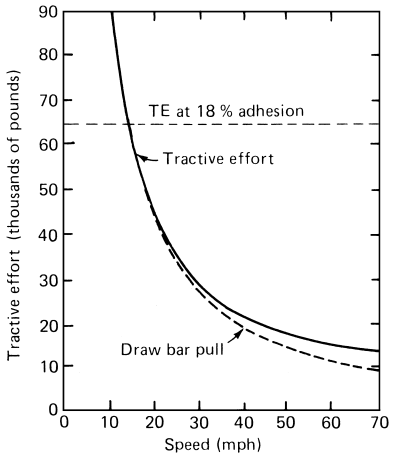


FIGURE 3 Tractive effort curve for a 3000-hp, 180-ton diesel-electric locomotive.

Rearranging terms and allowing for 18% losses, the tractive effort equation becomes $\text{TE} = 308 \times \text{hp}/V$, the expression used by the railroad industry. A tractive force curve for a 180-ton, 3000-hp diesel-electric locomotive is shown in Fig. 3. Some tractive effort must be used to overcome the locomotive's own resistance. Deducting this from the computed tractive effort gives the net tractive effort or drawbar pull.

Tractive effort at starting for a steam locomotive is usually taken as $\text{TE} = 0.85\ Pd^2s/D$, where P is the boiler pressure at starting, d is the cylinder diameter in inches, s is the length of stroke in inches, and D is the diameter of the driving wheel. For speeds above 15 or 20 mph, the mean effective pressure (m.e.p.) is substituted for P.

For electric locomotives, tractive effort is based on motor torque. It has the form $\text{TE} = 24TNe/D$, where TE is the tractive effort in pounds, T is the motor torque in pound-inches taken at a 12-in. radius from the motor shaft, N is the number of motors, D is the driving wheel diameter in inches, and e is the locomotive efficiency, 80–85%.

Electric operation is found in extensive use for rapid transit and by Amtrak in the Northeast Corridor between Washington, DC, and New Haven, CT. All other main line operations with the exception of a few short lines serving industries use diesel-electric motive power. Electrification of U.S. railroads is under continuing discussion and study, but the high initial costs and demand for a 5- to 10-year return of investment have prevented much action.

B. Diesel-Electric Locomotives

The diesel-electric locomotive is essentially an electric locomotive with a self-contained power plant, the diesel

engine. The engine, in turn, drives a generator or alternator that provides electric energy to the traction motors that propel the locomotive.

1. Diesel Engine

The prime mover for a diesel-electric locomotive is a compression-fired internal combustion engine. The four-stroke engine uses separate strokes for intake of fuel, compression (and ignition), power delivery, and exhaust or scavenging. The two-cycle engine combines intake and compression in one stroke and power delivery and exhaust in the second stroke. A modern 16-cylinder engine having a 10- to 11-in. stroke will have a cylinder displacement of 600–700 $in.^3$ with a 16:1 compression ratio.

The heart of the engine is the injector set in the head of the cylinder. A constant-stroke plunger pump operating from a camshaft delivers fuel to the cylinders in a measured atomized spray (about 0.03 oz) forced through the injector tips at 15,000–20,000 psi pressure. Efforts to improve fuel efficiency through computer control of speed, adjusting it to the load, and use of organic-based petroleum additives have effected a 5–14% increase in fuel efficiency in recent models.

Ignition occurs at a pressure of about 600 psi and a temperature of 1000°F. A fine clearance of about 0.000025 in., a large amount of air, and rugged construction are required of the injector. Additional air is obtained at high altitudes from a multibladed compressor or supercharger. The throttle that controls fuel admission and engine speed usually has eight settings or "notches." The eighth notch represents full engine output at engine speeds of 500–3000 rpm.

2. Starting Systems

To provide initial torque for compression, a gasoline engine is frequently used geared to the main shaft. More recent practice uses a storage battery to energize the generator, which then acts as a motor to turn the shaft. To avoid starting problems, the engine is seldom shut down and idles for long periods, but fuel conservation efforts have tended to reduce this practice.

3. Electric Coupling

The diesel engine is a constant-torque machine, turning at a high speed (500–3000 rpm) and with an optimum speed. Some method of coupling is necessary to reduce high rpm values to the more moderate wheel speeds. Mechanical and hydraulic coupling have been used in Europe but are costly to maintain under U.S. conditions. Because of its operating flexibility and relative economy, electric coupling is in universal use in the United States. The electric generator or alternator mounted on the engine shaft delivers electric energy to direct current (dc), series-wound motors geared to the driving wheel axles.

4. Electric Power System

The generator is usually self-excited and shunt wound. Two field strengths aid in transition. Output is controlled by varying the engine speed through the throttle or notch position. The exciter current automatically increases or decreases with the shaft speed, permitting the generator output to match engine output and capacity.

The dc motor requires a high current at low voltage to deliver the high torque needed at low speeds and starting. Torque varies with motor design (number of poles and windings) and with armature current.

In the motoring process, as speed increases, enough back electromotive force (emf) may be developed to oppose fully the terminal voltage applied to the motor. To obtain higher voltages (at less current) for higher speeds, a process of transition, usually automatic, changes the motor hookup from series to series–parallel or from series–parallel to a higher field strength. For still higher speeds, a shift to full parallel may be used to give maximum voltage across each motor. Recent locomotive designs have incorporated alternators capable of delivering enough voltage throughout the speed range to render transition unnecessary.

The introduction of three-phase induction-type traction motors is expected to bring higher tractive effort and horsepower capability. The traction alternator supplies dc output through solid state voltage inverters, which convert dc current from the alternator to the three-phase alternating current (ac) voltage and frequency required by the traction motors. These motors, first used in the Amtrak F60PH diesel-electric locomotive, have been modified to provide the high adhesion required for heavy axle load freight haul. Elimination of the dc motor commutator and brushes should reduce maintenance costs and problems from moisture and snow.

5. Gearing

The maximum and minimum speeds of a locomotive are determined by the characteristics of the motors and by the gearing. A gear-and-pinion combination connects the driving wheels to the motor shafts. The dc motor can withstand a brief overload, but the heavy current demand at low speeds can cause motor damage and failure. The motor speed must not drop below a predetermined minimum, usually 500–600 rpm for continuous operation. The maximum motor speed, about 2500 rpm, is limited by centrifugal force, which can cause the armature windings to work out of their slots at excessive speeds of rotation.

The relation between these effects is given by $V = (\text{rpm}/Dg)/336G$, where V is the locomotive rail speed in miles per hour, rpm is the motor speed in revolutions per minute, D is the diameter of the driving wheels in inches (usually 40 in.), g is the number of teeth in the pinion, and G is the number of teeth in the gear.

6. Ratings

Because of electric couplin, the diesel-electric locomotive is able to exercise full engine capacity over a wide range of rail speeds, a great help in starting and on heavy grades. There are, however, limitations. If full power is applied to the motor for more than a brief period, the heavy current flow will cause the motors to overheat, experience a flashover, and burn out. Consequently, locomotives have short-term and long-term ratings. The short term may be a 5- to 10-min period, sufficient to get a train in motion (whereupon resistance decreases) or to surmount a heavy grade. The ammeter before the engineman is sometimes divided into red, yellow, and green areas to indicate current ranges for allowable time periods: the green area represents continuous or cruising speed, red represents the short term, and yellow represents an intermediate time interval.

7. Adhesion

Another limitation on the effective tractive effort is adhesion, the ability of the wheels to grip the rails. Effective tractive effort is limited to μW, where W is the locomotive's weight on driving wheels and μ is the coefficient of friction between the wheel and the rail. The coefficient of friction is frequently taken as 0.25, but, to avoid difficulties in starting or on a heavy grade, it is held to 0.18. Thus, a locomotive weighing 240,000 lb would have an effective tractive effort of $240{,}000 \times 0.18$ or 43,200 lb. If the power plant can produce more tractive effort than the 43,200 lb, the wheels will tend to slip.

Adhesion also varies with rail conditions from 0.30 + for a clean dry rail to 0.15–0 for a greasy, frosty, or slightly moist rail. A rail washed clean by a heavy rain has a high value. A rail washed clean by chemicals can have an additional 17–22% of adhesion, but this is not carried out in practice. The use of sand to increase adhesion on heavy grades or in making station stops is traditional. Locomotives are equipped with sandboxes containing a dry, sharp sand. Flow is controlled by the engineman in the cab.

There is always some slip between the wheel and the rail. To avoid excessive slip, the difference between wheel circumference speed and rail speed should not exceed $1\frac{1}{2}$ mph. Slip may be reduced by monitoring individual axles and applying sand automatically if slip occurs or by adjusting generator output automatically to achieve a non-slip condition. One system measures ground speed with a small Doppler radar with electronic solid state sensors for comparison with motor speeds and current control.

8. Dynamic Braking

As a supplement to air brakes, dynamic braking assists in controlling train speeds. Leads to the generators can be reversed, causing the motors to act as generators and thereby exercising a retarding force on the wheels and axles of the locomotive. The current generated is wasted in heat from grids mounted on the cab roof. Dynamic brakes can permit descent of long grades without the need to stop and cool brake shoes or set retarders. Thermal cracks in wheels are reduced, there is less wear on brake rigging, and train schedules can be shortened.

9. Fuel Consumption

The diesel-electric locomotive has a thermal efficiency (percent of useful work derived from a unit of fuel) of 20–25%. Since a gallon of fuel oil contains approximately 130,000 Btu of energy, only 20–25% of this would be available to produce work, that is, pull the train. These values are in contrast to the 5–8% efficiency of a steam locomotive. The straight electric locomotive, taking gear and line losses into account, has a thermal efficiency of about 23%. The average consumption of fuel is 0.002 gal/GTM (or 500 GTM/gal). In 1979, U.S. railroads consumed approximately 43 billion gallons of fuel oil. The efforts to improve fuel use efficiency by improved engine design, the use of fuel additives, and a search for alternative fuels (including redesign of the coal-burning steam locomotive) indicate the interest in this area.

C. Electric Locomotives

The electric operation of main line railroads finds its principal use in the United States between Washington, DC, and New Haven, CT, in the Northeast Corridor. Drawing on the experience and facilities of the Pennsylvania and New Haven railroads, the line uses a 25-Hz current at 11,000 V supplied from overhead conductors supported from an elaborate system of steel towers and catenary cables. Current flows to the locomotive through a pantograph—diamond frame or single bent-arm type—on the roof of the cab. A plan to convert the system to 25,000 V at 60 Hz has been deferred due to lack of funds.

1. Motors

The dc series-wound motor used in diesel-electrics finds use in electric locomotives. A thyristor, solidstate,

silicon-controlled, on-board rectifier converts the ac supply to dc for the motors. Earlier systems used the bulkier ac, single-phase, series-wound motor, which lacked the low-speed capability of the dc motor and performed best at high speeds.

A few recently installed coal-hauling railroads in the United States and Canada (including the British Columbia Railway's Tumble Ridge Branch) are using a 50-cycle, 50,000-V supply.

2. Other Motor Systems

Until the advent of the diesel-electric motor, motor-generator sets were used to supply dc current to dc traction motors. The ac current supply from the transmission line feeds an on-board generator that produces dc current for the motors. In another system, small mercury are on-board rectifiers or ignotrons permit the use of 60-cycle commercial current as well as the 25-cycle, 11,000-V supply. The polyphase and split-phase induction motors found use prior to the diesel-electric on the Norfolk and Western and Virginian railways, where constant speed for slow heavy trains on long grades made them useful.

3. Torque Transmission

Torque is usually transmitted to the driving wheels by the same gear-and-pinion arrangement used on diesel-electric motors. An earlier variation had the gear mounted on a quill that fit over the axle. The quill engaged the driving wheel through spring connectors.

Gearless drives, such as those used on the former Milwaukee Road's bipolar motors, had the rotating armature mounted directly on the driving axles. Side rods have had only limited use in the United States. Their use permits motors to be mounted in the cab and the drive transmitted mechanically by the side rods to the axles. Side rods permit a more equalized torque to be developed on each axle.

4. Ratings

Electric locomotives have short- and long-term ratings based on heating of the motors. They are also given a 1-hr rating that permits a high output useful in negotiating long grades. The locating engineer should endeavor to keep long heavy grades within this limit, or the heavy grade should be followed by a lesser gradient or even a descending grade to permit cooling of the motors.

5. Tractive Effort

Tractive effort curves are usually provided by the manufacturer. Tractive effort can also be computed with a relation based on equating work performed at the rim of the driving wheels to that performed by torque exercised 12 in. from the motor shaft: $\mathrm{TE} = 24TNe/D$, where TE is the tractive effort in pounds; T is the torque in pound-inches (taken 12 in. from the center of the motor shaft); N is the number of motors; e is the locomotive efficiency, 80–85%; and D is the driving wheel diameter in inches.

6. Power Transmission

Main line railroads use an overhead electric power supply system. The principal supply is at high voltage from a central power station to transformer stations and then to a contact wire suspended above the track. Voltages are reduced by the transformers to that required by the traction motors. For light rail and streetcar lines a single contact wire or cable on brackets feeds the motors through a trolley wheel collector.

With heavier construction, the contact cable is a stranded or grooved copper wire suspended from a carrier cable that assumes the shape of a catenary curve between poles or towers. The contact cable is supported by $\sim\frac{1}{4}$-in. hanger rods from the catenary at a uniform distance above the track. The standard cable size is 12.24×12.24 mm. For economy in construction and maintenance, the trend is toward lighter, less elaborate support systems—wooden poles instead of the steel towers that have marked the Northeast Corridor.

To avoid loss of contact from vibration at high speed, a heavier cable, 12.24×17.27 mm (0.68×0.428 in.), is used. The contact carrier is often given a zigzag path above the track to equalize wear on the pantograph collector shoes. The copper collector shoe is mounted on a diamond or a single-arm pantograph that permits variations in the height of the arm to conform to the height of the contact cable. Spring pressure maintains contact between the shoe and the cable. Single shoes are used for currents up to 500 A; double shoes are used for heavier currents.

7. Direct Current Systems

A direct current (dc) supply is only used over short distances because of the large line losses. A conductor rail, similar to a running rail, offers a greater flow area and less resistance that a contact wire. The conductor or third rail is mounted beside the running rail on insulating stands or "pots" attached to an extended fourth or fifth crosstie. Contact shoes on the traction trucks collect current for the motors. The motor and third-rail voltages vary with individual systems from 400 to 600 V. Transformer substations convert high transmission voltages to those suitable for the motors. The location of these stations is an economic function of station cost versus power loss over the

intervening distance between stations. The allowable loss is also dependent on traffic density and speed.

Third-rail, direct current operation is confined almost exclusively to rapid transit lines where the distance to the central power plant is relatively short.

D. Acceleration and Deceleration

Because additional tractive effort is required to increase train speed, acceleration can be thought of as a form of resistance. By equating the work equivalent of kinetic energy to its value, $\frac{1}{2}mv^2$, expressing the several terms in railroad units, and adding 5% for the rotative energy for the wheels, the distance and time required to accelerate or decelerate from one speed to another are obtained. This is useful information in planning or evaluating train performance. The distance d in feet to accelerate from a speed of V_1 to V_2 mph is given by $d = 70W(V_2^2 - V_1^2)/F_a$, where W is the train weight in tons. The time t for that acceleration in second is $t = 95.6W(V_2 - V_1)/F_a$. In both equations, F_a is the accelerating force. A negative value for d or t indicates deceleration.

1. Accelerating Force

The accelerating force F_a is the residual tractive force remaining after all other resistance has been overcome. It is available in diminishing amount as the train accelerates until total resistance equals total available tractive effort. This is the maximum or balancing speed that can be obtained (Fig. 4).

Because both tractive effort and train resistance are changing at the same time that speed increases, the total distance and time to accelerate must be computed in incremental speed steps. The smaller the speed increment, the more accurate the computation becomes. Use of a computer permits incremental steps of 1.0–0.50 mph. In ascending a grade or in coasting, the final speed may be lower than the initial speed, thus giving a negative answer to the equation, indicating deceleration.

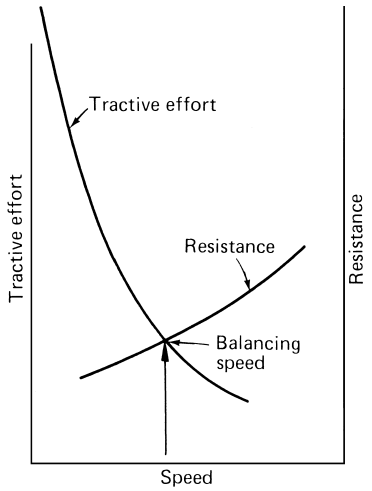


FIGURE 4 Balancing speed.

2. Speed Control

No speed is safe unless it can be fully controlled and the train stopped. Train speeds are controlled by air brakes. The brake system consists essentially of an air compressor on the locomotive that supplies air under pressure, 90–110 psi, to a train line running the length of the train. A similar pressure is maintained in an auxiliary reservoir on each car kept charged by the train line.

To apply the brakes, a reduction is made in the train line pressure that actuates a brake valve on the car, permitting air to flow from the auxiliary reservoir into a brake cylinder, where a piston is driven forward moving the brake rigging and applying brake shoes to the treads of the car wheels. Recharging the train line again actuates the brake valve, releasing air from the brake cylinder and recharging the auxiliary reservoir. As air is released from the brake cylinder, a spring moves the brake cylinder piston backward, releasing the brake shoe contact on the wheel treads. Because of this three-way function, the brake valve is sometimes referred to as a triple valve, but more recent designs have been given a variety of trade identifications.

The reduction in train line pressure occurs serially from car to car at an approximate rate of $\frac{1}{2}$ sec per car. On a 100-car train, the brakes would not apply on the last car of the train until 50 sec after the intial reduction on the locomotive. If an emergency application is made, there is danger of derailment form the rear cars running into the head end of the train.

To prevent such occurrences, a light or service application is customary, bringing the train gradually to a stop. Quick-acting features and a graduated release, which reduces brake shoe pressure after the intial application, contribute to smooth and safe handling.

Additional features include a "lap" position on the engineman's brake handle that permits the brakes to remain partially applied with no pressure change. Retainers, spring-operated valves, retain brake cylinder pressure at a 10-psi minimum while the engineman "releases" the brakes in order to rebuild train line and auxiliary reservoir pressures. Retainers are set by hand before descending a long steep grade and must be released when the descent is completed. Train delays ensue.

3. Braking Distance and Time

The equations used for acceleration can also be used for calculating braking distance and time, except that the

accelerating force F_a becomes the gross braking force F_b. Force F_b is equal to the brake shoe pressure plus train and curve resistance (if any) plus or minus grade resistance.

The braking force $F_b = Pef$, where P is the brake shoe pressure (total force on piston × brake rigging ratio), f *is* the coefficient of friction between brake shoe and wheel tread, and e is the efficiency of the system. The coefficient of brake shoe friction for cast-iron brake shoes is low at high speeds and high at low speeds, where least needed (Fig. 5). Composition brake shoes have been developed that give a more nearly uniform coefficient throughout the range of speeds except in the range 5–0 mph, where it also becomes high. Note that the product of brake shoe friction and efficiency is plotted in Fig. 5.

One other factor is the braking ratio by which brake force is multiplied. If F_b were based on loaded cars, the results of a brake application on empty cars could be disastrous—too much force, a very quick stop, flat wheels from sliding, and even derailment. To avoid these, the braking force is set for only a percentage of the empty car weight so that $P \leq \mu W$, where W is the empty car weight and μ is the coefficient of friction between the wheel and the rail. A braking ration R is usually taken as 0.90 for diesel-electric and electric locomotives and passenger cars. Freight cars have ratios of 0.50–0.80 based on their empty weight. From this, $F_b = Rwef$.

The final equations for distance and time to brake from one speed to another include the distance the train travels while the brakes are being applied serially throughout the train:

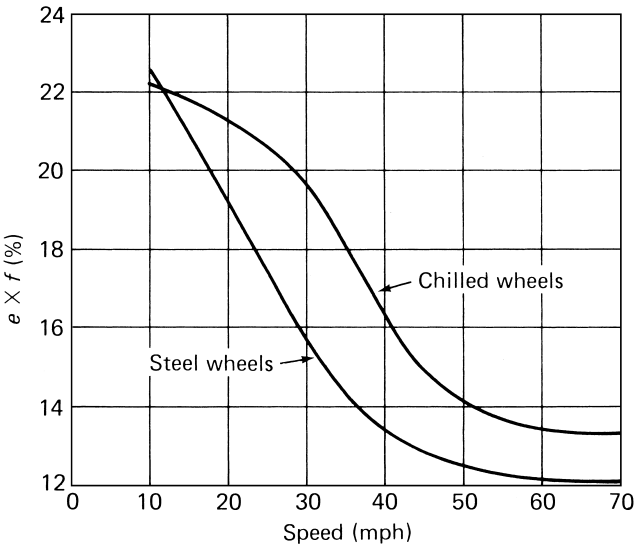


FIGURE 5 Coefficient of friction × brake shoe efficiency at various speeds for chilled iron and wrought steel wheels. [Based on results in Bull. 257 (1933) and Bull. 301 (1938), Univ. of Illinois Engineering Experiment Station, Urbana, IL.]

$$d_s = \left[1.47V_1t_1 + \frac{0.035W\left(V_i^2 - V_f^2\right)}{Rw'ef + R_t + R_c \pm R_g}\right] \times 1.3$$

and

$$T_s = \left[t_b + \frac{0.048W(V_i - V_f)}{Rw'ef + R_t + R_c \pm R_g}\right] \times 1.3,$$

where W is the train weight in tons, t_1 is the time for train brakes to apply serially throughout a short train (10 cars or less) or an initial 10±-car increment, w' is the empty car weights in tons, e is the brake rigging efficiency—often combined with f, f is the coefficient of brake shoe friction, V_i and V_f are the initial and final train speeds in miles per hour, R_t is the train resistance in tons, R_c is the curve resistance in tons, and R_g is the grade resistance in tons. The term 1.3 represents a 30% factor of safety.

These equations must be solved in incremental speed steps as in computing acceleration. For a short train of 10 cars, more or less, an average value can be taken for t_1, brake application time. For longer trains, an incremental number of cars for each step can be assumed based on the rate of brake pipe reduction over the preceding speed increment.

Railroads normally do not depend on computing values for braking distance and time. Instead, dependence is placed on tests made with representative trains. The equations can be useful, however, in projecting performance of trains for which tests have not been conducted.

4. Train Performance Calculators

The availability of the computer and acceleration/braking information enables the computation and graphical presentation of a train's performance in traversing a given track segment. The computer program must include the storage of all speed restrictions, grades, curves, station stops in order of occurrence, locomotive tractive effort at various speeds, and enough data on rolling stock so that train resistance can be computed. The program must permit calcualtion for each speed increment of all resistance, the net or residual tractive effort available for acceleration, and the braking distance and time required for each change in speed or stop. From these data, the performance of the train over each portion of the route and the time required to traverse the route are determined. A broad range of computers can be used or a special train performance calculator can be developed for the specific purpose.

5. Very High Speed

Conventional line haul train speeds will vary from 20 mph or less for low-priority freight branch lines to 50–70 mph

for main line freight and from 60 to 125 mph for passenger trains and other premium services. The higher speeds are found with the Japanese 125 mph Sinkhasin, on certain lines of the British and German railways, and on certain segments of the Northeast Corridor in the United States.

Normal speeds of 160 mph are operated on the TGV (Train de Gran Vitesse) system in France on tracks specially designed for very high speed. A top test speed of 300 mph has been attained. Operating speeds of 180 mph are planned for TGV service in Brittany. In Germany, tracks for 150 mph are under construction. In the United States, a maximum speed of 252 mph was attained with a linear induction powered car at the Pueblo Test Track. Plans for very high speed operation have been proposed for several locations in the United States but have not advanced beyond the talking stage.

For these premium speeds, a high degree of precision in design, construction, and maintenance is required. Curves should not exceed 30 min (radius of 11,460 ft). The maximum curvature for the TGV operation varies between $15\frac{1}{2}$ and 30 min. Long, flat, tangential turnouts with swing-nose frogs are used (but sparingly). Signal systems are designed for longer stopping distances, and there are no rail or highway crossings at grade.

E. Velocity Head

A train traveling at a certain speed contains, in addition to the energy from the locomotive, energy from momentum. The measure of the momentum energy is velocity head. From physics, velocity head is $h = v^2/2g$, where v is the velocity and g is the acceleration of gravity. Converting to miles per hour and adding 5% for rotative energy in the wheels, $h = 0.035V^2$ and $V = 5.35h^{1/2}$. In general, $h = 0.035\,(V_i^2 - V_f^2)$, where V_i and V_f are initial and final velocities.

Velocity head can be represented by the height a train would rise on an incline, ignoring friction, if all power were shut off (or just enough power applied to overcome train resistance). It also indicates the speed a train would acquire in coasting down a grade for a vertical distance h. If a train approaches a grade at 50 mph with the throttle open only enough to overcome train resistance, it will ascend the grade before stalling to a height of 0.035×50^2 or 87.5 ft above the base of the incline.

Velocity head has been used to overcome short, steep grades where only a portion of a long train would be on the grade at one time. Momentum must not enter into the design of long, heavy grades because velocity head is lost at a rate of about 30% per mile due to shock, sway, and concussion in the train. Also, if the train has to stop, all velocity head is lost.

F. Tonnage Ratings

A tonnage rating is the weight of train in tons a given locomotive or locomotive consist can haul under stated conditions of speed, gradient, and curvature. The stated conditions are usually based on the ruling grade, which may be defined as the grade that establishes the maximum tonnage that can be hauled up the grade at a stated speed by a given locomotive or locomotive consist. The ruling grade is not always the steepest on the territory. Momentum grades and grades using helper or pusher service may be steeper.

1. Simple Tonnage Ratings

The locomotive net tractive effort or drawbar pull divided by the unit resistance of train plus grade plus curvature gives a simple tonnage rating. A locomotive producing 60,000 lb of tractive effort could haul 10,000 tons if train resistance were 6 lb/ton. If, however, one assumes a 1% grade, the total unit resistance would be 26 lb/ton, and the locomotive could only haul 2300 tons. Hence, the importance of grades.

The speed at which tractive effort and resistance are computed may be the lowest speed maintained, 10–12 mph, for example, to negotiate the ruling grade, or it may be a higher speed based on the type of train and schedule to be maintained. Each type of train will have a different minimum speed, and each type of locomotive or locomotive consist will have a different tonnage rating.

2. Adjusted Tonnage Rating

The simple tonnage ratings just described are based on uniform train consists, that is, all cars in the train of the same weight and type and with the same weight of contents. This is usually found only with unit trains hauling bulk commodities. Trains carrying general freight will have cars and contents of various weights—empty cars, partial loads, and full loads. The train resistance will be different for each. An adjusted tonnage is required.

The adjusted tonnage is useful in building trains and matching train weight and resistance to locomotive capacity. An adjustment factor is added to the weight of each car to be placed in the train until the total car weights plus the adjustment for each car equal the adjusted rating.

The adjustment factor is obtained by equating a train of empty cars with a train of loaded cars:

$$W + KN = W_1 + KN_1$$

$$K = (W_1 - W)/(N - N_1),$$

where K is the adjustment factor to be added to the weight of each car, W is the weight of the train with empty cars in tons, W_1 is the weight of the train with loaded cars, and

N and N_1 are the number of cars in the W and W_1 trains, respectively. The expressions $(W_1 + KN_1)$ and $(W + KN)$ equal the adjusted tonnage rating.

3. Cost of Ruling Grade

The costs of grade (rise and fall) for moving tonnage in terms of cost per 1000 GTM have been considered. The procedure gives the cost of moving a given total tonnage over a line. The ruling grade, which limits train size, determines the number of trains required to move that tonnage in a stated time period, usually a day or a year.

The number of trains to move a given tonnage over a line is a function of the tonnage rating. The higher the rating, the fewer trains needed and vice versa. Thus, tonnage ratings are directly related to traffic density. The unit cost is the cost per train mile obtained by dividing total operating expenses by total train miles. Such data can be obtained from a railroad's records or from published statistics of trade and government agencies.

Not all items thus included in train mile cost pertain to actual train operation or to the number of trains, that is, traffic density. The costs of the legal and stores departments, for example, bear little relation to the number of trains operated. Statistical analysis demonstrates that the costs incurred due to traffic density are only 34.274% of the total. The cost of moving 20,000 gross tons in 4000-gross-ton trains would be $(20{,}000/4000) \times$ distance in miles $\times$ cost per train mile $\times$ 0.34274. Train mile costs can also be calculated by combining fuel, lubricants, and other train supplies with wages of train and enginemen, enginehouse expense, and a portion of track maintenance costs.

G. Location Problems

The actual location process involves a selection of control points—initial and final termini, mountain passes, river crossings, and so forth—reconnaissance to determine the "lay of the land" between the control points, a paper or preliminary survey, and final and construction surveys. In rapid transit location in urban areas, the underground environment with its sewers, water, gas, and electric and communication lines can be the most important feature. Tunneling effects on the foundations of adjacent buildings must also be considered.

1. Surveys

The procedures to mark out a new location or revise an existing location may, as in the case of certain Canadian iron mine railroads, involve all the traditional means—on foot or horseback, by dog sled, or by canoe. But, the same lines also used airplanes, helicopters, and aerial photography. Modern location bypasses much of the physical effort through the use of available geological survey maps supplemented by aerial surveys and photogrammetry, which permit a close approximation of final position and even the computation of earthwork quantities from which construction and operating costs can be evaluated. The conventional tools of surveying—transit, level, theodolite—have received an addition, inertial surveying. A vehicle carrying a cluster of three accelerometers establishes a base line. Travel times of the vehicle are converted into distances, and off-line objects are positioned by laser beams and coded for reference.

Several alternative locations are usually available between the control points. These are evaluated by using the costing procedures for distance, curvature, rise and fall, and ruling grade to obtain the most cost-efficient location. Aided by computer analysis, many such alternatives can be examined and compared. One study of alternative rail locations on the Niagara Peninsula compared over 60 different rail networks. Three of these were then selected for very detailed analysis.

2. Overcoming Elevation

Elevation and the means to overcome it are significant cost factors, in both construction and operation. Inclined planes, rack-and-pinion locomotives, switchbacks, helper locomotives, doubling (taking the train up the hill in two or more sections), tunnels, and development have all been used. Most still find some use in various parts of the world.

A basic principle enunciated by Wellington calls for penetrating a mountainous barrier as far as possible at the lowest possible elevation, and then overcoming the main barrier at one spot by one or more of the foregoing methods. Two principal methods find favor today.

1. Development introduces additional distance into the location by loops and spirals to keep the gradient within desired limits. Grades of more than 1% are considered excessive. Grades of less than 0.10% are the most favored. Spirals have even been constructed within tunnels to obtain additional distance. Development involves the costs of constructing additional mileage and the operating costs of additional distance and curvature.

2. Tunnels are expensive to build and maintain but offer a relatively trouble-free solution. Construction can involve blasting through rock or some form of shield boring through soft ground, often under pressure, with air locks leading to the tunnel face during construction.

Tunnels in stable rock are often unlined. In less stable ground a tunnel lining is required. Old tunnels have timber linings, which makes them susceptible to fire. Others are

lined with brick. More recent designs use concrete or steel ribs with steel plates or concrete block liners.

Ventilation and drainage are major problems. Air shafts and/or forced draft are used to clear locomotive gases. One system blows the gases back along the train. Another blows the gases ahead of the train. Tunnel doors, controlled by track circuits to open and close automatically, have improved efficiency in some tunnels. A break in grade near the midpoint of the tunnel permits gravity to assist in draining toward both ends. Ice formations in cold weather may call for heating units in the track ditches or labor forces to clear ice from walls and track. Moisture is still a problem; it forms ice in the winter and combines with locomotive gases to corrode rails and fastenings.

IV. SUPPORT SYSTEMS

A. Track Criteria

The fixed plant in the railroad system includes the right of way (land and fences), and track structure (rails, ties, and fastenings), and the subgrade-ballast support system, as well as turnouts and crossings, highway crossings, bridges, yards, and tunnels. The support system is of prime importance in meeting the requirements of a good railroad. The requirement for safety is always fundamental. Safety, in turn, is derived from track stability, the resistance of the track structure to deflection and permanent deformation under load. Deflection accentuates the differential movement between the several track components, resulting in wear and deterioration of surface, alignment, and the component parts. These criteria must be met with economy in construction and maintenance costs.

B. Modulus of Track Elasticity

The track must possess stiffness in itself and in its support system to minimize deflection. One measure of stiffness is the modulus of track elasticity as given by $p = uy$, where p is the applied load; y is the deflection; and u is the proportionality factor, the track stiffness modulus. Professor Arthur Newell Talbot defined the modulus as the load per inch of rail length in pounds required to depress that rail 1 in. It is also the load on one tie required to depress the rail 1 in. divided by the tie spacing. It can range in value from 500 to 3500 for ballasted track; 2000 is normally a good value.

The modulus is also a lumped parameter encompassing the elasticities of the crosstie fastenings, ballast, and subgrade. Various studies have endeavored to evaluate the elasticities of these elements individually, but with little success. The modulus also varies under dynamic loads and speeds. Mathematical models have been developed that relate support system behavior to traffic loadings (wheel loads, speed, and traffic density) and to the stresses and deflections arising in the track support and the rails. Difficulty is usually found in getting suitable values for the numerous terms in such models. One of the most useful of these, however, is the Illitrack model developed by Professor Marshall Thompson and associates at the University of Illinois. Actual field data can be introduced and, by use of suitable algorithms, numbers applicable to in-track conditions can be derived in terms of track deflection, stress, resilience, and permanent deformation.

C. Talbot Equations

Reliance is still placed on the equations and procedures of Dr. Talbot and his AREA Committee on Stresses in Railroad Track. By considering the track as a continuous and elastically supported beam, a series of working equations have been developed:

Y_0	$= P/(64EIu^3)^{1/4}$	maximum deflection in inches
M_0	$= P(EI/64u)^{1/4}$	maximum bending moment in pound-inches
s_0	$= P/2$	maximum shear in pounds per square inch
P_0	$= P/(u/64EI)^{1/4}$	maximum reaction in pounds

where P is the applied load in pounds, a static value that must be augmented for dynamic (moving load) conditions; u is the modulus of track elasticity in pounds per inch of rail length per inch of deflection; E is the modulus of elasticity of rail steel $= 30 \times 10^6$; and I is the moment of inertia of the rail in (inches)^4.

These relations are represented graphically in Talbot's master diagram (Fig. 6), based on values of x_1, the distance from the point of load application to the point of zero bending, where $x_1 = (\pi/4)(4EI/u)^{1/4}$. Note that x_2, the distance to the point of zero deflection, is $3x_1$.

The values of Y_0, M_0, S_0, and P_0 are maxima occurring where the load is applied. Other points on the bending and deflection curves are found by multiplying the maximum values by the ratio between x_1 and the distance x to the point desired. For example, the value for the deflection (or bending moment) at a point 45 in. distant from the loading point would be $45/x_1$ to obtain the abscissa position of the desired point. The ordinate value for deflection (or bending moment) at that point is read from the ordinate scale as the proportional part or percent of the maximum deflection (or bending moment) at the 45-in. point.

For multiple wheel loads, as in the two- and three-axle trucks of cars and locomotives, the values for individual wheels are computed and superimposed algebraically

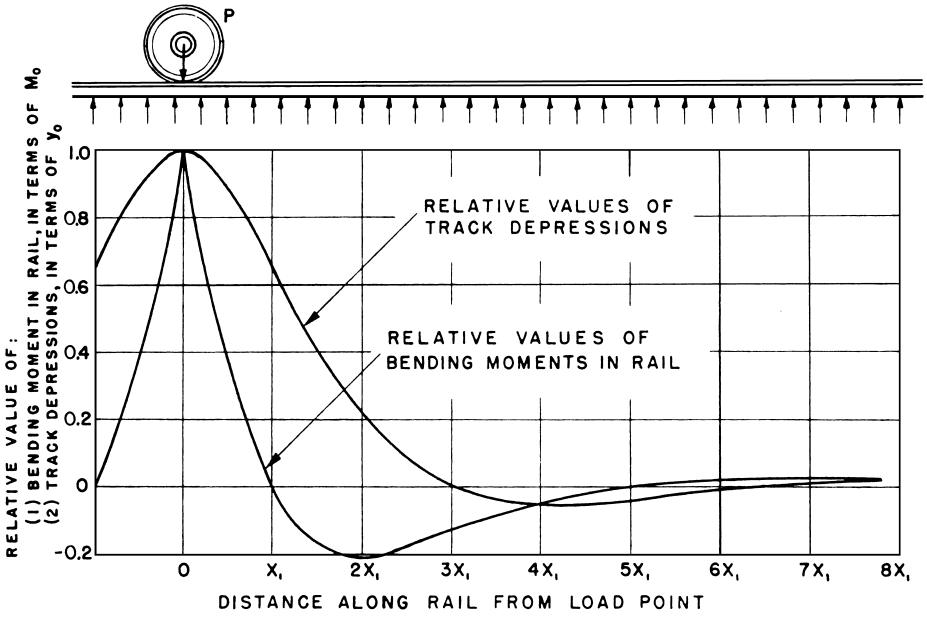


FIGURE 6 Talbot's master diagram. [From Talbot, A. N., Chairman. (1918). "First Progress Report, Special Committee on Stresses in Track," *Proc. Am. Railw. Eng. Assoc.* **19**.]

at the appropriate axle positions to obtain composite deflection and bending moment curves (Fig. 7).

D. Bending Stress

The bending stress is computed for the rail by the equation $S_b = Mc/I$, where S_b is the bending stress in pounds per square inch, M is the maximum bending moment in pound-inches, I is the moment of inertia of the rail in (inches)^4, and c is the distance in inches from the base of rail to its neutral axis.

The yield point of rail steel is approximately 70,000 psi. The allowable working stress is, however, much lower. The

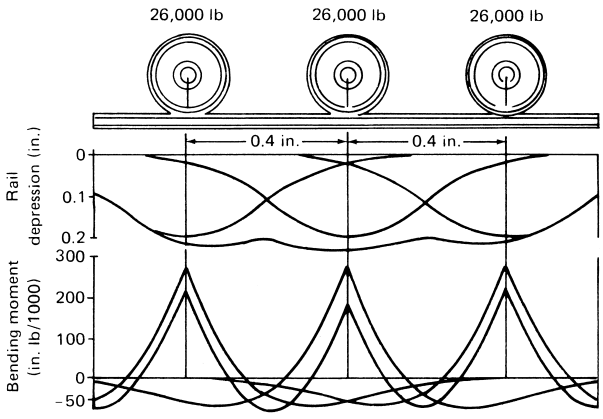


FIGURE 7 Six-wheel truck deflection and bending diagram. [From Sillcox, L. K. (1943). "Talbot and Track," Watertown, New York.]

70,000 psi is first reduced by 7000 psi for jointed rail and by 20,000 psi for continuous welded rail (CWR) to allow for thermal stresses. The result is then divided by the product of several factors of safety: 20% for lateral bending, 15% for rail wear, 15% for unbalanced superelevation, and 25% for track conditions (rough track). The rounded-off stress thus becomes 32,000 psi for jointed track and 25,000 psi for CWR track.

E. Dynamic Effects

The Talbot procedure is basically a static method. Factors to convert to dynamic loads of moving trains have been devised. The AREA suggests an increase or impact factor, based on a 33-in. wheel, of $\theta = 33V/100D$, where θ is the impact factor, V is the speed in miles per hour, and D is the wheel diameter in inches. Thus, if P is the static wheel load, the dynamic load is $P_d = P + \theta P$.

F. Support versus Rail Stiffness

The Talbot equations show that the stiffness of the support system is more significant in reducing track deflection and bending stress than increasing rail weight and stiffness. This is shown in the plots of Fig. 8. The effects on deflection of changes in rail weight and stiffness (the moment of inertia) are much less than those of changes in track modulus values. Which should govern in a particular situation is largely a question of comparative costs. A stiff track modulus is, however, desirable not only to reduce deflection,

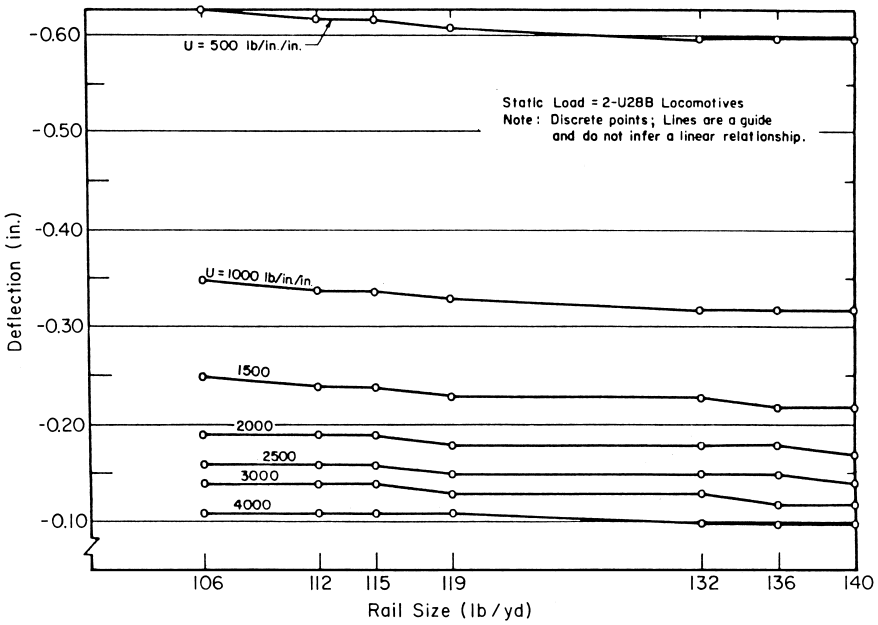


FIGURE 8 Effects of rail size on deflections for various track moduli. [From Butler, A. B. (1969). Master of Science thesis. Civil Engineering Dept., Univ. of Illinois.]

but also to protect whatever rail is in place and to reduce train resistance through the wave action in the rail. For heavy haul railroads, both a stiff support and a heavy stiff rail are necessary. Nevertheless, a certain amount of resilience is necessary to absorb impact and avoid rail and wheel breakage. This is seldom a problem.

G. Transmission of Pressures

Wheel loads are transmitted, by rails, ties, and ballast, to the subgrade with diminished unit pressure. The subgrade must have a load-bearing capacity equal to or greater than the load transmitted through the ballast; otherwise, the ties and ballast will settle in a nonuniform state of permanent deformation, that is, rough track.

Studies by Talbot led to an equation showing the pressure beneath the centerline of the tie (under the rail) as $p_c = 16.8p_a/h^{1.25}$, where h is the depth of ballast between the bottom of tie and the subgrade and p_a is the unit distribution of the tie load over the bearing surface of the tie. Because only the outer two-thirds of the tie length receive compaction, p_a becomes $(P'/2)/bL) \times 2/3$ or $3P'/bL$, where P' is the axle load in pounds, b is the tie width in inches, and L is the tie length in inches.

Talbot's equation for p_c can be rewritten as $h = (16.8p_a/p_c)^{4/5}$ to give the depth of ballast needed to bring the applied load within the bearing capacity of the subgrade. The soil-bearing capacity (2.5 times the unconfined compressive strength) can be determined by the soils engineer's test for unconfined compressive strength or by a recently devised British Rail soil stress index test. The AREA has noted 20 psi as representing a normally stable soil material, but soils with only 5–10 psi of support are not uncommon. Charts developed by Dr. Nathan Newmark at the University of Illinois can also be used to determine the distribution of pressures at a given depth.

The effect of ties on load distribution will be discussed in Section IV.L.

H. Subgrades

The subgrade must have adequate supporting capacity for the loads imposed through the ballast. An inadequate subgrade is characterized by soft and subsiding track, mud infiltrating the ballast; pumping ties, ballast penetration into the subgrade, forming water-filled ballast pockets, slides, and even complete failure.

Subgrade stability is a function of soil type, water content, degree of compaction, mineral content, and other characteristics. The desired goal in construction and maintenance is to secure high internal friction and shearing strength in coarse-grained materials and high cohesion in fine-grained soils. Clays and silts give the most trouble. With these, the soil particles should be so tightly compacted as to squeeze out all excess moisture and crowd the particles so closely together as to preclude the entry of excess moisture. There is an optimum moisture content at which maximum compaction or density can be achieved. This must be determined through soil engineering procedures. Good drainage is essential to stable track.

Soils with a high organic content—swamp muck and muskeg—should be avoided or wasted, as should clays with a high montmorillonite mineral content and false shales. These soils have a tendency to swell when wet. If their use is necessary, they can be put on the side slopes of the fill where the swelling will do little harm or at the bottom of the fill where the weight of material above will restrain the swelling. Sensitive clays and silts, those which can liquefy under impact or shock, are also poor fill materials, as are the varved clays. Coarse granulars are best. Good soil engineering practice must be followed to have a stable subgrade. Railroads have not generally adopted the layered designs and analyses of highway engineering.

I. Roadbed Stabilization

A variety of stabilizing procedures are available for unstable subgrades. Foremost of those is subdrainage, to be discussed later. Other procedures include (1) mechanical support by piles and pole driving in the shoulders and across the track section under the ballast; (2) toe walls; (3) excavation and backfill with select material; (4) sand-filled spud and blast holes; (5) reinforced earth—granular soil in frictional contact with metallic strips on the back of precast concrete panels; (6) admixtures in the form of cement grout and lime injections; and (7) electric osmosis by passing a current between two electrodes driven into the subgrade, which causes moisture to seep into well-points (one of the electrodes) from which it is pumped away. Other methods used depend on the ingenuity of local forces.

J. Drainage

Adequate drainage is essential for stability, to prevent excess moisture in the subgrade soils, and to prevent washouts and derailments. Surface water must be diverted from the track structure by track ditches parallel to the track and just below the ballast section, by berm ditches at the top of the cut slopes, and by toe ditches at the foot of the subgrade fill. Water from side ditches and small streams is carried through the subgrade in culverts and small bridges.

1. Surface Drainage

Principles of hydrology and hydraulics are involved in the design of railroad drainage structures. The size of the opening in the structure must be sufficient for the quantity of water that will reach the subgrade from a drainage area. That quantity will be based on selecting a design storm and using its intensity, frequency, duration, and amount of runoff. The choice of a design storm is a safety and economic problem. Storms of 25–50 years frequency are often selected for main line routes and those with passenger service. Lesser storms that occur more frequently can be used for low-density branch lines, where less hazard is involved and traffic delays from washouts would not be serious.

The foregoing factors are contained in the equation $Q = AiR$, where Q is the quantity of water flowing from a drainage area in cubic feet per second, A is the drainage in acres, i is the rainfall intensity of the design storm in inches per hour, and R is the rate of runoff. Runoff will vary from 90% for steep, rocky slopes or paved and built-up areas to 50–90% for industrial areas and 20–40% for rail yards and cultivated fields.

The area of the opening is $A_c = QV$, where A_c is the bridge or culvert opening in square feet, Q is the quantity of flow as above, and V is the velocity of flow, sometimes taken as 10 ft^3/sec maximum.

2. Subdrainage

Vertical and lateral seepage and percolation introduce excess moisture into the subgrade. Corrugated and perforated metal pipes laid in trenches with a porous, granular backfill are among the various methods and materials used to drain large flat areas such as yards or to drain wet cuts or fills. Cross drains placed at the ends of wet cuts can reduce seepage into the adjoining fill area. Where a weak subgrade has caused the ballast to penetrate the soft materials to form a ballast pocket, a pipe laid transversely to the track beneath the pocket and backfilled with a sand or gravel will drain moisture and help keep the pocket dry.

K. Ballast

Ballast is a porous, granular material such as crushed rock, crushed slag, gravel, sand, or a similar material placed between the tie and the subgrade and around the tie. Ballast distributes the tie load to the subgrade, anchors the tie (and track) against movement, facilitates maintenance, provides immediate drainage for the track structure, and discourages vegetation.

1. Requirements

A ballast material must be hard, tough (nonshattering), noncementing, stable, sterile, and free from dirt, clay lumps, and friable pieces. It also must provide good drainage. It should be easily worked, readily available, and have a minimum overall economic cost. Some of these requirements are contradictory, hence, a need to determine the cost effectiveness of a particular ballast source.

Hard trap rocks, granite or basalt, are preferred materials, especially with concrete ties, but they are not cheap

TABLE I Recommended Ballast Gradations[a]

Size no.[b]	Nominal size, square opening	Percent passing									
		3 in.	$2\frac{1}{2}$ in.	2 in.	$1\frac{1}{2}$ in.	1 in.	$\frac{3}{4}$ in.	$\frac{1}{2}$ in.	$\frac{3}{8}$	No. 4	No. 8
24	$2\frac{1}{2}''$–$\frac{3}{4}''$	100	90–100	—	25–60	—	0–10	0–5	—	—	—
25	$2\frac{1}{2}''$–$\frac{3}{8}''$	100	80–100	60–85	50–70	25–50	—	5–20	0–10	0–3	—
3	$2''$–$1'$	—	100	95–100	35–70	0–15	—	0–5	—	—	—
4A	$2''$–$\frac{3}{4}''$	—	100	90–100	60–90	10–35	0–10	—	0–3	—	—
4	$1\frac{1}{2}''$–$\frac{3}{4}''$	—	—	100	90–100	20–55	0–15	—	0–5	—	—
5	$1''$–$\frac{3}{8}''$	—	—	—	100	90–100	40–75	15–35	0–15	0–5	—
57	$1''$–No. 4	—	—	—	100	95–100	—	25–60	—	0–10	0–5

[a] From American Railway Engineering Association. (1988). "Manual for Railway Engineering (Fixed Properties)," p. 1-2-6, Washington, DC.
[b] Gradation numbers 24, 25, 3, 4A, and 4 are main line ballast materials; gradation numbers 5 and 57 are yard ballast materials.

and may involve high transportation costs from available sources. Certain slags, such as copper, and some limestones are also good, but limestone and slag with high cementing characteristics should be avoided. Gravels, chat, sand, and similar materials find use in low-density lines.

2. Specifications

The AREA and the American Society for Testing and Materials (ASTM) have established specifications for size, that is, the percentages by sizes of the mix in each grade of ballast. The AREA No. 4 with 3/4 to $1\frac{1}{2}$ in. as the principal size range has widespread use, but the No. 24 with sizes up to $2\frac{1}{2}$ in. is often used for heavy haul lines and with concrete ties. Tables I and II show the AREA specifications.

Specifications are also related to weight and density, rate of abrasion, and durability (resistance to the freeze–thaw cycle). These are quality specifications, but, they have no direct bearing on the material's stability in track. One measure used by some engineers is the flakiness index, the ratio of length to width. Long slender particles give initial stability but tend to break and lose that stability. AREA specifications limit flat elongated particles to 5% of the ballast volume, defining them as having a length equal to or greater than five times the average width or thickness. One Canadian railroad limits such particles to 30% of the total.

The particle index (PI) has also been used as a measure of ballast stability and geometry based on angularity or sharpness (of edges), shape, and surface texture. With polished steel balls having a PI of zero, the range can vary from a PI of 5.0+ for smooth, rounded gravels to 20.0+ for crushed rock and slag. Unfortuantely, the rough edges that give stability tend to wear away and stability is lost.

TABLE II Recommended Limiting Values of Testing for Ballast Material[a]

Property	Ballast material							
	Granite	Traprock	Quartzite	Limestone	Dolomitic limestone	Blast furnace slag	Steel furnace slag	ASTM test
Percent material passing No. 200 sieve	1.0%	1.0%	1.0%	1.0%	1.0%	1.0%	1.0%	C117
Bulk specific gravity[b]	2.60	2.60	2.60	2.60	2.65	2.30	2.90	C127
Absorption percent	1.0	1.0	1.0	2.0	2.0	5.0	2.0	C127
Clay lumps and friable particles	0.5%	0.5%	0.5%	0.5%	0.5%	0.5%	0.5%	C142
Degradation	35%	25%	30%	30%	30%	40%	30%	—[c]
Soundness (sodium sulfate) 5 cycles	5.0%	5.0%	5.0%	5.0%	5.0%	5.0%	5.0%	C88
Flat and/or elongated particles	5.0%	5.0%	5.0%	5.0%	5.0%	5.0%	5.0%	USACE CRD-C119

[a] From American Railway Engineering Association (1988). "Manual for Railway Engineering (Fixed Properties)," p. 1-2-6, Washington, DC.
[b] The limit for bulk specific gravity is a minimum value. Limits for the remainder of the tests are maximum values.
[c] Materials having gradations containing particles retained on the 1 in. sieve shall be tested by ASTM C535. Materials having gradations with 100% passing the 1 in. sieve shall be tested by ASTM C131.

The best stability relation has been the British Standard 81-24 crushing index. The more crushing that occurs in a measured sample the less stable the material; that is, a low crushing value correlates with high stability.

3. Ballast Section

The ballast section (Fig. 9) brings the ballast to the top of the ties. Some roads keep it an inch or so below. The AREA recommends a 6-in. shoulder beyond the ends of the ties, but that is often increased to 12 in. for heavy haul service under 100-ton cars or where continuous welded rail or concrete ties are used. The sides extended on a 2:1 slope from shoulder to subgrade.

As noted earlier, the depth beneath should be sufficient to reduce the tie load to within the bearing capacity of the subgrade. A subgrade bearing capacity of 20 psi is considered normally firm. Such a subgrade under the loads imposed by a 30,000-lb wheel load would require 24 in. of ballast depth for load distribution. If, as often happens, the soil-bearing capacity is only 10 psi or less, at least 36 in. of ballast are required for distribution. Dr. Talbot's pressure studies, later confirmed by Dr. M. F. Salem, show that uniform pressure distribution occurs at a depth equal to the center-to-center tie spacing plus 3 in. or simply the center-to-center spacing for a well-compacted, firm subgrade. Both studies show that pressure distribution follows this pattern regardless of the type of ballast in use. Not all of the ballast depth need be of high-grade material. Twelve inches of gravel or a similar material may serve as a subballast for distribution purposes.

4. Compaction

Ballast must be well-compacted for stability. In new construction, ballast is applied in 4- to 6-in. layers. The first layer should be spread and compacted before any load is placed on the track to prevent the ties from being forced into the subgrade surface to form incipient water or ballast pockets.

Existing track is also tamped (surfaced) to remove deformations and when new ballast is added after cleaning, raising, sledding, or plowing operations. Tamping, once done by hand, is now performed by mechines that apply pressure and vibration electrically or hydraulically. Such machines automatically raise and tamp to an established grade and often include track lining capabilities. Machines are available to tamp ballast side slopes and to reproduce "train tamping" effects on new ballast to reduce slow-order delays.

5. Cleaning

Ballast loses resilience and drainage qualities as it becomes worn and dirty. Internal friction causes powdering and sometimes cementing of the particles and is a source of muddy track. Dirt also comes from sand used in braking or on heavy grades, from dust and droppings from cars carrying granular commodities, and from

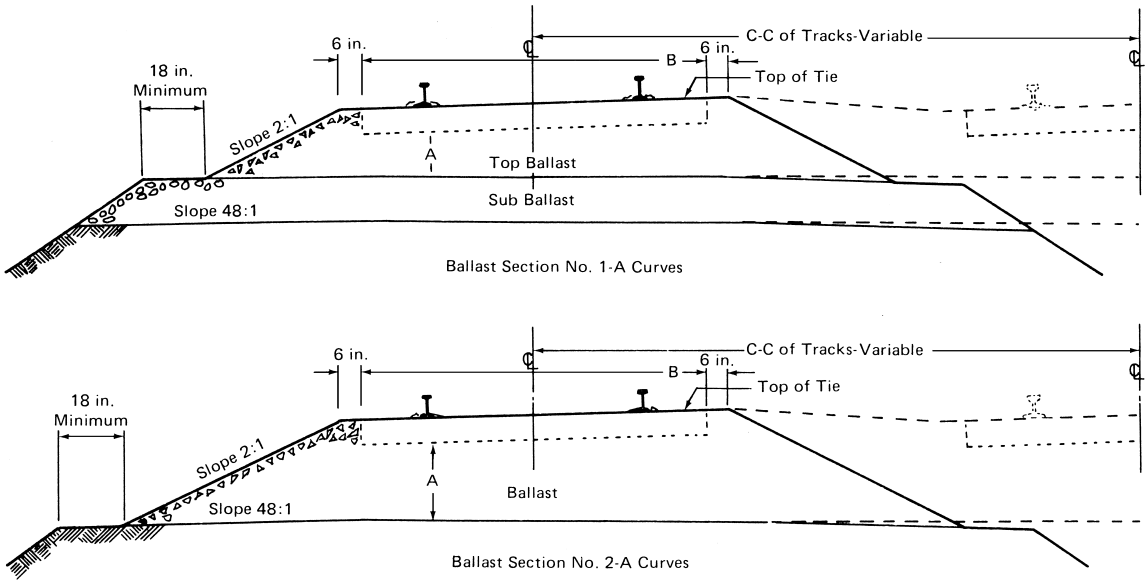


FIGURE 9 Recommended ballast sections. The depth of the ballast section used will depend on conditions peculiar to the railroad or location. Sections apply to all types of ballast and are for use with jointed or continuous rail. The top of the ballast is determined by use of various mechanized ballast distributing operations. [From American Railway Engineering Association (1978). "Manual for Railway Engineering (Fixed Properties)," p. 1-2-8, Washington, DC.]

windblown dust from freshly plowed fields. The track settles in soft spots and under joints. Permanent deformation occurs.

Crushed rock, slag, and some prepared gravels can be cleaned by machnies that remove the shoulder ballast and the intertrack material and carry it on elevating belts or conveyors to vibrating screens; the dirt is shaken out, and the clean ballast is returned to the track. Old ballast may be removed entirely, or pushed to the shoulder for cleaning, by any of several types of plows pulled by a winch or locomotive. Or the track may be raised, and the crib material between the ties leveled to form a subballast beneath a new ballast application.

Much ballast is now being cleaned by undercutting machines that pass a continuous belt, with digging buckets attached, under the track to dig all ballast and elevate it to vibrating screens for cleaning. The proper depth of cut is still under debate. Several roads cut 6–8 in., but depths of 12–18 in. are not uncommon. Another school says to dig until clean ballast is reached. Others argue against disturbing the compacted ballast any deeper than is necessary. It is a matter of costs and benefits, since the costs of undercutting rise sharply with the depth of cut.

6. Ballast–Subgrade Separation

The repetitively applied loads of passing wheels cause a patting action that brings moisture to the surface of fine-grained soils. As much as 1 in. or more of slurry can be formed at the ballast–subgrade interface. Ballast penetrates downward into the subgrade to form ballast pockets; subgrade particles infiltrate the ballast to form a muddy, pumping track. Several methods are at hand to prevent such intermingling of subgrade soils and ballast.

1. Use of a graded granular material as a filter blanket on top of the subgrade has been effective. The gradation must be such as to prevent infiltration of soil particles into the ballast, but coarse enough to permit drainage and keep ballast particles from the subgrade. Filter designs are available in soil engineering textbooks. An 8-in. filter blanket can also serve as a subballast.

2. A relatively recent innovation is the use of geotextiles laid between the ballast and subgrade. The geotextile fabric forms a physical barrier between the two (or between clean and dirty ballast), while permitting drain-age through the fabric and, with some types, a wick action that drains laterally from the interface to evaporation at the edge of the ballast section.

Geotextiles are nonwoven polyester or polypropylene fibers, heat bonded or interlocked mechanically by needle punching. Rolls vary from 300 to 1000 ft in length and 10 to 17 ft in width; they weigh 280–1200 lb. Thickness varies from 0.08 to 0.20 in. The size selected depends on soil type and condition and traffic type and density. The rolls are spread on top of a new subgrade or, with old track, inserted beneath a freshly plowed or undercut track before reballasting and lowering the track to its final position. There must be sufficient depth of cover, 8 in. or more, to prevent puncture by tamping machines. Some use has been made of a protecting layer of sand next to the fabric. Results with geotextiles have been generally good. Geotextiles are subject to puncture, tear, and deterioration from sunlight and certain chemicals. Their economic life is still undetermined for railroad use.

3. A lime-stabilizing layer of subgrade, 8 in. thick is another procedure that has been used to reduce the depth of ballast. About 5% of lime by weight is blended into the upper 8 in. of a compacted subgrade by disking and blading. Water to the optimum content is added and the mixture is rolled. After setting, the process is repeated once more before ballast is applied.

Another process blends cement, about 3% by weight, into the top layer of a clay subgrade. Water is added up to the optimum and the mixture is rolled. It sets to a hardness of 2000–3000 psi.

Hot asphalt, as both a binder in the ballast section and a covering material over the subgrade, is under study and test. For the ballast, the asphalt can be hot-mixed with the ballast or sprayed on the ballast in place. It serves as both a binder and a waterproofing cover. Its placement as a subgrade cover is much like laying an asphalt-bound macadam highway.

L. Crossties

Crossties perform the triple functions of load distribution, supporting the rail to gauge ($56\frac{1}{2}$ in. measured 5/8 in. below the gauge corner of the rail), and anchoring the track against movement in all three planes. Because the loads are applied under each rail, the ballast tends to compact more at those points until actual contact between tie and ballast is lost and the point of support moves toward the center between the rails. The track, now centerbound, wobbles slightly about the midtrack pivot. Some form of surfacing is required, preferably before the center-bound condition is reached.

1. Wood Ties

Main line wood ties are usually 7–8 in. thick, 8–9 in. wide, and $8\frac{1}{2}$–9 ft long. The larger sizes are used in heavy haul territory. Slightly smaller sizes are assigned to low-density branch lines, yards, and some sidings. The $8\frac{1}{2}$ ft tie is the optimum size in terms of bending stress, but the longer ties

offer better load distribution and lateral restraint. Specifications for size and condition of the wood are found in the AREA manual and in the standards of the trade associations. White oak, red oak, mixed hardwoods, and Southern pine are in common use. The oaks are preferred.

a. Life and treatment. Wood ties have an expected in-track life of 15–35 years, depending on specie of wood, initial preparation and preservative treatment, traffic loading, type of ballast, moisture conditions, and care in handling. A life of 30–35 years can be anticipated with 15–20 million gross tons of traffic annually, but only 15–20 years with 30 or more gross tons annually. Tie end splitting, which occurs during the drying process, can be restrained by inserting tie irons in the ends or by steel dowels driven diagonally from side to side. Splitting is also reduced by some form of vapor drying process that greatly shortens the drying period. Another process boils the tie in creosote oil for quick drying before treatment.

Ties are protected against insect attack and decay by a preservative process whereby, after drying, the wood cells are impregnated with a coal tar creosote and petroleum mixture or pentachlorophenol in an oil solution at about 200°F and 200 psi pressure. The earlier use of zinc chloride has been largely discontinued. Specifications for treating compounds are given in the AREA manual.

b. Armor. Ties are protected against crushing and shearing forces and abrasion by placing a steel tie plate having a bearing seat for the rail between the rail and the tie. Plate size should be based initially on the crushing strength of wood (~400 psi for hardwoods, ~250 psi for softwoods) to prevent plate cutting and crushing of the tie wood. For main line use, plates $7\frac{3}{4}$ in. wide range from 12 to 14 in. in length with 16- and 18-in. plates on heavy haul lines, especially on curves over 6°. Special designs of plates, but serving essentially the same purpose, are used on bridges, railroad crossings, and turnouts.

c. Movement. Differential movement between the plate and the tie increases tie wear, so the plate is held tightly to the tie by plate-holding spikes away from the rail (or by screw spikes or studs). The bearing area may be further protected by composition tie pads, sometimes glued to the ties, to reduce wear and moisture penetration. Excessive driving and redriving of spikes, as in track gauging or frequent rail renewals, "spike kills" the tie by destroying the wood fibers and their plate-holding power and permitting moisture to penetrate the wood interior.

d. Load distribution. By a further manipulation of the Talbot equations, the tie load is found to be $2Q_0 = 0.782PS/x_1$ and the load under one rail is $Q_0 = 0.391PS/x_1$, where P is the wheel load (static or dynamic as desired); S is the center-to-center tie spacing in inches; and x_1, as before, is the distance in inches from point of load application to point of zero bending. As a rule, for a $19\frac{1}{2}$-in. tie spacing, the wheel load carried by one tie is 40%, but it can vary from 38% for a $17\frac{1}{2}$-in. spacing to 56% for a $27\frac{1}{2}$-in. spacing.

The effect of tie spacing on track stiffness and the stiffness modulus may be taken as $u = 128A_b/S$, where u is the track stiffness modulus; A_b is the effective bearing area under each rail, usually one-third the tie face area; and S is the tie spacing in inches. When K, the track constant, is taken as $S \times u$, the track deflection $Y_0 = Q_0/K$ should not exceed 0.25 in., so that $Q_0 = 0.25K$.

e. Tie renewals. Two systems of tie renewals are practiced: spot renewals and out-of-face renewals. Spot renewals require a close semiannual inspection to identify the ties that will not serve through another 12 (or 24) months. The worn-out ties are removed by tie gangs or, in some instances, by local forces. Such renewals are frequently made on a 2-year basis to make large mechanized production gangs cost-effective. The gangs are equipped with machines to remove the ties in one piece or sawed into thirds and to insert new ties. Also included are spike pullers, spike driving machines, and tamper/liner machines to restore the disturbed track to good surface and line. The track is kept close to but never at a 100% tie condition.

In out-of-face renewals, all ties are removed over a given track segment at the end of a projected life regardless of the condition of individual ties. The period starts with all ties in 100% condition, but the track quality deteriorates as the renewal time approaches. Some life is lost from early removal of sound ties. In the United States, out-of-face renewals are usually restricted to tunnels, highway crossings, and station platforms, where the costs of frequent track openings become excessive.

2. Concrete Ties

Concrete ties have long been used in Europe, but extended use in the United States has come about only over the past 30 years and through the introduction of prestressing to resist failure in bending. The current tie design is a monoblock with a rail bearing seat and shoulders cast in the tie surface (Fig. 10) and with two rows of prestressed members, four strands per row, extending the full length of the tie. Specifications of the AREA call for 81,000 lb of prestressing tension; bonding length is 8–10 in. Deformed

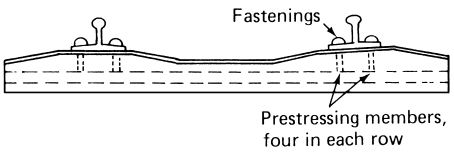


FIGURE 10 Concrete tie.

or chemically treated strands are used for better bonding. The wheel load distribution on ties of 60% is taken at a 30-in. spacing, decreasing to 45% for a 20-in. spacing. Usual spacings are 24–26 in. center to center. An impact factor of 150% is specified. Reinforcing rods under the rail give a flexural strength of 300,000 lb-in. and 200,000 lb-in. of negative center bending.

Rail fastenings have long been a problem. In the United States, rails are attached to the ties by direct fixation without intermediate tie plates. A threaded socket is cast in the tie into which some form of spring or lug bolt is inserted, or the holder for a steel coil or spring plate to bear on the rail base is cast in the tie. Bolts, lugs, clips, or coils bear on the base of the rail to give lateral restraint and longitudinal anchoring. The AREA specifies an insert depth of $2\frac{3}{4}$ in. with a pull-out resistance of 12,000 lb. Concrete ties weight 600–800 lb, compared to a wood tie weight of slightly over 200 lb.

With the cut spike and wood tie combination, the rail is free to rise from the tie plate bearing area and experience some reverse vertical bending ahead of the wheel. With concrete fastenings there is generally a rigid connection between the rail and the tie. Wave action in the rail can cause the tie to move upward with the rail, although the tie weight tends to act as a restraining force.

Some vertical freedom is had in some fasteners. A lug-type device permits $\frac{1}{8}$ to $\frac{1}{16}$ in. of clearance between the lug and the base of rail. Flexible spring clips with elastic washers on the unit bolts also allow some vertical play. Fasteners should exert a longitudinal restraint of 2400 lb or more on the base of the rail per tie to act as rail anchors.

Fasteners must also restrain rail rollover. Pads are placed between the base of the rail and the tie surface. A soft pad will compress under load, thereby reducing the anchoring force of a rigid fastening. A hard pad would be more suitable for heavy wheel loads. Soft pads with elastic clips for fasteners do well under moderate loads and high-speed passenger service. Amtrak has used a soft pad $6\frac{1}{2}$ mm thick. Hard pads, often 5 mm thick with elastic clips under heavy loads, restrain rail rollover while placing more strain on the ties and ballast. Not infrequently, tie pads crush or work loose.

No problems are experienced in placing concrete ties out of face; however, the stiffer modulus given by concrete ties makes for uneven track support if these are spotted among wood ties. Concrete ties should be installed in groups. Sound wood ties thus displaced can be salvaged for use elsewhere.

Concrete ties have found extensive use on the Florida East Coast Railroad, the Northeast Corridor, heavy haul portions of the Canadian railroads, and in rapid transit. Almost every railroad has one or more test sections laid with concrete ties, but an all-out trend is yet to come.

A different design of tie, the block tie, has found extensive use in France and other countries using French technology. The block tie consists of two concrete blocks with bearing seats for the rails with a steel tee section connecting and embedded in the blocks. A reinforcing coil surrounds each end of the tee section in the blocks, and reinforcing rods are placed near the bottom of the blocks.

3. Steel Ties

Steel ties, frequently found in Europe, Africa, and Asia, have little use in the United States. The wear and loosening of fastenings, corrosion, and difficulties in securing adequate insulation in track circuit territory have been major obstacles. Some effort is being directed toward their use in nonbonded branch lines, for temporary tracks, and for industrial tracks. The steel tie, being a hollow rolled section that fills with ballast, has the virtues of good load distribution and stability.

4. Reconstituted Ties

Goals of economy and conservation and difficulty in disposing of old ties have directed attention to the reuse of tie wood fibers in a reconstituted tie. The old ties are put through a rigorous cleaning and grinding process, blended with a proprietary resinous compound, and baked under pressure in a mold. Reinforcing rods are placed two at the top and two at the bottom. Production time is about 7 hr. Three old ties are required to produce two reconstituted ties. These ties, 7 in. $\times$ 9 in. $\times$ 9 ft in length, are somewhat heavier than an oak tie and not quite as strong. The estimated life, still unconfirmed, is 50 years. The future of this technology is still unclear.

5. Slab Track

Experiments have been in progress for several years on the design of a concrete or concrete–asphalt slab to which the rails are attached with fasteners similar to those used with concrete ties. One design uses a 6-in. cellular slab placed longitudinally under each rail laid on a concrete or asphaltic base (sometimes on crushed rock). The slabs are connected transversely every 5 or 6 ft with reinforced concrete connectors. Another design uses a thin shell 4 in. thick under the rail and 2 in. thick in the space between rails. This is laid on a compacted ballast bed.

M. Fastenings

The devices used to join rails together and hold them in place on the ties are known as fastenings and, for accounting and storekeeping purposes, as other track material (OTM). These devices include joint bars and track bolts for the bars, tie plates, track spikes, rail anchors, gauge rods, and compromise bars to join rails of dissimilar weight or section, as well as turnouts and crossings.

1. Track Spikes

The conventional cut spike with a rectangular and pointed shank and a head for driving and withdrawing is almost universally used with wood ties. The spikes have little or no hold-down power, but hold the rails to gauge and prevent rollover. Spikes range from $5\frac{1}{2}$ to 6 in. in length, $\frac{9}{16}$ to $\frac{5}{8}$ in. in thickness, and have a head $1\frac{1}{4}$ by $1\frac{5}{16}$ in. wide and $1\frac{3}{16}$ to $1\frac{1}{2}$ in. long.

Screw spikes have circular heads to bear on the rail about 2 in. in diameter, with a threaded shank $5\frac{13}{16}$ in. long and a blunt end. Holes must be bored in the ties, and the spikes screwed into place. A $\frac{3}{4} \times \frac{3}{4}$-in. head above the circular portion receives the wrench. Screw spikes, often used for plate holding, have good initial restraint and hold-down capabilities, but they lose these with the passage of traffic and require more time and equipment for their use than do cut spikes.

2. Tie Plugs

When a spike is withdrawn, a hole is left in the tie that could permit moisture to enter and cause decay if left open. Tie plugs, usually of treated wood and 4 or 5 in. long, are driven into the spike holes. Even if a spike is redriven in the same hole, the tie plug improves the friction grip on the shank. A plastic compound recently on the market can be squeezed into the spike hole to perform the same function as the plug.

3. Joint Bars

Unless continuously welded, rails are joined together by pairs of joint bars, one bar on each side of the rail, that fit between the underside of the head and the top of the base. The bars may have four holes, but more often today have six holes to receive the joint bar bolts that pass through both joint bars and the rail. Joint bars must not only hold the rails together with perfect gauge corner alignment, but also provide girder strength in bending equivalent to the rail section. Specifications for joint bars and their manufacture are found in the AREA manual. In track circuit territory, the circuits are electrically insulated by insulated joints that have insulating pads, thimbles, and end posts or an insulating coating to isolate each circuit.

4. Rail Anchors

Rail anchors, attached with a friction grip to the base of the rail, are used to transmit longitudinal forces from the rail to the tie and thence to the ballast. The wave action in rails under passing wheels and thermal expansion and contraction cause the rails to "run," that is, move longitudinally. Rail anchors restrain that movement, especially with continuous welded rail. The anchor is usually made in one piece and is forced over the base of the rail with a special wrench or driven on with a hammer. The anchor has a flange or thickness deep enough to bear against the tie. Anchors should have a restraining force of 2400 lb.

5. Other OTM Items

Gauge rods are attached to the base of each rail on sharp curves to assist in holding the gauge. Derails are used on side tracks to derail cars that might otherwise roll onto a main track. Their use with railroad crossings at grade has been largely discontinued except where a minor branch line crosses a high-density line. Derails at the ends of passing sidings are almost obsolete.

V. RAIL

A. Rail Designations

The prime element in the railroad track structure is rail rolled from steel having the characteristics and properties shown in Tables III and IV. The steel may be of open hearth, basic oxygen, or electric furnace origin. Rails are usually rolled in 39-ft lengths. Newer mills now roll lengths of 80–85 ft. Rails are designated by section and weight in pounds per yard, ranging from 60 lb/yard or even less to 132, 140, and even 152 lb/yard. The lighter rails are seldom rolled today except for mining, logging, and industrial tracks or for export, but they are still to be found in low-density branch lines and industry tracks, often handed down from main line use. The AREA manual recommends weights of 90, 100, 115, 119, 132, 133, 136, and 140 lb. The heavier rails are used in high-density and heavy haul lines and for high-speed territories. Lighter rails are in low-density lines and rapid transit.

Various agencies have contributed to rail design—the ASCE rails by the American Society of Civil Engineers, the RA rails by the American Railway Association, and the RE sections by the American Railway Engineering Association. The New York Central Railroad originated the Dudley section, the Pennsylvania Railroad originated the PS group, and the Colorado Fuel and Iron Steel

TABLE III Chemical Composition in Weight Percent for Steel Rail[a]

Element	Chemical analysis weight percent nominal weight lb/yard		Product analysis weight percent allowance beyond limits of specified chemical analysis	
	90–114	115 and over	Under min.	Over max.
Carbon	0.67–0.80	0.72–0.82	0.04	0.04
Manganese	0.70–1.00	0.80–1.10[b]	0.06	0.06
Phosphorus, max.	0.035	0.035	—	0.008
Sulfur, max.	0.037	0.037	—	0.008
Silicon	0.10–0.50	0.10–0.50	0.02	0.02[c]

[a] From American Railway Engineering Association (1998). "Manual for Railway Engineering (Fixed Properties)," p. 4-2-1, Washington, DC.

[b] The upper manganese limit may be extended to 1.25% by the manufacturers to meet the hardness specifications. When the manganese exceeds 1.10%, the residual alloy contents will be held to 0.25% max. Ni, 0.25% max. Cr, 0.10% max. Mo., and 0.03% max. V.

[c] Product analysis for continuously cast steel shall be 0.05% over the maximum limit for silicon.

Corporation originated the CF & I rails. All designs are variations of a girder with differences in weight and weight distribution, width of head, head curvature radius, height, base width, web thickness, and fillet corner radii.

Rails are also designated by grade at the mill. No. 1 rails are fully acceptable. No. 2 rails are also acceptable but contain slight surface imperfections. "A" rails come from the upper end of the ingot, where inclusions and blowholes are most likely to occur. Their use in so-called hazardous locations has been restricted on some railroads, but with improved manufacturing procedures, such restrictions are few. High-carbon rails are those in the upper five points of the carbon content range. Special metallurgy rails are designated by the alloy material—high-silicon, chrome–vanadium, or chrome–molybdenum. Almost all main line rails of today are control cooled to prevent the formation of shatter cracks and transverse fissure developing therefrom. Rails may also be heat treated. One process treats the entire rail; a second, by electric induction, treats only the head.

B. Rail Design

Rails are designed as girders with a flat base; a narrow, deep head; and a slight curve across the head to rounded gauge corners. Stiffness varies with the square of the weight but with the cube of the height, so modern rails for heavy traffic emphasize height to increase stiffness and reduce deflection. Characteristics of rails in common use are given in Table V.

TABLE IV Hardness Properties[a]

	Standard rail		
	90–114 lb/yd	115 and over lb/yd	High strength rail
Brinell hardness	248 min.	285 min.	341–388

[a] From American Railway Engineering Association (1988). "Manual for Railway Engineering (Fixed Properties)," p. 4-2-2, Washington, DC.

C. Rail Life

Safety and economy are closely allied in rail life. Rail life, usually measured in millions of gross tons carried, has averaged about 600 million gross tons for conventional traffic in jointed main line tracks, 400 million gross tons for heavy haul movements, and 1000 million gross tons for general traffic on continuous welded rail.

Rail life is dependent on rail steel, traffic loading (wheel load, speed, density), incidence of defects, condition of support, and whether jointed or continuous welded. An equation frequently used is $L = KWD^{0.565}$, where L is the rail life in millions of gross tons, W is the weight of rail in pounds per yard, D is the traffic density in millions of gross tons per year, and K is a "level of maintenance" factor, sometimes taken as 0.545. It can also be computed if L is known from experience or projected wear.

The life of jointed rail is limited by the batter of rail ends at the joints, by side wear on curves, and by vertical head wear.

Head wear that permits the wheel flanges to touch the joint bars is considered a condemning wear by some railroads. Others consider a 25% loss of head metal to be condemning. An average wear of 0.03 in./100 million gross tons has been stuggested in AREA studies.

Battered rail ends can be rebuilt by gas torch or electric welding and grinding to correct contour. Longitudinal flow and rail end chipping can be reduced by

TABLE V Properties of Typical Rail Sections[a]

Section[b]	Weight per yard (lb)	Moment of inertia	Base to neutral axis (in.)	Section modulus (base)	Area (in.^{2})	Height (in.)	Base width (in.)	Head width (in.)	Head radius (in.)
133 RE	133.4	86.0	3.20	27.0	13.08	$7\frac{1}{16}$	6	3	10
132 RE	132.1	88.2	3.20	27.6	12.95	$7\frac{1}{8}$	6	3	10
131 RE	131.2	89.0	3.20	27.0	12.90	$7\frac{1}{8}$	6	3	14
130 RE	130.0	77.4	3.03	25.5	12.71	$6\frac{3}{4}$	6	$2\frac{15}{16}$	14
120 RE[c]	120.9	67.6	2.92	23.1	11.85	$6\frac{1}{2}$	$5\frac{3}{4}$	$2\frac{7}{8}$	14
115 RE	114.7	65.6	2.98	22.0	11.25	$6\frac{5}{8}$	$5\frac{1}{2}$	$2\frac{23}{32}$	10
112 RE	112.3	65.5	3.00	21.8	11.01	$6\frac{5}{8}$	$5\frac{1}{2}$	$2\frac{23}{32}$	14
110 RE[c]	110.4	57.0	2.83	20.1	10.82	$6\frac{1}{4}$	$5\frac{1}{2}$	$2\frac{25}{32}$	14
100 RE	101.5	49.0	2.75	17.8	9.95	6	$5\frac{3}{8}$	$2\frac{11}{16}$	14
90 RA A	90.0	38.7	2.54	15.23	8.82	$5\frac{5}{8}$	$5\frac{1}{8}$	$2\frac{9}{16}$	14
75 AS	75.2	23.0	2.36	10.0	7.38	$4\frac{13}{16}$	$4\frac{13}{16}$	$2\frac{15}{32}$	12
106 CF&I	106.6	53.6	2.85	18.8	10.45	$6\frac{3}{16}$	$5\frac{1}{2}$	$2\frac{21}{32}$	14
119 CF&I	118.8	71.4	3.12	22.9	11.65	$6\frac{13}{16}$	$5\frac{1}{2}$	$2\frac{21}{32}$	14
136 CF&I	136.2	94.9	3.35	28.3	13.35	$7\frac{5}{16}$	6	$2\frac{15}{16}$	14
155 PS	155.0	129.0	3.38	36.7	15.20	8	$6\frac{3}{4}$	3	—
152 PS	152.0	130.0	3.50	37.0	14.90	8	$6\frac{3}{4}$	3	—
140 PS	140.6	97.0	3.37	29.0	13.80	$7\frac{5}{16}$	6	3	10

[a] From Hay, W. W. (1982). "Railroad Engineering," rev. ed., p. 494, Wiley, New York.

[b] RE, American Railway Engineering Association design; RA, American Railway Association (forerunner of AAR) design; CF&I, Colorado Fuel & Iron Steel Corp. design; PS, Pennsylvania Railroad design; AS, American Society of Civil Engineers design.

[c] Not presently being rolled.

slotting or grinding a slight bevel on the ends of the rails.

Whether rails are jointed or welded, the incidence of rail failures often determines the limit to rail life, that is, the number of defective rails per mile of track per year as the accumulated gross tonnage increases. When the failure rate reaches an established value, the rail is removed from the track. The condemning value, a management decision, will range from 1.25 to 2.0 or more, sometimes simply until the situation becomes intolerable.

D. Wear versus Fatigue

Studies by Zarembski and the Association of American Railroads indicate that wheel loads have a marked effect on rail life and on the incidence of defects. As seen in Fig. 11, rail life for wheel loads up to 26,000 lb is largely determined by wear and, with CWR, can provide a service life of 1000–1300 million gross tons. Wear includes batter, abrasion, and associated defects. With wheel loads over 26,000 lb, fatigue is the controlling factor. A wheel load of 33,000 lb (that of a 100-ton car) tends to limit service life to ~350 million gross tons. Fatigue failures include those associated with contact stresses between the wheel and the rail such as shelling, head checking, spalling, flaking, and transverse defects that develop from these.

E. Rail Defects

Rail defects arise from many causes—service wear, fatigue and contact stresses, rolling mill practices, and heavy wheel loads.

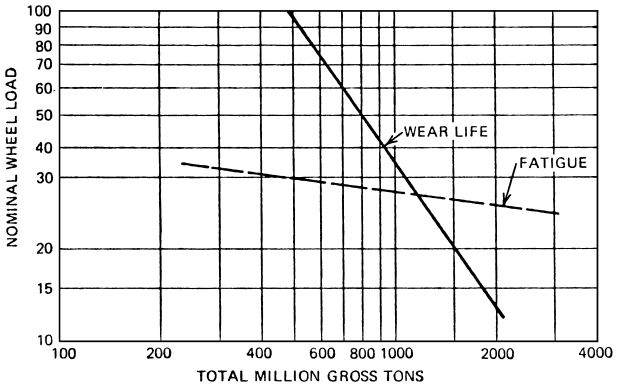


FIGURE 11 Fatigue life and wear life for various wheel loadings. The ordinate represents the largest static wheel load in a given mix of traffic consisting of a range of wheel loads. For the 100-ton axle load mix about 10% of the traffic was actually 100-ton cars. [From Zarembski, A. M. (1979). "The effects of rail section and traffic on rail fatigue life." *Am. Railw. Eng. Assoc. Bull. No. 673* **80,** 514–528.]

1. Service Wear

Abrasion, rail end batter, bolt hole breaks, and corrosion are typical of service defects. Abrasion on curves can be lessened by curve oilers, pots of graphite grease or a similar material, that pump a small amount of lubricant onto the gauge corner of the rail as passing wheels depress a movable ramp or treadle. Tests are underway on devices mounted on the locomotive or on oiler cars that apply, even on tangents, lubrication from the rail vehicle. A significant reduction in train resistance results as well as reduction in wear, but there can be an increase in fatigue failures. Waste oil is sometimes sprayed on rail to reduce corrosion from salt air, chemical fumes, or the weak sulfuric acid formed by gases in tunnels.

2. Mill Process Defects

Any impurities in the molten steel such as slag at the top of the ingot or blowholes (the "A" rail situation) can be the source of a transverse defect that grows and breaks under traffic. Blowholes from hot torn steel have the same effect. Transverse defects are particularly serious because they usually originate within the rail and, unless special equipment is used, are not detected until the rail breaks. Loss of correct rolling temperature or other sources of incomplete "welding" can lead to seams, segregations, and splits in the metal and cause head and web separations, vertical split heads, pipes, and horizontal fissures. Refinements in the manufacturing process are working to reduce these types of failures. Vacuum degassing and continuous casting that goes directly to the bloom stage, omitting the casting of ingots, are examples of progress.

For many years shatter cracks caused by hydrogen gas entrapped in the rail during cooling caused transverse fissures that grew under traffic to cause rail breakage and serious accidents. Controlled cooling has largely eliminated this problem. Rails are placed in an insulated cooling box before cooling below 700°F and left there until their temperature has dropped to 300°F before removal.

3. Fatigue-Related Defects

This group of defects, due largely to the advent of 100-ton cars, arise from high contact stresses between the wheel and the rail. Head checks have a scalelike appearance on the gauge corner of the rail. Head checks can represent a phase of the "wearing in" process or a prelude to spalling. Spalling is the breaking out of large chunks of metal from the rail head due to heavy wheel loads and to brittleness of metal. Flaking is a breaking out of small pieces of metal about the size of a fingernail from the running surface of the rail. These give the rail a poor appearance but pose no hazards. Shelling defects, which occur mostly but not always on the high rail of curves, are also fatigue induced. Shells appear as a crack or separation of gauge corner metal that drops slightly from wheel contact and, collecting grease, looks like a black blister. The crack may originate within the rail or may work in from the outside as rail metal flows under heavy pressure. Shells represent an accumulation of residual stresses that are relieved by crack formation.

Shells began to appear in numbers following the introduction of curve oilers, which preserved the gauge corner metal and allowed stresses to accumulate. Without oilers, the gauge corner metal wore away so that stresses could not accumulate.

The shell itself is not hazardous, but it can develop a transverse component leading to a transverse defect, a detail or progressive fracture. Wheel loads above 26,000 lb exceed the shearing strength of the metal (usually taken as 50,000 psi) and are prime contributors to shells. Shelling and transverse defect cracks appear about $\frac{3}{8}$ in. below the rail surface, where shearing stresses are a maximum.

F. Contact and Shearing Stresses

The Hertzian rolling cylinder approach is used to compute shearing stresses in the rail head. A simplified formula based on the theory of two rolling cylinders (the wheel and the curved rail head) in contact is the following:

$$\tau = \frac{23{,}500P^{1/3}}{2(R_1/R_2)^{0.271}(R_2)^{2/3}},$$

where τ is the contact stress in pounds per square inch, which should not exceed 50,000 psi; P is the dynamic wheel load in pounds; R_1 is the radius of the larger cylinder in inches; and R_2 is the radius of the smaller cylinder in inches.

This equation is admittedly of limited accuracy. It has only two parameters: the wheel load and the radii of the two cylinders. It does not take into account orthogonal forces, and the metal in the rail and the wheel is certainly not isotropic and homogeneous as required by strict adherence to the Hertzian theory. Nevertheless, it can be useful.

One measure of load severity is the wheel load divided by the wheel diameter, since larger wheels spread the load over a greater area. This measure indicates a value of 800 lb of wheel load per inch of wheel diameter for a 33-in. wheel, 810 lb for a 36-in. wheel, 820 lb for a 38-in. wheel, and 825 for the conventional 40-in. locomotive driving wheel. These limits, while desirable, are not enforced by the AAR Rules for Car Interchange. They are generally exceeded by the 100-ton car now favored for heavy haul and general use. There is a requirement that cars of 70 tons or over

TABLE VI Contact Stresses for 30,000-lb Wheel Load[a, b]

Wheel–rail combination	P_{dyn} (lb)	R_1 (lb)	R_2 (lb)	T (psi)
New wheel–new rail	36,900	18.0	10.0	71,973
New wheel–old rail	36,900	18.0	11.5	67,887
Old wheel–new rail	36,900	24.3 (R')	18.0	52,200
Old wheel–old rail	36,900	35.5 (R')	18.0	47,297

[a] At 30 mph on a 132-lb rail with 36-in. wheels.
[b] From Hay, W. W. (1982). "Railroad Engineering," p. 517, Wiley, New York.

built since 1961 have 36-in. wheels to reduce load on the bearings.

A modification is required for worn wheel and rail conditions. Wheels tend to develop a concave pattern in the tread with a constant radius of ~17 in., and the rail head wears to a constant ~11.5 in. radius. The modification, in effect, "extends" the worn rail radius to conform to the radius of concavity in the wheel tread. To make this revision, the R_1 or larger radius is set equal R', whereby $R_1 = R' = R_h R_r/(R_h - R_r)$, where R_h is the hollow worn radius of the wheel and R_r is the radius, new or worn, of the rail head.

Contact stresses can have values for four possible situations: new wheel on new rail, new wheel on old rail, old wheel on new rail, and old wheel on old rail. The last combination gives the lowest stress value because of the greater area of wheel–rail contact and lower stress distribution; see Table VI.

G. Corrugated Rail

A special problem encountered by heavy haul railroads is corrugations on the running surface of the rails. Corrugations occur as surface undulations with depths of 0.02–0.35 and even 0.05 in. and with crests 2–60 in. apart. "Roaring track" with crests 2–3 in. apart occurs under high-speed passenger trains. Under heavy haul conditions, crests of 8–12 in. are common and, at high speeds, crests of 12–60 in. can be found. Corrugations are formed on either the high or the low rail of curves, but primarily on the high rail; the low rail experiences crushing and flow. Corrugations also occur under heavy wheel loads on frozen roadbeds, especially with the continuous welded rail, where there are no joints to interrput rail vibrations. The only practical correction available is running a grinding train to grind about 0.01 in. off of the crests of the corrugations at a train speed of $1\frac{1}{2}$ mph or running a shaping train that mills the rail head to something approaching its original contour. Corrugations can be damaging to both the rail and the roling stock; the wheel–rail impact can sometimes be increased 100%.

Numerous solutions to the contact stress problem have been considered—revised wheel and rail contours to obtain a larger wheel–rail contact area, high-cant tie plates (1:14 versus 1:40) for the same purpose; lighter wheel loads (a problem in car design and car loading), and the use of tougher steels obtained by alloying with silicon, chrome, manganese, or molybdenum and heat treatment.

H. Rail Detector Cars

The detection of internal and other defects by rail detector cars is a standard feature of track maintenance. Search coils on the detector car are passed over the rails to monitor any variations in the inductive field or residual magnetic field surrounding the rail. Such variations, indicative of a defect, are recorded on a tape that gives continuous rail-by-rail record. Hand tests measuring variations in current flow at the indicated defect give a more nearly precise location of the defect and its magnitude. Quartz crystal transducers are used by some systems to give an audio reflection of the vibrations from sound waves passed through the rails. By properly positioning the crystals, a test can be made throughout the rail and through the joint bar area, where the presence of bolts and bars pose difficulties for the inductive and magnetic systems. Defective rails can then be removed or protected by slow orders, or other precautions can be taken.

I. Continuous Welded Rail

The problems of rail end batter, joint ties and joint support, joint bars and bolt tightening, and numerous other cost items virtually disappear when the joint is eliminated by welding the 39-ft rails into long continuous lengths. This is accomplished by one of three methods: Thermit welding, oxyacetylene pressure butt welding, or electric flash butt welding. The Thermit process, the oldest, is now used principally to join in the field strings of electric flash butt-welded rails, usually 1440 ft long. The 1440-ft strings are welded at a centrally located welding plant and hauled on a rail train to the installation site. The long strings are then field welded as installed into lengths of many miles, interrupted only by insulated joints for automatic signals and grade crossing protective devices. Welding has been successfully carried through turnouts.

1. Expansion Control

The problem of controlling expansion and contraction is overcome by a liberal use of rail anchors to transfer thermal stresses from the rails to the ties and into the ballast. Using the unit expansion value for steel and the modulus

of elasticity of rail steel, the thermal stress is found from the relation $0.00000065\,\Delta t = S/30 \times 10^6$, where Δt is the temperature change in degrees Fahrenheit and S is the thermal stress, which becomes 195 psi for a 1° change in temperature. The total force to be restrained is $195 \times \Delta t \times A$, where A is the cross-sectional area of the rail.

The restraining force per tie per rail has been given as 500–2000 lb (the latter with a frozen ballast section). A value of 500–800 will do for a 22-in. tie spacing. The number of ties to be box anchored (a rail anchor on each side of the tie at each rail) can be computed. Usually six 39-ft rail lengths at each end of the long welded rail are box anchored. The same nuber are box anchored at approaches to short open deck bridges, where rail anchors cannot be used. For long open deck bridges some form of expansion joint is used. Similar anchoring is required where tracks cross at grade. Joint restraint is usually ignored because of its varying amount.

The general procedure is to box anchor every tie for six to eight rail lengths beyond the joint at each end of the long rail, then box anchor every other tie (or every third tie) between to cover possible rail breakage and normal running. Some movement will occur at the ends of the long strings, because full restraint is not developed until the entire six boxed rails are involved. That movement can be calculated from the expression $M = NTL/2AE$, where M is the rail end movement in inches, T is the restraining force per tie per rail in pounds, L is the length in inches for the six rail lengths to secure full restraint ($L = Ns$, where N is the number of ties and s is the tie spacing in inches), A is the cross-sectional area of the rail in square inches, and E is the modulus of rail steel, equal to 30×10^6.

2. Adjustment for Temperature

To keep equal the forces of expansion and contraction, rail should be laid at the median temperature, which for a 0–120°F range would be 60°F. But rail must be laid under a variety of changing temperature conditions. An adjustment is made in the rail length to correct for actual laying temperatures.

For example, using the foregoing temperature range, if the rail is being laid at 80°F, it should be shortened to what it would be at 60°F. The change per foot of length for the 20° difference would be 20×0.0000065 or 0.00013 in./ft. The total required reduction in length should be $1440 \times 0.00013 \times 12$ or 2.25 in. The rail is made to contract this amount by applying cold water or cold air. The rail anchors are installed as soon as the shortening has been accomplished. If the rail were laid at a lower temperature than the median, it would be lengthened by heating, probably with a gas torch or special burner. Vibrating the rail helps to overcome resistance in sliding across the tie plates. Rather than aiming at an exact adjustment, most roads stay within 10%, while others stay 10–15% above to avoid the danger of sun kinks.

It is important to note that the temperatures in question, measured with a rail thermometer, are rail steel temperatures and not air temperatures. These are taken at least four times during the workday as the temperatures change.

3. Other Field Procedures

A full ballast section with a 6- to 12-in. shoulder beyond the tie ends is required to withstand thermal forces in CWR and lateral and vertical displacement or sun kink. Surfacing and tie renewals should not be undertaken when the rail is in compression (trying to expand), or at least only a few feet of track should have ties and ballast disturbed at a time. When a sun kink occurs, the rail is shrunk by cooling or a portion is cut out and a field weld or joint bars is applied. Pull-aparts (usually in cold weather) are repaired by applying joint bars, by heating and expanding, or by cutting in a short section of rail (10+ ft) held by joint bars or a Thermit weld.

VI. TURNOUTS AND CROSSINGS

A. Turnouts

Turnouts are a part of the track structure that permits diversions from one track to another. They are used in crossovers from one track to another, from main tracks to sidings and side tracks, and from ladder tracks to body tracks in yards.

Turnouts include a switch that initiates the diversion, a frog and guard rail assembly to permit wheel flanges to cross intervening rails, and closure rails that connect the switch with the frog (Fig. 12a). Turnouts are variously designated by the frog number, the frog angle, the length of switch rails, or the lead distance—from point of switch to point to frog.

The commonly used frog number is the ration between unit spread of the frog legs and distance to the frog point from which the unit spread is taken (Fig. 12b). The relation between the frog number and the frog angle comes from the equation $N = \frac{1}{2}\cot F/2$, where N is the frog number and F is the frog angle. The theoretical point of frog, being a knife edge, would batter and break, so it is cut back to where the spread is $\frac{1}{2}$ in. This is termed the half-inch or actual point of frog.

Frog numbers 12–20 find use in main line tracks. Numbers 10 and smaller are used in sidings, yards, branch lines, and industry tracks. Rolled section rigid frogs are made from rail sections sheared, planed, and bolted together.

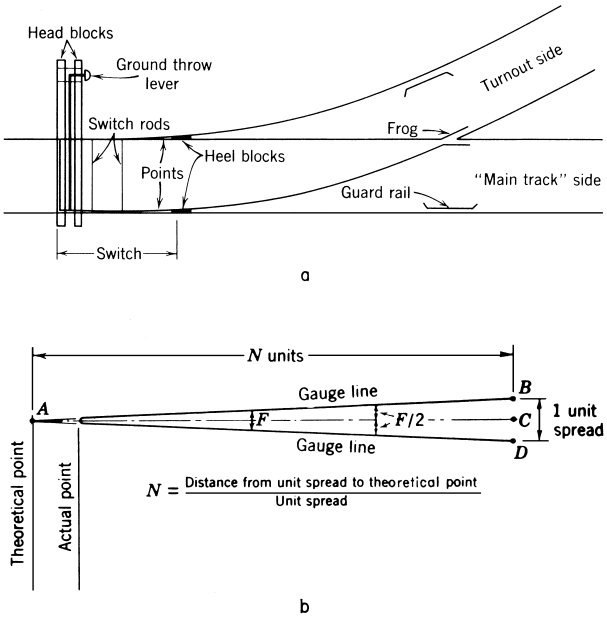


FIGURE 12 (a) Split switch turnout. (b) Frog number and angle. [From Hay, W. W. (1982). "Railroad Engineering," Wiley, New York.]

Rail-bound frogs have a cast manganese center containing the throat and point held in place by bolted rail sections that form the legs. Spring frogs have a movable wing or throat section that is held closed by a spring to provide solid support for wheels making a main line movement. The wing can be pushed open by wheel flanges in a diversion move.

With high-numbered frogs, the diversion angle F is very small and the gap the wheel must jump from throat to frog point is large, with batter and possible derailment potential. Movable point frogs are then used, hand or machine operated, to provide a solid support for the wheel.

Guard rails, set opposite the frog point, reduce impact and batter on the point and throat area and guide the wheels in the proper path. The inner face of the guard rail must be set 4 ft, $6\frac{1}{2}$ in. from the gauge corner of the frog point to prevent the wheel flange from striking the guard rail and to ensure a flangeway opening of $1\frac{7}{8}$ in. between the guard rail and the running rail.

Based primarily on experience, it is held that a smooth ride and diversion occur when the diversion or switch angle made by the switch point with the stock rail is one-fourth the frog angle or $s = F/4$, where s is the switch angle. The switch point is pivoted at the heel adjoining the closure or connecting curve to permit the left–right movement of the switch points. The heel spread, gauge of stock rail to gauge of switch point rail, is usually $6\frac{1}{4}$ in.

Rather than having a multitude of switch point sizes, a few are selected for manufacture: 11 ft for frog numbers 4, 5, and 6; 16 ft, 6 in. for frog numbers 8, 9, or 10; 22 ft for frog numbers 11, 12, and 14; and 30 ft for frog numbers 15, 18, and 20. Longer points and flatter frog angles can be found in special installations.

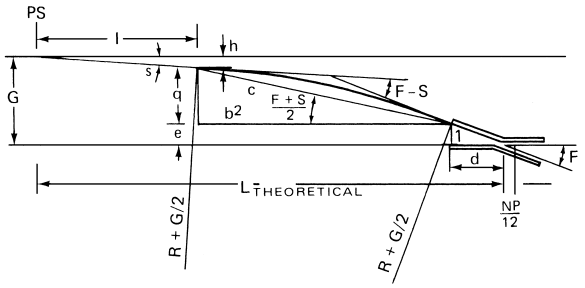


FIGURE 13 Turnout design. [After Pickels, G. W. and Wiley, C. C. (1949). "Route Surveying." Wiley, New York.]

Figure 13 shows the turnout design procedure. The lead distance from point of switch to actual point of frog is $l + b + d + Np'/12$, where l is the length of switch, equal to $\sin s/(h - t)$; t is the thickness of switch point at point of switch, $\frac{1}{8}$ to $\frac{1}{4}$ in.; h is the heel spread, usually taken as $6\frac{1}{4}$ in.; c is the long chord of the connecting curve; b is the horizontal projection of the long chord; d is the horizontal projection of the frog leg; F is the frog angle; f is the frog leg or toe distance; G is the track gauge, $56\frac{1}{2}$ in. for standard gauge track; $q = G - h - e$; R is the radius of the centerline of the connecting curve; $G/2 + R$ is the radius of the outer rail of the turnout curve; and $R - G/2$ is the radius of the inner rail of the connecting curve.

B. Crossings

Railroad crossings have two end frogs similar to those used in turnouts and two center frogs, unless the tracks cross at 90°, when all four frogs are identical. Crossing frogs can be of the rail-bound type, manganese steel inserts, or heat treated, bolted rigid. The prime requisite for crossings is a strong, stable subgrade; a good depth of high-grade ballast, perhaps a geotextile spread; and excellent drainage. If the crossing angle is very flat, movable points are used for the center frogs. Crossings normally permit movement only in the directions of the intersecting tracks, but when one or more turnouts are superimposed, a single or double slip or "puzzle" switch is obtained that permits multiroutes through the structure.

C. Rail–Highway Crossings

Highway grade crossings have advanced from timber flangeways with gravel or cinders between the timbers to all-timber or plank crossings, asphalt, concrete slab, steel plate, and, more recently, rubberized and composition plates. A stable subgrade and excellent drainage are fundamental needs. The approach slope should not be

steep, visibility should be clear in all four quadrants, and the crossing itself should be as wide as or slightly wider than the highway approach. All crossings are protected by signs. For principal highways and streets, protection in the form of signs, watchmen, crossing gates, illumination, and flashing lights have had a place. A successful device is the automatic short-arm crossing gate, which protects the approach lane yet leaves room for the motorist to escape around it if caught in its descent. These gates are usually also protected with flashing lights.

VII. TRACK–TRAIN DYNAMICS

A. Track–Train Relations

The track and trains running over it form a closely related system. Well-maintained track geometry in terms of profile, gauge, curvature, superelevation on curves, and cross level on tangents is essential to safe operation and the avoidance of excessive wear on track materials and rolling stock. Cars should be designed to run safely over any track regardless of the track condition.

B. Car Design

The conventional car consists of a steel center sill running the length of the car, to which are attached transverse members or floor beams to hold the car body. The car assembly is supported at each end of four-wheeled car trucks and connected to the car body by a center plate on the trucks and a center pin, fitting into the plate from the car body, that permits the cars to swivel on curves. The coupling devices and shock-absorbing draft gear are mounted at the ends of the center sill. The train line, air brake equipment, and brake rigging are mounted under the car body or, with some hopper car designs, at the end of the car body under the slope of the hopper.

Particular attention has been given to improving the safety and strength of cars carrying hazardous materials in a derailment. Strong outer shells and end shields and so-called shelf couplers to guard against coupler puncture in the event of derailment are among these improvements.

The conventional three-piece truck includes two side frames set over the journal boxes that provide bearings for the axles. Each frame also contains a nest of springs. The side frames are connected transversely by the truck bolster, which holds the center plate and side bearings. The center pin attached to the car body bolster fits into the center plate.

Not all cars follow this basic design. Four-wheeled cars with single-axle trucks have been introduced in trailer-on-flatcar service (piggyback), and tank cars and hopper cars have been designed with no center sills. The car body rests on stub end sills over the trucks and longitudinal strength is obtained through the car body.

Other trailer-on-flatcar units have been designed to operate in 10-car units for hauling highway trailers, and another design stacks container boxes two tiers high on one car. Covered rack cars carry automobiles on two or three levels, protected from the elements and vandalism.

C. Weight

A major problem today is simply one of car weight. This has already been noted in connection with the adverse effects on rail life of the 100-ton car, which, when loaded, places a weight of approximately 264,000 lb on the rails. Cars of 70 tons (220,000 lb on rail) or less stay within the 26,000 lb per wheel range, but those of greater capacity exceed that range, even with 36-in. wheels.

Nevertheless, the operating economies of high-capacity cars—lower unit tractive resistance, shorter trains (and sidings) for the same tonnage, more tons per train and engine crew, shorter loading and unloading time—are attractive, especially for coal and ore haul and twin stack container service, even calling for a 125-ton car.

The advantages of heavy axle loads may be offset, in part, by more rapid care and track deterioration and the costs of maintenance thereby incurred. Rail wear and fatigue failures occur at an earlier stage, especially on the high rail of curves and on the gauge face. Rapid abrasive wear can have a greater effect on rail life than fatigue under these loads. Corrugated rail ensues, concrete ties experience more cracking failures, and ballast and wood ties abrade more readily. Bridges and car wheels are also weight sensitive.

Continued efforts are being directed toward developing technologies that will permit heavier car loadings. A design of car truck that permits the two axles to assume a radial position has helped prolong rail and wheel life. Rail lubrication, both by train-mounted and trackside devices, has had similar effects.

Numerous attempts are being made to reduce car weight. Elimination of the center sill is one such effort. The use of aluminum and other lightweight metals in car construction is receiving attention, and prototypes weighing only 51,000 lb but carrying 105 tons of cargo have been constructed and are undergoing service tests.

The use of alloy and heat-treated rails and frequent rail grinding and contouring are required. Granite and trap rock ballasts tend better to resist abrasion. Concrete ties, using elastic fastenings, are beneficial. Frequent adjustment of surface and alignment are essential. If deferred maintenance is allowed to accumulate, there will be a rapid and costly increase in rail corrugations and loss of line and surface.

The problem is essentially economic. Cars of 105- to 125-ton capacity are being operated using the foregoing technologies, but at a cost that must be weighed against their advantages. The AAR is now conducting tests at its Pueblo, CO, test track in search of hard numbers for track and car component life and costs.

D. Unit Trains

The increase in heavy haul traffic places severe demands on the track. Heavy haul is usually performed in unit trains, trains in turnaround service dedicated to hauling a single bulk commodity between designated points, usually in 100-ton cars. Such trains have a uniform consist. Each car responds repetitively and in the same manner as the car before it to any track irregularity, thereby causing further deterioration at the irregularity. The long trains of 60–100+ cars apply that impact loading many times at one spot, giving the track no time to respond elastically. Permanent deformation is hastened.

E. Rock and Roll

The car body on its center plate acts as a giant inverted pendulum as it pivots above the car trucks. It has a side-to-side or lateral swaying motion that can become violent on rough track. Low or battered joints or uneven cross level are sources of excessive side sway. The natural frequency of a 100-ton hopper car is about $\frac{1}{2}$ Hz. If matched by a uniformly periodic disturbing force, a resonance situation develops, adding to the amplitude of the roll. Rough spots, low or battered joints, can provide that disturbing force. The critical roll period occurs at a speed of about 15 mph when the car is moving 22 ft/sec, so that a full cycle is accomplished every 2 sec, that is, every 44 ft. With the conventional 39-ft rail and $19\frac{1}{2}$ ft of stagger between joints on opposite rails, both car trucks are at a joint area and, with a high center of gravity, the wheels can lift off the rails and derailments occur.

Solutions to these problems include (1) good track maintenance, (2) elimination of joints by using continuous welded rail, (3) restricting the height of the cars' center of gravity to 96–98 in., (4) maintaining a uniform load distribution within the car, and (5) proper suspension design.

Because 15 mph is a critical speed for rock-and-roll effects, most railroads prohibit slow orders at that speed and use 10 or 20 mph instead.

Improved car suspension also helps. Better tracking characteristics are obtained by proper center plate lubrication and by the use of manganese center plate liners, which reduce wear and permit the trucks to swivel freely on curves. Truck stiffness has been reduced by adding elastomeric pads to the bearing supports of roller bearing cars. The elastomeric pads give additional play, permitting greater adaptability to rough track. AAR Interchange Rules also call for $3\frac{11}{16}$ in. of vertical spring travel (compared to an earlier $2\frac{1}{2}$ in.) along with supplemental snubbing devices.

A further attempt to reduce stiffness and lower the angle of attack made by the wheel flange with the rail on curves is the use of a so-called self-steering truck. With more play in the bearing boxes and a lever-arm arrangement between the axles, the axles can assume a radial position on the curve, reducing both curve wear and resistance.

F. L/V Ratio

A measure of the derailing tendency of a car is the L/V ratio, that is, the ratio of the lateral to the vertical component of the wheel load. It has become especially significant with the introduction of the long 85-ft cars and 100-ton cars. Generally accepted values are

$L/V = 0.64$: the resultant passes through the outside base of rail; vertical instability ensues with a likely rollover of an unrestrained rail.
$L/V = 0.75$: wheel flange may climb a worn rail.
$L/V = 0.85$: the outside wheel on a curve may lift with the flange rising above the gauge corner of the rail.
$L/V = 1.00$–1.20: the derailing region; old wheels may climb a new rail; a new wheel reduces the wheel-climbing threshold.

G. Hunting

The theoretical path of a wheel set (wheels rigidly pressed onto an axle) is a smooth sine curve with a crest-to-crest amplitude λ, in feet, of $\lambda = 2\pi(R_w G/2T)^{1/2}$, where R_w is the wheel radius in feet, G is the track gauge in feet, and T is the tangent of wheel coning (1 in $20 = 0.50$).

There is a tendency for coned wheels to assume an equilibrium position between the rails. Therefore, any disturbing force (from track or truck) causes the wheel set to shift laterally and roll at the same angular speed on a different tread diameter. One or both wheels will slip and the truck will tend to oscillate about the center pin, causing the wheel to move from the larger coned diameter and to swing toward the opposite rail. The action becomes repetitive. Irregularities in truck or track accentuate the violence of the reaction. The truck tends to "lozenge" or "parallelogram," thereby increasing flange pressure between the wheel flange and the rail. Hunting leads to gauge widening, excessive rail head and gauge corner wear, wheel tread and flange wear, and excessive wear of the car structure. The problem has been particularly acute with the empty return movements of 100-ton cars in unit trains at critical speeds.

Critical speeds vary with car types. For hopper cars with new wheels the critical speed is 45–50 mph; with worn wheels the critical speed is 30 mph. Boxcars will hunt at 80 mph with new wheels, but at 50 mph when the wheels are worn. Trailer on flatcar (TOFC) equipment can hunt at 75 mph, but will drop to 40 or 45 mph after 25,000 miles of running and 35 mph after 100,000 miles of running.

Corrective action includes increasing the rotative resistance between the car body bolster and the center plate (but with a reduction in tracking ability), introducing elastomeric pads in the bearing assembly, using self-steering trucks, and improving the design of suspension elements.

H. Track Geometry Cars

For safety and economy, track irregularities should not exceed 1/2 in. in gauge or alignment over a 62-ft distance or in superelevation in the same distance. For a high-speed track (over 70 mph), 1/4 in. is the desired limit. Track safety rules of the FRA set specific standards on track classifications by speed. Standards of individual railroads may be equally or more severe.

Track geometry cars are a means of determining the track geometry and support conditions under load and at various speeds. Some cars are self-propelled, and others are hauled by a regular or a special train. They measure and give a mile-by-mile printout of the actual gauge, profile, curvature, superelevation, cross level, and warp. There are two general types of cars: (1) a contact car that measures by actual mechanical or electromechanical feelers touching the rail and (2) a noncontact type that makes use of strain gauges, transducers, laser beams, and accelerometers to obtain signals for translation into linear displacements.

From the car printouts, hazardous conditions needing immediate attention can be identified. The extent to which established criteria are met in the track can be noted and priorities can be set for the maintenance budget. Territorial comparisons and before-and-after readings on the work of production gangs can be used for quality control. The FRA uses its cars to check adherence to Track Safety Standards and for setting track classifications.

I. Derailment Analysis

The costs in life and dollars of derailments, the scrutiny of the FRA and the National Transportation Safety Board, and a better understanding of track–train dynamics have forced a broader approach to the investigation of derailments and their causes—to the end of a reduction in their numbers. The traditional single cause for a derailment is giving way to a recognition that a combination of factors involving track, equipment, and the way the train is handled can combine as causes. Manuals have been prepared by some railroads that outline a step-by-step procedure for analyzing conditions and events prior to a derailment and the combined effects of irregularities in several system components—a worn wheel flange and a curve worn rail, for example. Personnel are instructed in these procedures.

J. Track Geometry

Rail lines are composed of straight track (tangents) and curves, arcs of circles, to which the straight portions are tangent. Curves have the obvious purpose of permitting changes in direction and extend from the point of curve (PC or TC) to the next tangent point (PC or CT). In the United States curves are usually designated by the degree of curve, that is, the amount of central angle subtended by a chord of 100 ft (or an arc of 100.007 ft). Some rapid transit and most foreign railroads use the radius as the designation. A convenient relation between the degree and the radius is $R = 5730/D$, from which $D = 5730/R$.

To offset a train's centrigugal force in rounding a curve, the outer rail is superelevated above the inside rail by an amount $e = 0.0007\,DV^2$ in., where D is the degree of curve and V is the speed in miles per hour (Fig. 14). This e value is the equilibrium superelevation; that is, the car weight is equally distributed on both rails. However, because trains run at different speeds on the same track, a certain amount of unbalanced superelevation is permitted whereby $e_c = 0.0007DV^2 - u$, where e_c provides a comfortable speed elevation with an imbalance of u inches. The value of u is normally taken as 3 in., but it is frequently reduced to 2 or even 1 in. where slow-moving heavy haul trains operate.

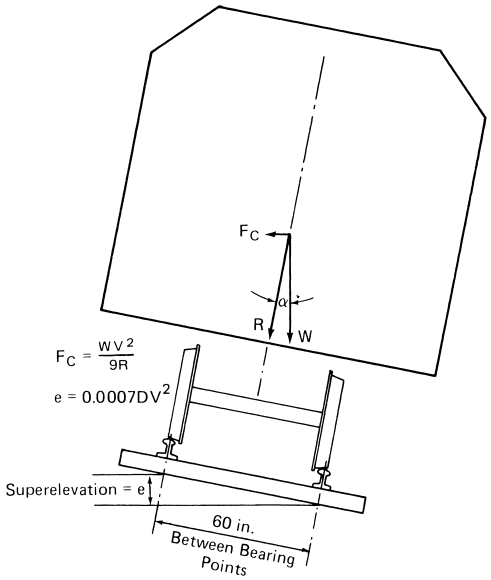


FIGURE 14 Superelevation.

A stable condition prevails when the resultant of car weight and centrifugal force falls within the middle third of the interrail space; otherwise, there is an unstable, possibly derailing, situation. Curvature thus has a limiting effect on speed. The maximum speed permitted by FRA rules is given by the equation, based on 3 in. of imbalance, $V_{max} = [(E_a + 3)/0.007D]^{1/2}$, where e_a is the actual superelevation and D is the degree of curve.

To avoid an abrupt change from level track to full superelevation and from zero to full lateral acceleration, and to provide a smooth transition from tangent to full curvature, an easement or transition curve is placed at each end of the simple curve. The cubic parabola is used in which the degree of spiral increases directly with the distance along the spiral until the full degree of curve is attained at the spiral to curve (SC). The rate of change in degrees per 100-ft station is k, so L_s, the spiral length in feet, is equal to $100\ D/k$. Additional geometry is given in Fig. 15.

VIII. BRIDGES

A. Bridge Use and Types

The complete subject of bridges is beyond the scope of this article. It lies within the province of bridge and structural engineering. Some aspects are, however, pertinent to this presentation in connection with track and location.

Bridges are used to carry the track structure over valleys, ravines, streams, rivers, other railroads, and highways. Unlike culverts, an opening through the subgrade, bridges replace a portion of subgrade. The type selected for a particular site depends on the width of opening to be bridged, the approach terrain, the volume of flow, the required vertical and lateral clearances, the soil and foundation characteristics, and the relative cost of alternative designs.

The simplest design for small openings uses a single girder span of 20–40 ft with a supporting bridge abutment at each end. For wider openings, an additional set of girders and a supporting pier in the middle may be added. Trestles consisting of a number of short grider spans supported by pile bents are in frequent use. Early types used timber bents of five to seven piles with timber caps, but the axle loads of modern trains cause too much crushing of timber caps and piles. Steel trestles with steel piles and steel or concrete caps and decks are finding favor. The piles can be pipe or heavy rolled girder sections. H-piles or channels with cover plates serve as caps, or concrete caps can be used.

For wide openings and unobstructed channels, some form of steel truss is used. The truss is composed of tension and compression members riveted, welded, or bolted together; rivets find little use tody. The trusses are supported at the land end by bridge abutments with bridge seats and in the stream portion of the structure by piers. The bridge deck on which the track is carried can be supported at the bottom of the truss (a through truss), midway between top and bottom (a half-through truss), or at the top. Suspension bridges, used for very wide opening, support the girders

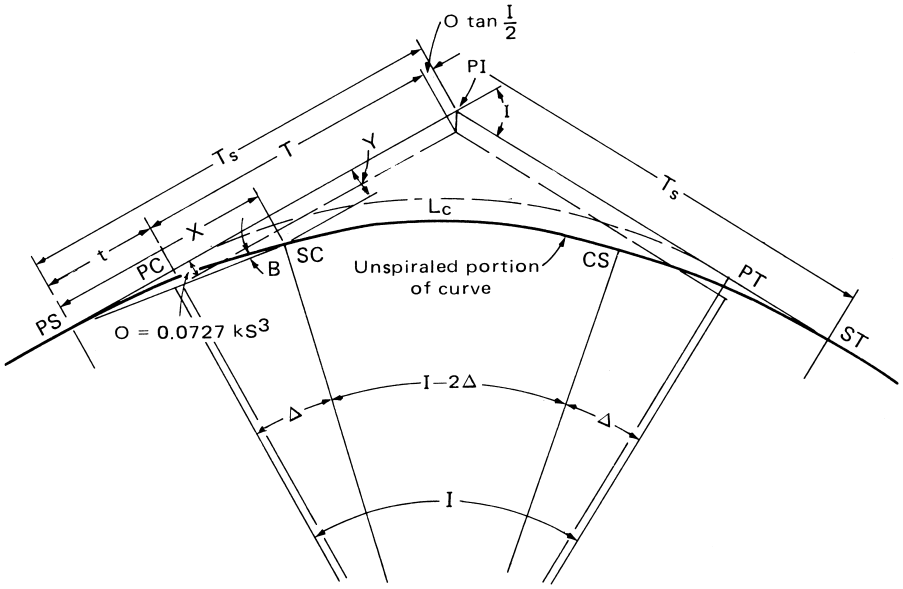


FIGURE 15 Spiral geometry. S = length of spiral in stations; k = change in degree of spiral per station; $D = kS$; $L_c = I/D - S = (I - 2\Delta)/D$; a = spiral deflection $= \frac{1}{6} ks^2$; A = total spiral deflection $= \frac{1}{6} kS^2 = \Delta/3$; $B = 2A$; $t = 100S/2 - 0.000127\, k^2 S^5$; $X = 100S - 0.000762\, k^2 S^5$; $Y = 0.291\, kS^3 - 0.00000158\, k^3 S^7$. [From Hay, W. W. (1982). "Railroad Engineering," Wiley, New York.]

and deck by heavy cables hung from heavily anchored land-based towers with deep foundations. The Brooklyn Bridge, which carries rapid transit as well as street traffic, is of this type.

Where navigation must be maintained and vertical clearance is insufficient, some form of movable span bridge must be used. Where channel width and foundation conditions permit, a vertical lift bridge or a bascule type (hinged and counterbalanced at one end) is used. For wider channels, a swing span mounted on a pivot pier and powered by a steam engine (largely obsolete) or electric motors are used. The pivot pier must be well protected by a fender system of shear fencing and pile clusters against damage by marine craft. The increasing size of towboats, barges, and tows on inland rivers have rendered channel clearances at some pivot pier sites inadequate, and collisions with and damage to the fender system are not uncommon. Navigation interests endeavor to secure wider channel openings whenever a bridge at such a site is renewed. Signal protection for both rail and water traffic is a vital adjunct.

B. Bridge Decks

The railway engineer's chief concern with bridges is the bridge deck. Two types of decks are in use: the traditional timber open deck and the more recent ballast deck. The ballast deck consists of a timber, steel, or concrete trough supported by the girders beneath. A typical concrete trough is about 36 in. deep, 12 ft wide, and supported by one or two sets of girders. The trough is filled with ballast to receive the rails and ties, and the track performance and maintenance are the same as when track is laid on a soil subgrade. Drainage of the trough is important, and the parapet walls are kept high enough to restrain the ballast in the event of a track raise.

The open deck design has one or a pair of girders under each rail. The rails are supported and attached in the conventional way on bridge ties (usually somewhat wider and deeper than standard crossties). The ties are held in place against shifting and bunching by guard logs spiked or bolted to the ends of the ties. Hook bolts are passed through holes drilled in every third or fourth tie or in the guard logs and are secured by a friction grip under the girder head. The bridge ties may be dapped to various depths to give proper surface and cross level, or shims at the girder ends on piers and abutments can give the required adjustments. Rail anchors can be used on ballast decks, but must be omitted on open decks to prevent tie bunching. With continuous welded rail, six rail lengths on each side of short bridges must be box anchored. For longer brides, some form of expansion joint is used.

IX. YARDS AND TERMINALS

A. Yard Functions

Yards are groups of parallel tracks interconnected and used for the classification of cars, their makeup into trains, car storage, interchange, and loading and unloading. Yards are one of a wide variety of facilities for originating, terminating, and interchanging freight and passengers and for servicing the cars and locomotives.

B. Yard Elements

The parallel tracks in a yard are the body tracks. Access to the body tracks is via a ladder track containing turnouts leading to each body track. The ladder tracks connect with a yard lead and running tracks. Yards are double ended with a ladder at each end or stub ended with a ladder at one end and a bumping block at the other. Small yards are flat switched; that is, a locomotive remains attached to a car or cars throughout any movement. Large classification yards depend on gravity to move the cars into the designated tracks. Gravity can be provided by the natural slope of the ground or, more often, by an artificial hump constructed above the entrance to the classifying or sorting unit. Such yards are referred to as hump yards.

C. Classification Yards

Classification yards are used for sorting cars into predetermined groups called blocks and for breaking up trains and building trains from the blocks. The typical classification yard is composed of three units: a receiving yard, where trains are brought from the main tracks to await classification and the locomotive is cut off and sent to an engine house; the classification or sorting unit, where cars are sorted onto tracks corresponding to the designated block; and the departure yard, where blocks from the various body tracks are brought together to form trains for road movement. It is here that the locomotive and caboose (if still in use) are attached, air brakes are tested, and the train is held ready for road movement at the dispatcher's discretion.

D. Car Speed Control

Cars are moved through the receiving yard and the departure yard by yard locomotives. In flat-switched yards, the classification is also performed by locomotives. In gravity yards, however, cars are pushed to the top of the hump by a locomotive and then cut singly or in groups to roll by gravity into the designated tracks. Normally, only one block is assigned to a track. If more than one

block is on a track, that track must be reclassified. Blocks can be by destination (ultimate or intermediate), by commodity grade (as in ore or coal yards), or by continuing rail line.

Control of car speed down the hump incline and into the body tracks is essential to achieve a safe coupling speed (of 4 mph) and to avoid excessive impact and damage to cars and lading.

Yards of an earlier day relied on car riders operating hand brakes for speed control, but today's yards have computer-controlled retarders for the purpose. Two conditions determine the height of the hump and the degree of speed control to be exercised. The first is the hard-rolling car—an empty car, in poor mechanical condition, in cold weather, and with a head wind. The second is the easy-rolling car—a loaded car, in good mechanical condition, in warm weather, and with a tail wind. The hump must be high enough to impart sufficient velocity head to get the hard roller past the clearance point of the body tracks. But the hump cannot be so high as to give the easy roller excessive speed that leads to damaging impacts with cars already standing on the body tracks.

Speed control is exercised by retarder units that apply pressure mechanically via electrically or electropneumatically operated shoes pressed against the sides of the wheel treads. The braking pressure is expressed in feet of velocity head and depends on the amount of resistance versus the effects of gravity that will be encountered between the hump crest and the leaving end of the last retarder and between the last retarder and the clearance point of the body tracks.

Using data on the train's consist received from the arriving train's departure terminal, a sorting or switch list is prepared and placed on a tape or stored in a computer memory that designates the block, body track, and switch routing to that body track, and switch routing to that body track. These data are then combined with the ambient situation—car weight, instantaneous car speed down the hump incline, temperature, wind velocity and direction, and number of cars already on the assigned track—to give continuous and automatic control from the moment a car is separated by the pin puller at the hump crest until it couples with the cars already in the body track. Switches are automatically aligned ahead of the car.

Gradients encountered during the descent are essential elements. These include a steep (4.0%) breakaway grade beyond the hump to separate the cuts of cars, a much flatter grade over the track scales (if any), an accelerating grade through the master retarder, and a somewhat flatter grade through the group retarders. Beyond the clearance point, a nonaccelerating grade may be used to maintain the release speed from the group retarder. Based on an average resistance of 6 lb/ton, this would be a 0.30% grade. Improvements in car rollability and track stiffness have permitted grades as low as 0.20 or even 0.10%.

Some roads maintain a flat grade throughout the yard with a higher release speed from the group retarders. Others continue the nonaccelerating grade almost to the lower end of the yard. In either design, the trend is to put a short decelerating grade at the end of the body tracks to prevent cars from rolling out onto the ladder. Such a yard is referred to as a "bowl." Inert retarders, functioning only by springs through which cars can be pulled, are usually placed below the lower clearance point for the same purpose.

A recent British innovation places a series of hydraulic cylinders and plungers beside the gauge corner of the rails on the hump descent, so that the wheel flanges come in contact with the plungers. Using computerized controls as with conventional retarders, the pressure required by the wheel flanges to depress the plungers is varied to give the desired speed restraint.

E. Terminal Facilities

A variety of facilities are used for loading, unloading, and transferring freight and passengers and for servicing cars and locomotives.

1. Locomotive Terminals

Facilities for servicing and making light running repairs to locomotives are a necessary adjunct to a classification yard. These are usually located within the yard complex or near the coach yard of a passenger (Amtrak) operation. Servicing facilities include those for fuel, sand, water, washing, and repairs. The repair capability can range from simple adjustments and replacements to major overhaul and replacement of engines, generators, and power trucks. A current problem is salvaging spilled fuel and avoiding pollution through the disposal of spilled fuel, old engine oil, and locomotive wash rack fluids.

2. Intermodal Terminals

The increase in TOFC or piggyback traffic has led to the design and construction of terminals where trailers and containers can be transferred between rail and highway vehicles or between rail and waterborne craft.

The general requirements are for paved driveways between track pairs for trucks; parking areas for tractors, trailers, and containers; and load–unload facilities. These may include simple ramps for circus-type roll on, roll off (largely obsolete); gantry cranes moving up and down a line of tracks; individual trailer/container lift cranes for side loading; or, at water terminals, gantry cranes

extending over both tracks and vessels at the wharf side. Note that the use of lifting devices has called for a stronger type of trailer/container construction than would otherwise be required. Other important features are lighting for night operation and security, protection in the form of fences and check-in gates, and communications via computer and other means with the traffic and operating offices of the railroad and truck and shipping interests.

3. Public Delivery Tracks

One or several tracks, depending on traffic volume, with adjacent driveways have traditionally been available for shippers or consignees who do not have their own side track yet deal in carlod quantities.

4. Bulk Cargo Transfer

Coal, ore, grain, and other granular bulk commodities are loaded into cars; cars are unloaded into ships, barges, or stockpiles with automated car dumpers that turn a car upside down to empty it and with a system of conveyor belts to move the commodity from the dumping pit to a further destination. Coal, ore, and grain storage facilities and retrieval equipment are part of the system.

5. Car Repair Shops

The car repair shop or "rip track" where light running repairs are made on cars is also part of a classification yard complex. Once left uncovered and exposed to the elements, sheds and shop buildings now tend to be covered, well lighted, heated, air conditioned, and equipped with overhead cranes, power jacks, and pits where wheels can be changed, brakes inspected and cleaned, and other sometimes heavy repairs performed. A machine shop and (sometimes) a wheel turning lathe are necessary adjuncts for full service. One or more tracks for car storage are also required.

6. Passenger Stations

The magnificent marble structures of the steam era passenger service have given way, for reasons of economy, to small, minimal facilities, or, as with Penn Station in New York and Union Station in Chicago, the railroad facilities are concentrated in a small area to make room for a variety of commercial services and enterprises. With the advent of mergers and Amtrak, many individually owned stations have been abandoned and the services have been concentrated in one central complex. When a station serves both commuters and intercity traffic, the flow lines for each should be kept separate and not cross or intersect.

X. SIGNALS AND COMMUNICATIONS

A. Train Operation

A single train moving up and down a rail line presents few operating problems. When two trains are on the same line, problems of train interference, safety, and efficient use of track capacity arise. Many of the solutions to these problems lie within the realm of electrical engineering and electronics and, more recently, communications and computer engineering.

The oldest as well as the newest forms of train direction and control can be found in revenue service on various parts of U.S. railroads.

B. Train Orders and Book of Rules

Before the advent of signals, trains ran on schedules set forth in the employees' timetable. Meeting points between trains were denoted, and a book of rules established arbitrary superiority and precedence at meeting points and for situations where trains did not adhere to schedules.

The book of rules also prescribed procedures for running trains not authorized by the timetable schedule, that is, for extra trains aided by train orders. The system was slow, cumbersome, and hazardous. It depended on the attentiveness of operating personnel and the accuracy of their watches.

C. Manual Block

In a block system, the track is divided into segments or blocks. A signal is set manually by a block station operator in conjunction with the adjoining block operator to stop a train or permit it to enter and move through the block until stopped or permitted to proceed by a signal at the next block station. Communication was first by telegraph and later by telephone. The operator keeps a block sheet, a record of whether or not a block is clear of trains and when they entered and left the block.

D. Automatic Block

Automatic block systems in general use depend on track circuits. With no train in the block, the circuit is energized, and a green light or other aspect permits the train to enter and move through the block at maximum authorized speed. When a train is in the block, the power supply is shorted through the wheels and axles of the train, and a relay is deenergized and changes contacts to permit current to flow through the red or "stop" aspect (Fig. 16).

More complex systems permit several different aspects and indications to tell not only of block occupancy or nonoccupancy, but also the speed to run and the condition of the next block or two or three ahead. A broken rail will

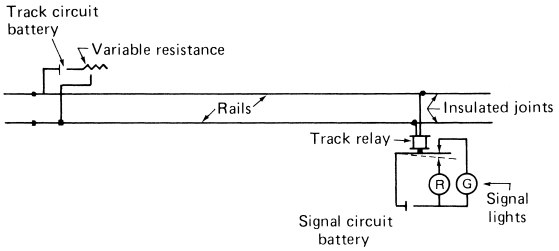


FIGURE 16 Simple closed track circuit.

interrupt the track circuit and cause the signal to display a stop indication.

Even greater selectivity can be had with coded track circuits. A coding device, responding to a feedback from the block ahead, sends any of several codes or current interruptions per minute through the rails. A code-following relay responds to the code and activates the corresponding signal relay and aspect.

Automatic block, simple or coded, is adaptable to single track or multiple track and to operation in both directions on multiple track.

E. Centralized Traffic Control

Centralized traffic control (CTC) concentrates the operation of signals, switches, and crossings over hundreds of miles of railroad in the hands of one individual, who works through a control console before an illuminated track diagram board that shows by appropriate lights the position of each switch and signal and the location and movement of all trains. This operator makes judgments regarding the meeting and passing points for trains and sets the switches and signals accordingly. Safety, better utilization of track capcity, and faster train movements are obtained. With recent innovations, a computer receives these data, analyzes them, and sets the optimum routing for the trains. The operator or dispatcher monitors the performance with override capabilities.

Relays, the heart of an automatic block or CTC system, are being replaced on a trial basis with computerized control that perfors all the necessary steps of analysis and decision making along with "failsafe" and redundancy features.

F. Automatic Train Control (ATC)

ATC, as it is often called, provides for an automatic response to a restricting signal by the train if the engineman fails to make the proper response within a given time period. Without that response, the brakes are automatically applied and speed is reduced or the train is brought to a stop, depending on the type of restriction in effect.

Two general systems are in effect with variations of each. The intermittent system depends on an inductive field interaction between a receiver mounted on the locomotive and a transmitter unit or wayside induction ramp mounted at a signal location. A code that actuates the control system is picked up from the wayside ramp.

The continuous train control system depends on an inductive pickup, by a receiver mounted on the locomotive, of the codes from a coded track circuit. Any change in track occupancy, favorable or otherwise, is thus known by the control system for immediate response. Cab signals that reproduce wayside aspects and indications within the cab may be included with ATC or may operate separately.

G. Communications

A railroad system cannot operate safely and efficiently without adequate communications. The telegraph has been replaced throughout with telephone and, in many areas, particularly those subject to heavy ice storms, with microwave radio. Buried cables, for both signals and communications, are used in locations subject to hurricanes.

Radio communication may be only between the wayside operator and train and engine crews or between the dispatcher and train and engine personnel directly. This is especially good in concentrating all communication within the locomotive cab when the caboose is no longer used.

Yards must be kept informed of the consist and movement of cars and trains so that switch lists can be made in advance of a train's arrival and the yard routing and consist of departing trains can be prepared and sent to the next terminal point. Car tracing and reporting for shippers becomes part of the process.

Within the classification yard communication is required between the yardmaster's office and the hump office, retarder operators, car inspectors, and the locomotives in various parts of the yard. Telephone, radio, walkie-talkies, and fixed talk-back systems are used for these purposes.

An interesting development in communication is the use of fiber optics. Fine glass fiber cables permit carrying many communication circuits in a cable very much smaller than wire cable. Such a cable is being installed on the Northeast Corridor along the right of way with Amtrak being allotted sufficient capacity to meet its communication needs in that territory. Other railroads are entering into similar agreements to permit fiber-optic cables on their rights of way and using part of the capacity for their railroad purposes.

H. Advanced Train Control Systems

Advanced train control systems (ATCS) combine the most recent developments in computerized circuitry,

communications, train control, and signaling in the decision-making processes of train dispatching.

Situation features such as track layout, gradient, curvature, speed restrictions and slow orders, train identification (train number, tonnage, number of cars), and train priorities are entered into the computer system. Train location and movements are received from transponders sited at frequent intervals along the track.

The computer, following its program of priorities and schedules, makes decision as to the optimum meeting, overtaking, and passing points for the trains. Each train is advanced with delay time and speed as set by established priorities. The program sends appropriate instructions to the field equipment (signals and switches), setting these to accomplish the desired train movements. The dispatcher can follow these movements on a display screen in a monitoring capacity. He will only enter the decision-making process when there is a change—a change in routing or priority, for example. A complete or partial track/train display panel shows the position of trains, switches, and signals, with or without a supplementing monitor panel on which a more detailed portion of the track system may be selected for display.

ATCS can be installed in several stages or layers of sophistication. An additional feature will permit the engineman to receive a visual display of the track and signals ahead and to be in direct contact with the dispatcher. The possibility also exists of extending control of the train, starts, stops, and reductions and increases in speed, with the engineman functioning only in a monitoring capacity with override capability.

Another assist to the engineman is the monitoring of locomotive performance as it relates to the train's tonnage, speed, and ambient conditions. Locomotive maintainers can be alerted to the need for maintenance before delays or disasters occur. Dispatchers and/or dispatching systems are warned of possible delays, and adjustments are made according to the pattern of operation.

An adjunct to ATCS is the automatic train inspection, whereby roadside monitoring devices check such items as dragging equipment, hot wheels, hot bearings, flat wheels, and oversized loads, alerting the train crew and the dispatcher to permit adjustments to equipment and schedules before failure and delays occur.

ATCS finds a place in centralized dispatching in which the operational control of the entire railroad system is consolidated in one location. One railroad has established centralized and computerized control of 20,000 miles of line.

The use of radio contact between trains and roadside stations is under the jurisdiction of the Federal Communications Commission, which has, in turn, designated the AAR as frequency coordinator and assigned frequencies in the 890–935 range for ATCS use.

SEE ALSO THE FOLLOWING ARTICLES

ELASTICITY • ELECTRIC MACHINES, ROTATING, CONSTRUCTION AND THEORY • INTERNAL COMBUSTION ENGINES • MECHANICS OF STRUCTURES • NONLINEAR DYNAMICS • POWER TRANSMISSION, HIGH VOLTAGE • SURVEYING • VIBRATION, MECHANICAL

BIBLIOGRAPHY

Anderson, M. J., Walbrun, M. C., *et al.* (1985). "Passenger facilities," *Am. Railw. Eng. Assoc. Bull. No. 699* **86,** 76–90.

A.R.E.A. Committee 16, Economics of Train Speed. (1989). *American Railway Engineering Association Bulletin* **90**(721), 272–278.

Blaine, D. G. (1980). "Modern Freight Car Air Brakes," Simmons-Boardman, New York.

Bosserman, B. N. (1981). "Status Report on the Ballast Experiments at the Association of American Railroads Facility for Accelerated Service Testing," prepared for the Transportation Research Board, Pueblo, CO.

"CSX Gains 20,000 Mile Control." *Progressive Railroading* (May 1989), pp. 32–36.

Effect of heavy axle loads on track. (1975). *Tech. Proc. 12th Ann. Railroad Eng. Conf.*, Federal Railroad Administration, Washington, DC.

Hanna, A. N. (1975). "Railway track research—theoretical and experimental," *Research and Development Bull. RD 030,01R*, Portland Cement Association, Skokie, IL.

Law, E. H., and Brand, R. S. (1973). "Analysis of the non-linear dynamics of a railway vehicle wheelset," *J. Dyn. Syst. Meas. Control* **95**(1), 28.

Lewis, W. F., Cerny, L. P., Hay, W. W., Hutcheson, T. B., Reiner, I. A., and Zarembski, A. M. (1981). Cost Analysis Organization's subcommittee report on the "Economics of '80' vs '100' Ton Cars: Final Report of the Engineering Panel," sponsored by the Association of American Railroads, Washington, DC.

Mitchell, R. J., and Pak, K. (1979). "Model studies of track support systems," *Transp. Res. Rec. 733, Transp. Res. Board.* pp. 21–28.

" Proceedings of the Association of American Railroads Conference on Track-Train Dynamics Interaction" (1971). Vols. 1 and 2, Research and Test Department, Association of American Railroads, Chicago, IL.

Tayabji, S. D., and Thompson, M. R. (1977). "Considerations in the analysis of conventional railway track support system," *ASCE Transp. Eng. J.* **103**(Te-2), 279–292.

Raymond, G. P. (1977). "Railroad wood tie design and behavior," *ASCE Transp. Eng. J.* **103**(TE-4), 521–538.

Reed, H. D. (1983). "Crosslevel safety performance index," *Am. Railw. Eng. Assoc. Bull. No. 693* **84,** 347–363.

Reinschmidt, A. J., Choros, J., and Shafarenko, V. (1985). "Track gage characteristics as measured from a moving vehicle on main track," *Am. Railw. Eng. Assoc. Bull. No. 701* **86,** 158–180.

Steele, R. V., and Stone, P. H. (1986). "Developments in railroad rail," *Am. Railw. Eng. Assoc. Bull.* **87**(707), 311–358.

Sundberg, J. M., Lowe, D. A., and Ogden, P. R. (1989). "Panel on laying and maintenance policies on continuous welded rail," *Am. Railw. Eng. Assoc. Bull.* **90**(717), 353–371.

Ward, E. J., *et al.* (1985). "Economics of proposed new classification yard," *Am. Railw. Eng. Assoc. Bull. No. 701* **96,** 210–226.

Zarembski, A. M. (1981). "Effect of Increasing Axle Loads on Rail Fatigue Life," Report 485, Association of American Railroads Technical Center, Chicago.

Zarembski, A. M. (1989). "The evolution and application of rail profile grinding," *Am. Railw. Eng. Assoc. Bull.* **89**(718), 471–487.

Zarembski, A. M., *et al.* (1987). "Corrugation behavior—The freight railroad environment," *Am. Railw. Eng. Assoc. Bull.* **88**(712), 303–332.



Surveying

Jack B. Evett
The University of North Carolina at Charlotte

I. Introduction
II. Basic Measurements in Surveying
III. Basic Operations of Surveying
IV. Types of Surveying
V. Additional Topics in Surveying

GLOSSARY

Azimuth Indicator of the direction of a line by giving the clockwise angle the line makes with a north line.
Bearing Indicator of the direction of a line by giving the acute angle between the line and a north–south line and the quadrant in which the line lies.
Error Deviation of a measured value from the true value as a result of conditions basically beyond the control of the person making the measurement.
Geodetic surveying Surveying in which the approximately spheroidal shape of the earth is taken into account.
Local attractions Electromagnetic phenomena that can cause a compass needle to deviate from north.
Magnetic declination Difference between the direction of a compass needle when lined up in the earth's magnetic field and a true north line.
Mistake Deviation of a measured value from the true value as a result of conditions basically within the control of the person making the measurement.
Photogrammetry Means of surveying by photographing land from an airplane and preparing maps from the photographs obtained.
Plane surveying Surveying in which the approximately spheroidal shape of the earth is not taken into account and the earth is assumed to be flat over the area being considered.
Transit Instrument used in surveying that can perform virtually all of the operations normally done in surveying.
Traversing Surveying of any number of consecutive lines by determining the length and direction of each line.

SURVEYING can be broadly defined as the science and art of determining relative locations of points above, on, or below the earth's surface. It involves, basically, the measurement of distances and determination of directions. Surveyors make certain types of measurement by various methods, use such measurements to perform appropriate computations, and often display results in the form of maps, plats, diagrams, and so forth.

I. INTRODUCTION

A. History of Surveying

Surveying originated many centuries ago (most likely around 1400 B.C. in Egypt) when land was divided into

tracts for delineation of ownership and taxation purposes. Such surveying was, of course, crude by today's standards. Some improvement in surveying techniques occurred during the heyday of the Roman Empire in association with the large-scale construction of that period.

It was not, however, until much later—the eighteenth century—that significant advances in surveying techniques began to occur. The need for more accurate nautical maps for navigation may have provided the initial impetus, but the necessity for better delineation of property boundaries as the value of property increased, as well as the construction of canals and railroads and, later, highways, also contributed to the development of new surveying techniques.

The twentieth century has seen rapid and major improvements in surveying techniques, largely as a result of military requirements during the two World Wars and two "conflicts" (Korea and Vietnam) and of the space program. Modern surveying techniques include the use of electronic measuring devices, lasers, aerial photogrammetry, satellites, and computers, to name a few.

B. Plane Surveying versus Geodetic Surveying

It is important to differentiate between plane surveying and geodetic surveying. In *geodetic surveying*, the approximately spheroidal shape of the earth is taken into consideration, and horizontal distances and angles are projected onto the surface of a spheroid. In actuality, all surveying should be considered geodetic surveying. If, however, a given surveying job covers only a very small portion of the earth's total surface (as is often the case), that surface may be assumed to be flat without appreciable error. Surveying done under this assumption is called *plane surveying*, and horizontal distances and angles are projected onto a horizontal plane.

Inasmuch as plane surveying is much simpler than geodetic surveying, local surveying is generally done under the concept of plane surveying (i.e., a "flat earth"). Geodetic surveying is required for surveys over larger areas (such as a state or country).

C. Units of Measurement

In the United States, the unit for measuring distance has long been the *foot* (ft), with distances given in feet and decimal fractions thereof. In other countries, where the International System of Units (SI) is used, the unit for measuring distance is the *meter* (m). For long distances, units of *miles* (mi) or *kilometers* (km) may be used, where 1 mi = 5280 ft and 1 km = 1000 m.

If distances are measured in feet, a bounded area can be computed in *square feet* (sq ft). Land in the United States is often bought and sold, however, on the basis of area expressed as *acres*, where 1 acre = 43,560 sq ft. For distances measured in meters, a bounded area can be computed in *square meters* (sq m). Large areas may be expressed as *hectares* (ha), where 1 ha = 10,000 sq m.

If distances are measured in feet, an associated volume could be computed in *cubic feet* (cu ft). In common surveying usage, however, volumes are often expressed as *cubic yards* (cu yd). When SI units are used, volume can be computed in *cubic meters* (cu m).

Angular measurement is commonly determined in units of *degrees*, *minutes*, and *seconds*, where 1 degree is $\frac{1}{360}$ of the total angle (i.e., one complete rotation) about a point, 1 minute is $\frac{1}{60}$ of a degree, and 1 second is $\frac{1}{60}$ of a minute. Degrees, minutes, and seconds are commonly denoted by the symbols $^\circ$, $'$, and $''$, respectively. Hence, an angle of 22 degrees 19 minutes 45 seconds can be written $22^\circ 19' 45''$.

D. Errors and Mistakes

Surveying measurement and computation must be accurate. Land is often bought and sold for thousands of dollars per acre. It is the surveyor's job to determine the area of land, and any inaccuracies can be costly to the buyer or seller. Similarly, landowners may be compensated for earth being removed from their property for use in nearby highway construction on the basis of so much money per cubic yard of earth removed. Again, it is the surveyor's job to determine the volume of earth removed, and inaccuracies can be costly.

Inaccuracies occur in surveying measurements as a result of errors and mistakes. An *error* is a deviation of a measured value from the true value as a result of conditions generally beyond the control of the person making the measurement. In other words, the person did the best job possible in the circumstances, but the measured value is not the true value. Errors can result from instrumental causes (e.g., a defect in the scale used to evaluate angles), personal causes (e.g., one's inability to maintain the required steady tension on a tape when determining a horizontal distance), and natural causes (e.g., inaccuracy in observing a line of sight through a transit or level telescope as a result of heat waves). Errors tend to be either systematic or random. In the case of systematic errors, the total error in a measurement can be evaluated and the measured value corrected to give the presumed true value. Random errors, on the other hand, tend to be compensative.

A *mistake* is a deviation of a measured value from the true value as a result of conditions basically within the control of the person making the measurement. In other words, the person "goofed." Examples of mistakes are misreading a tape, level rod, or vernier scale (e.g., reading

4.97 instead of the correct value of 3.97 on a level rod) and transposing digits when recording measured values (e.g., recording a measured distance of 324.7 ft as 342.7 ft).

Surveyors should always strive to minimize errors and avoid mistakes by concentrating on the job at hand, exercising care, using good judgement, and checking and rechecking whenever possible.

E. Surveying as a Profession

Although some individuals have, in the past, become surveyors through the "school of hard knocks," becoming a professional surveyor today generally requires some higher education as well as practical experience. The necessary higher education is often obtained as part of a baccalaureate engineering program, notably in civil engineering. An alternative is completion of a 2-year associate degree program emphasizing surveying.

To perform surveying and be in responsible charge, a person is required by the states to be a registered (licensed) land surveyor. Registration (licensing) is carried out by the states. Although specific requirements vary somewhat from one state to another, the requirements generally involve satisfactory education, experience (under the supervision of a registered land surveyor), and good character.

Some surveyors are self-employed; others work for surveying companies, engineering consulting companies, construction companies, and the like. Also, surveyors are employed by various government agencies, such as the Coast and Geodetic Survey and highway departments. Surveying is a good career choice for individuals who are seeking a challenge, are willing to work hard, and enjoy being out-of-doors.

II. BASIC MEASUREMENTS IN SURVEYING

A. Horizontal Measurement

Perhaps the most basic operation in surveying is the measurement of *horizontal distances*. In addition to simply measuring horizontal distances, horizontal angles can be evaluated by measuring three horizontal distances and employing the law of cosines. Hence, a considerable amount of surveying can be done with only a horizontal distance measurement device.

As related in Section I, in plane surveying horizontal distances are projected onto a horizontal plane. A piece of land, no matter what its relief may be, is analyzed and evaluated as if it were projected onto a horizontal plane. Accordingly, whenever the linear distance between any two points on the earth's surface is referred to, it is understood that it is the horizontal distance between the points, regardless of the elevations of the end points of the line, unless specified otherwise. Therefore, in measuring the distance between two points (such as the length of a side of a lot), either the measurement is made along a horizontal line between the points or the measurement is made along an inclined line between the points, with the horizontal distance computed and recorded. If the first method is used, the position of one or both end points may have to be projected vertically upward to the horizontal line of measurement. This can be achieved in some cases by means of a plumb bob.

Horizontal distances can be determined in a variety of ways. Some common means are pacing, stadia, taping, and electronic measurement.

Pacing is a crude means of finding horizontal distances, but it may be useful if rough values of distances are acceptable. The first step in pacing is to calibrate one's pace (step). The typical adult's pace ranges from about $2\frac{1}{2}$ to 3 ft. Once the pace is determined, a distance can be estimated simply by walking along the line, counting the number of paces (steps), and multiplying that number by the length of one's pace. An experienced pacer should be able to determine the value of a distance of $\sim$100 ft within 2 or 3 ft either way.

Stadia is a means of determining horizontal distances that is more accurate than pacing, although it is not accurate enough for some purposes. Stadia is carried out by viewing a leveling rod through a transit's telescope. Two crosshairs in the telescope appear to intersect the rod at two different points, and the distance between these two points on the rod is directly proportional to the distance from the transit to the rod. Hence, by noting where the two crosshairs intersect the rod, determining the distance between these points, and making appropriate computations, one can evaluate a horizontal distance. Stadia is generally accurate to the nearest foot.

Taping is a common means of determining horizontal distances. The general procedure, which is well known to laypeople, is to stretch a calibrated tape along the line to be measured and read the value of the distance off the tape. Although this is a simple procedure that anyone can do, several points must be considered in order to attain the accuracy generally required in surveying.

To begin with, as discussed previously, in measuring the distance between two points, either the measurement is made along a horizontal line between the points or the measurement is made along an inclined line between the points, with the horizontal distance computed and recorded. If a measurement is being made along a straight, horizontal surface, it is no problem to keep the tape horizontal, since it can be stretched tightly along the horizontal surface. For other situations, one can achieve horizontal

alignment of the tape, at least approximately, by using a plumb bob at one or both ends of the tape and stretching the tape tightly.

Another point to consider is that the tape must be stretched in a straight line between the end points of the line. Although this would appear to be a simple matter, it can be difficult in some cases, such as in measuring distances through heavy underbrush. One must also exercise care when measuring distances longer than a single tape length. When this occurs, temporary points must be marked as successive end points of the tape are encountered, and each such marking presents an opportunity to incur error.

If extremely accurate measurement must be done by taping, it may be necessary to compute corrections for various phenomena and apply them to a measured length to obtain the correct length. For example, a given tape may be slightly too short or too long as a result of imperfect manufacture, wear, and so on. Obviously, a distance measured with a tape that is too short or too long will give an erroneous value. In addition, a measured distance is subject to error caused by expansion and contraction of the tape as a result of change in temperature. Error in measurement also results when the tape is pulled either too tightly or not tightly enough. Finally, error in measurement occurs if the tape sags when it is supported only at its end points.

In each of these cases, the actual amount of error in a measured value is ordinarily quite small and is often neglected. However, if extremely accurate measurement is required, formulas are available for computing corrections for each case, and the total correction can be applied to the measured distance to determine the correct value.

More recently developed devices for measuring distances are known as *electronic distance-measuring* (EDM) *devices*. They find lengths based on phase changes that occur as electromagnetic energy (a "signal") is sent from one point to a second point and back to the first point. With the wavelength and length of time of travel known, the distance between the points can be computed. In practice, an EDM device, an example of which is shown in Fig. 1, is placed over one point and a reflecting target is placed over the other point. The signal is sent, the time recorded, and the distance computed. (It should be understood that this description of the operation of an EDM device is oversimplified.)

In the first EDM devices, the "signal" was a light beam. Subsequently, high-frequency microwaves and lasers were used. More recent advances have made the devices smaller, lighter, easier to use, and capable of giving a direct readout of the distance.

EDM devices can be used to measure very long distances while giving extraordinarily accurate results—of the order of 50 mi $\pm$ 2.6 ft (80 km $\pm$ 0.80 m). They are exceedingly useful in measuring long distances and others that would be difficult to measure otherwise (such as the distance across a large lake).



FIGURE 1 Electronic distance-measuring device.

B. Vertical Measurement

Another basic operation performed in surveying is the measurement of *vertical distances*. This type of information is often needed in engineering applications to find differences in elevations of various ground points as a means of computing slopes. Determinations of elevations and slopes are important in many contexts, including, to name a few, topographic mapping, design of highways, drainage systems, and sewers.

In some cases, vertical measurements can be made in the same manner as horizontal ones. For example, the height of a building might be found by stretching a tape from the roof to the ground. Often, however, this kind of direct measurement of vertical distance is not feasible because the end points are not located directly above and below one another. This is true, for example, in determining the difference in elevation between two ground points at different locations on the earth's surface. Differences in elevation can be evaluated by a procedure known as *differential leveling*.

Differential leveling is carried out using two devices: a level and a level rod. The basic element of a level is a telescope that can be adjusted so that it is always aligned in a level plane with its line of sight along a level line. It is mounted on a tripod when in use. A level rod is simply a rod that is calibrated for easy viewing and reading through the level's telescope. Examples of levels and a level rod are given in Figs. 2 and 3, respectively.

To begin the leveling procedure, a level rod is held vertically over a point of known elevation (a bench mark). With a level set up and properly leveled, a sighting is made on the rod and a reading made. This reading is known as a backsight; and when it is added to the elevation of the bench mark, the sum gives the height of the instrument



FIGURE 2 Top: Level; Bottom: automatic level.

(i.e., the telescope). The level rod is then moved to a point the elevation of which is desired, a sighting is made on the rod, and a reading made. This reading is known as a foresight; subtracting it from the height of the instrument gives the ground elevation where the rod was held.

The preceding describes the standard procedure for determining the elevation of a second point, given the elevation of a first point. This procedure can be repeated (taking a backsight on the point the elevation of which was just determined) as many times as necessary to obtain elevations of additional points. A check for accuracy is commonly made by continuing levels so as to finish on the starting point. The elevation of the starting point as computed from running the levels should, of course, be the same as its beginning (known) elevation.

Other methods of determining elevations include stadia and the use of EDM devices (see Section II.A) as well as barometers.

C. Directional Measurement

Sections II.A and II.B discussed the basic measurement of distances—including both horizontal and vertical ones. It is often not sufficient, however, to know only how far away something is or how long a line is; one must also know "in what direction it lies." Hence, another basic measurement made in surveying is that of direction. *Direction* is documented by measuring horizontal angles, vertical angles, bearings, and azimuths.

The instrument used in surveying for measuring angles, bearings, and azimuths is known as a *transit*, an example of which is shown in Fig. 4. A transit is a versatile device and exceedingly important to surveyors. It is probably the one piece of surveying equipment most commonly identified with surveyors by laypeople. Transits can be used to perform virtually all basic operations done in surveying, including measuring horizontal distances (by stadia), vertical distances (by stadia and running levels), angles in both horizontal and vertical planes, and directions (bearings and azimuths) as well as aligning points on a straight line.

Like levels, transits have a telescope that can be adjusted so it is aligned in a level plane; hence, transits can be used to run levels. Transits are far more than leveling devices, however. A transit's telescope can be rotated upward and downward to enable surveyors to evaluate angles in a vertical plane as determined from a circular scale. Transits



FIGURE 3 Level rod.

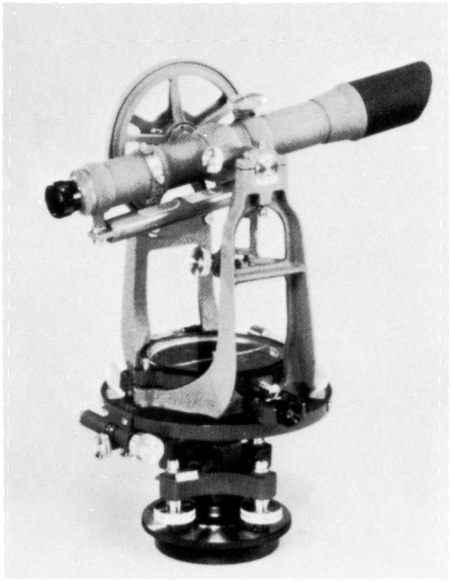


FIGURE 4 Transit.

have another circular scale that makes it possible to measure angles in a horizontal plane. Also, transits have a compass needle and associated scale, which allow evaluation of magnetic bearings and azimuths.

The transit shown in Fig. 4 is an *American transit*. A *European transit* is shown in Fig. 5; it is commonly known as a *theodolite*. In addition to the obvious differences in appearance between the two instruments (see Figs. 4 and 5), theodolites have horizontal and vertical circular scales made of glass, with graduation lines and numerals etched on the glass surface, as opposed to the metal scale on American transits. As a result, more accurate angular determinations can be made with a theodolite.

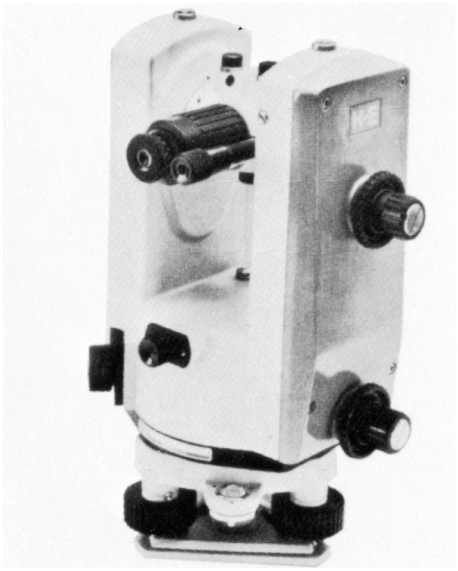


FIGURE 5 Theodolite.

An example of a *horizontal angle* in a surveying context is the angle formed by two adjacent property lines that meet at a property corner. Such an angle can be measured by setting up a transit over the corner, sighting first along one property line and then along the other, and reading the value of the angle off the appropriate circular scale.

Another means of indicating the direction of a line is by use of *bearings*. The bearing of a line is indicated by giving the acute angle between the line and a north–south line (meridian) and the quadrant in which the line lies. For example, a line making an angle of 44°30′ with a north–south meridian and being in the northeast quadrant would have a bearing of N 44°30′ E. A line making the same angle with a north–south meridian but being in the northwest quadrant would have a bearing of N 44°30′ W.

Bearings are evaluated by reading a compass needle that rests on a pivot in a recessed well in the horizontal plate of a transit. To determine the bearing of a line, the transit is set up over one end of the line and a sighting is taken along the line. The compass needle is then allowed to swing back and forth until it comes to rest. At this point, the needle is pointing toward north and the telescope is pointing in the direction of the line. The value of the bearing is read where the end of the needle intersects the associated circular scale.

Bearing determinations are subject to error as a result of what is known as *local attraction*. This refers to various electromagnetic phenomena that can cause compass needles to deviate from north. Examples are electric power lines, local ore deposits, and magnetic effects of fences, automobiles, and wristwatches.

Another problem encountered with bearings is known as *magnetic declination*, which results because compass needles do not generally point to the true (geographic) north pole. Instead, they align themselves in the earth's magnetic field, which varies over a period of time. The difference between the direction of a compass needle when aligned in the earth's magnetic field and a true north line is the magnetic declination. It varies in amount from one location to another, and as noted above, it varies over a period of time for a specific location. If the amount of magnetic declination is known, a true bearing can be converted to a magnetic bearing, and vice versa. The direction of true north (and therefore the declination) can be determined by sighting on the North Star (Polaris).

A final means of indicating the direction of a line is by use of *azimuths*. The azimuth of a line is indicated by giving the clockwise angle the line makes with a north line. Hence, the azimuth of a line with a bearing of N 44°30′ E is simply 44°30′; the azimuth of a line with a bearing of N 44°30′ W is 315°30′. Inasmuch as azimuths are tied to compass readings, most of the comments concerning

bearings (local attraction, magnetic declination) are applicable to azimuths.

III. BASIC OPERATIONS OF SURVEYING

A. Traversing

Traversing refers to the surveying of a number of consecutive lines by determining the length and direction of each line. There are basically two kinds of traverse in surveying: open and closed. An *open traverse* is one that does not return to its starting point; a *closed traverse* does return to its starting point. Examples of open traverses are surveys to lay off highways, pipelines, and power transmission lines, none of which generally returns to its starting point. An example of a closed traverse is a boundary survey of a piece of property.

Field surveying involved in traversing is essentially the measurement of lengths and directions, in accord with procedures related in Section II. Points where adjacent lines connect are marked by "hubs"; often, a hub is a wooden stake with a tack on it for open traverses or a driven metal pipe or rod in the case of closed traverses.

For closed traverses, there is an excellent scheme for checking the accuracy of field measurements made. The first step is to compute latitudes and departures for all traverse boundary lines. The *latitude* of a line is the component of the line in the north direction; it is computed by multiplying the length of the line by the cosine of its bearing angle. The *departure* of a line is the component in the east direction; it is the product of the length of the line and the sine of the bearing angle. The check on accuracy of field measurements is made by finding the sums of both latitudes and departures, for they should, in theory, both add to zero. In actuality they may very well not sum to exactly zero because surveyors, being human, cannot make every measurement perfectly. The question then becomes: How far from zero may the sums be with the survey results still considered to be within acceptable accuracy? This is answered by computing the value of a parameter called the *precision* and seeing if it falls within reasonable limits.

If a latitude and departure analysis indicates unsatisfactory results, field measurements should be checked. If results are deemed to be acceptable, the survey can be "balanced" by adjusting latitudes and departures to force them to sum to zero. From adjusted latitudes and departures, revised values of lengths and bearings of all lines can be computed.

B. Triangulation

Triangulation provides a convenient means of establishing precise locations of a number of points (i.e., providing horizontal control) over a large area. From these points, future surveys can be begun. For example, triangulation might be carried out over an area of considerable length through which a new highway is to be built, or it could be used to provide horizontal control over an entire county or state.

As suggested by its name, triangulation involves setting up a system of overlapping triangles covering an area under consideration. By measuring a number of angles and sides of these triangles, one can compute precise locations of successive points by trigonometric means. Upon occasion, more angles than sides are measured because it is sometimes easier and more expedient to measure angles than distances over large areas. (This was particularly true before the advent of EDM devices.)

To attain the high accuracy desired in triangulation, it is, of course, necessary that measurements of angles and distances be made with extreme care. As one means of securing the required accuracy, measurements are sometimes made at night to avoid certain atmospheric conditions that occur during sunlight hours and cause problems in sighting long distances through a telescope.

C. Mapping

A great deal of surveying is done for the purpose of preparing maps for certain areas or plats of property. Topographic maps show, in addition to natural and artificial features of the land, relief (variation of elevation of the earth's surface). They are widely used in engineering work.

Anyone needing a map or plat of an area should be aware of the fact that many maps are readily available at nominal cost. The U.S. Geological Survey has large-scale topographic maps covering virtually all of the United States. State highway departments are good sources of state, county, and other maps. County offices contain thousands of plats, which are a matter of public record. If, however, a map or plat of a specific area is not available for a given purpose or need, a surveyor may be employed to collect necessary field data and prepare the required map or plat.

The first step in preparing a map is to establish some kind of horizontal control. This can be done by traversing (Section III.A) or triangulation (Section III.B). The results can then be plotted on the map.

Another step is to locate and plot any significant "details," that is, important points or features on the earth's surface. Whereas maps of single lots might show houses, driveways, power poles, and individual trees, maps of larger areas would probably not feasibly show such detail.

As indicated previously, topographic maps indicate ground relief. This is most commonly done using contour lines. A *contour line* is an imaginary line along the earth's surface, all points on which are at the same elevation. This imaginary line is drawn as a solid line on a topographic

map. A series of contour lines provides invaluable information for engineers, planners, and others in many kinds of analysis and design problems. For example, if one traces along a contour line on a map, all points covered are at the same elevation. Hence, it is known, in this case, that water would not flow along the corresponding line on the ground. On the other hand, a line traced on a map from a contour line labeled 97 ft to one labeled 96 ft represents a vertical drop of 1 ft on the ground. By determining the horizontal distance between the two contour lines, which can be scaled off the map, the slope of the ground can be computed. Furthermore, it is clear that closely spaced contour lines indicate steep slopes, whereas widely spaced ones indicate flat terrain.

To draw contour lines on a map, it is necessary first to determine ground elevations of a number of points over the area to be mapped. This can be accomplished by differential leveling or stadia. Then these points must be plotted accurately and the ground elevations written beside respective points on the map. The next steps are to select a suitable contour interval and then draw the actual contour lines. Contour lines are drawn freehand among the plotted points, with ground elevations written beside them. For example, if two adjacent points have elevations of 97.5 and 93.8 ft and contour lines are being drawn at 1-ft intervals, it is clear that four contour lines (the 94-, 95-, 96-, and 97-ft ones) must pass between these points. With practice, one can become proficient at drawing contour lines among labeled, plotted ground points. (Contour lines can also be drawn from aerial photographs.)

Once boundaries, details, contour lines, and so on have been plotted, a map is completed by including a title, the date, a legend, a scale, the name of the person (or agency) preparing the map, and any other information that may be of interest.

Nowadays, many maps are prepared by computers. However, it remains the surveyor's responsibility to provide accurate input data to a computer and to verify the accuracy of maps thus prepared.

D. Area Determination

As related in Section I.D, an important task for surveyors is to determine the area of land, which must be done with considerable accuracy. There are several methods for evaluating area; these are summarized here.

If a parcel of land is relatively simply shaped, such as triangular, rectangular, or trapezoidal, or can be partitioned into several such shapes, and if sufficient data are known, the parcel's area can be computed simply by applying appropriate formulas.

If latitudes and departures (Section III.A) of a tract have been determined, its area can be computed by the *double meridian distance* (DMD) method. The meridian distance of a line is the distance from the line's midpoint to the north–south meridian. Obviously, the DMD is twice the meridian distance. The DMD method for computing area involves finding the DMD of each line, multiplying each line's DMD by its latitude, summing the products, and taking half the sum. The result is the tract's area in square feet (or square meters) if the lengths of the lines are in feet (or meters). In practice, the DMD method is carried out by a systematic procedure with the data in tabular form.

If coordinates of a tract have been determined, its area can be computed by the *coordinate method*. This is achieved by multiplying the x coordinate of each point by the difference between adjacent y coordinates, summing the products, and taking half the sum. The y coordinates must be taken in the same order around the traverse when obtaining the difference between adjacent y coordinates. As with the DMD method, the result is the tract's area in square feet (or square meters) if the coordinates are in feet (or meters). The coordinate method is also carried out in a systematic procedure with the data arranged in a certain format.

Both the DMD and the coordinate methods require straight-line boundaries. Sometimes, one or more of a tract's boundaries are curved lines. For example, a meandering creek or river may form one boundary of a tract. If such a tract is drawn to scale on a map, its area can be determined using a device called a *planimeter*.

An example of a mechanical planimeter is shown in Fig. 6 (top). It is a machine that has two arms, one of



FIGURE 6 Top: Mechanical planimeter. Bottom: Electronic planimeter.

which has an anchor point at one end, while the other has a pointer. It has a wheel on its underside that is rotated as the operator runs the pointer around the boundary of a tract. The area is found by reading a value from a scale on the planimeter. In practice, one generally traces the boundary with the planimeter several times and takes an average value of planimeter readings as the tract's area. (Sometimes planimeter readings must be scaled to obtain the correct area.) More modern planimeters, like the electronic one shown in Fig. 6 (bottom), give direct readouts of areas.

E. Volume Determination

Another important task for surveyors is computation of earth volumes. Many earth volume determinations are associated with highway construction. Two examples, which are covered here, are volumes of cuts and fills in a roadway between adjacent cross sections and volumes of earth taken from borrow pits.

The volume of cut or fill in a roadway segment between adjacent cross sections can be computed by multiplying the average of the crosssectional areas at the ends of the segment by the segment's length. This computation of volume assumes a linear variation in cross-sectional area from one end of the segment to the other. A more accurate determination can be made by adding the two end areas to the cross-sectional area at the middle of the segment multiplied by four, dividing the sum by six, and multiplying the quotient by the segment's length.

Sometimes, earth to be used in a highway fill can be obtained from nearby land. Of course, the landowner must be compensated for earth removed, and such compensation is usually based on the volume of earth removed. The area from which earth is removed is known as a *borrow pit*.

The volume of earth removed from a borrow pit can be determined by laying off a grid and running levels before and after earth is removed to find the initial and final elevations of each grid point. The depth of cut at each grid point can be computed by subtracting its final elevation from its initial elevation. The approximate volume of earth removed within each of the grid's squares or rectangles can be computed by multiplying the average of the depths of cut at each corner of the square or rectangle by the area of the square or rectangle. The sum of volumes from all squares or rectangles gives, of course, the total volume of earth removed from the borrow pit.

IV. TYPES OF SURVEYING

A. Topographic Surveying

Topographic surveying is used to gather data for preparation of topographic maps (see Section III.C). Topographic surveying is conducted by field or aerial methods. Aerial methods (Section V.C) are often used when large areas are to be mapped; field methods (Section III.C), for smaller areas. Whenever aerial methods are used, some fieldwork is usually done to establish control and check for accuracy.

B. Construction Surveying

Probably a majority of all surveying done is associated with construction. Surveying of some type is required in several phases of most kinds of construction. To begin with, a topographic survey and map showing the location of whatever is to be constructed are required before construction begins. During the construction process, surveyors lay out the structure and perform other tasks as needed. After construction is completed, a final map may be prepared to document any changes that may have been made in original design plans.

Several examples of specific types of construction surveys are the staking-out of buildings, highways, and pipelines. Staking out a building involves precisely locating each corner and setting its floor elevation. These tasks are effected and documented by placing batter boards. Staking out a highway entails setting sufficient stakes along the center line and "slope stakes" to each side to indicate how much cut or fill must be made by a contractor in shaping the highway. Staking out a pipeline involves locating the pipeline, such as by placing "line stakes," and setting "grade stakes," which the contractor can use to find the proper pipeline grade.

C. Land Surveying

Land surveying is employed primarily to delineate boundary corners and lines of tracts of land. Specifically, it involves finding old property corners or establishing new ones and then marking the boundary lines that connect the corners. After surveying a tract of land, a surveyor will often prepare a map (plat) showing the tract. He or she may prepare a legal description of the tract if it is to be sold.

In theory, land surveying generally requires only the knowledge and ability to measure and lay off distances, angles, and bearings. In practice, however, good land surveyors must have experience, patience, good judgment, and some knowledge of legal principles and local customs. It is essential that a novice work under an experienced land surveyor for a period of time to become a true professional.

D. Mining Surveying

Mining surveying is a specialized type of surveying done primarily beneath the earth's surface, usually in mines. It is performed to determine locations of underground features relative to aboveground features, to measure quantities of

material removed, to establish line and grade for construction or mining operations, and so on.

Like other types of surveying, mining surveying involves measurements of distances—both vertical and horizontal—and directions. Because of the underground nature of mining surveying, however, some unusual features are encountered and must be dealt with. For example, underground traverses sometimes follow very steeply inclined courses, and traverse segments are often straight for only short distances. Traverse hubs are commonly placed in a mine's roof (e.g., a hook driven or screwed into a timber) to prevent their being disturbed by the mining process. Sometimes a line of levels must be carried down a long, vertical shaft; this can be achieved by lowering a tape vertically. Cramped conditions may hamper instrument setups, and problems of darkness and dampness can cause further difficulties. Mining surveying is probably performed under the most dangerous conditions ordinarily encountered in any type of surveying.

E. Route Surveying

Route surveying refers to the surveying necessary for locating and constructing highways, railroads, canals, pipelines, and transmission lines. Initially, topographic surveys must be done and maps prepared for preliminary route location and project design. Land surveying is required in connection with land and right-of-way acquisition. Before construction begins, the final route must be staked out, grade stakes set, and so on. During construction a surveyor must be available to replace line and grade stakes that are disturbed by construction or provide additional stakes that may be needed. Final survey effort is necessary on completion of construction to document work done.

Route surveying also involves making earth volume computations (Section III.E) and design and field layout of vertical and horizontal curves. Vertical curves are generally designed as parabolas; horizontal ones, as spirals and/or parts of a circle.

Actual surveys to stake out the center line of a route (e.g., a highway or railroad) constitute open traverses. Positions along an open traverse are referred to by their distance from the starting point, with each hundred feet (or meters) being designated as a "station" and the distance beyond the last full station as a "plus distance." Hence, the location of a point along a route at a distance of 2265.4 ft from the starting point would be identified as station $22 + 65.4$.

F. Hydrographic Surveying

The prefix *hydro* refers, in general, to water; hence, *hydrographic surveying* is done on, in, or near water to secure information about the physical features of the water area. A major purpose of hydrographic surveying is to obtain data necessary for preparing nautical charts, which show water depths, navigation channels, structures (such as piers), breakwaters, and so on and which are used by mariners. Hydrographic surveys are also useful in determining information needed for design and construction of structures adjacent to or under water. Such information is also needed when dealing with silting and dredging of lakes and channel bottoms. Hydrographic surveying is also employed in work with drilling platforms, underwater cables, and so on during offshore oil and gas exploration. The latter case is referred to as *marine surveying*.

The main measurement performed in hydrographic surveying is water depth. It can be determined in several ways. If the depth is not too great, a calibrated stick or pole can be lowered in the water until it strikes the bottom. Such a stick or pole is known as a *sounding pole*. For greater depths, a chord that is calibrated and has a weight attached can be lowered until it strikes the bottom. It is called a *leadline*. Water depths can also be determined with an echo sounding device known as a *fathometer* and a laser device called *LIDAR*.

If water depths are determined in a lake one day and then again on another day when the water surface has risen several feet, the difference in surface elevation must certainly be considered in mapping the lake's bottom. Hence, some type of vertical control is necessary. It can be accomplished by finding the surface elevation with respect to some reference datum whenever water depths are being determined and taking this into account when analyzing the raw data.

Some kind of horizontal control is also necessary when water depths are being taken. In mapping a lake, it is essential that the location (in a horizontal plane) of each point for which a water depth is determined be known to plot the map or nautical chart. Horizontal control can be accomplished by traversing or triangulation, but the actual procedure is complicated by the fact that depth determinations are generally made from nonstationary boats, precise locations of which may be difficult to determine over larger water bodies. A navigational device known as a *sextant* is used to measure horizontal angles from a boat to various objects on land, and from these values various locations of the boat (i.e., of the depth soundings) can be determined.

Often associated with surveying and mapping of water areas is the topographic surveying of the shoreline needed to show it on the map or nautical chart. This can be done by transit–stadia ground surveying or by aerial photogrammetry.

G. Municipal Surveying

Municipal (or *city*) *surveying* is an overall term used to describe both surveying done in a municipal setting and

that performed by a municipality in fulfilling its service functions. Examples of the former are surveys for delineating property lines and for locating buildings and retaining walls that are to be constructed. Instances of the latter are surveys for streets, curbs, bridges, sewers, and storm drainage systems.

Of primary importance in municipal surveying is a horizontal and vertical control system. This calls for the establishment of numerous monuments, precise locations and elevations of which are known, throughout the municipal area. They provide numerous potential starting points from which future municipal surveys can begin with the assurance that all surveying done will be tied in to a common reference.

Also of importance in the realm of municipal surveying are the gathering of data for and preparation and dissemination of maps of various kinds. These include base maps, topographic maps, property maps, and subdivision maps, as well as maps showing locations of utility lines and underground features (sewers, water pipes, gas mains, subways, etc.), to name a few.

V. ADDITIONAL TOPICS IN SURVEYING

A. State Plane Coordinate Systems

The precise location of any point on the earth's surface can be specified by giving its latitude, longitude, and elevation above mean sea level. Thousands of points across the United States have had their latitudes and longitudes determined accurately and marked with ground monuments by the National Geodetic Survey (formerly the Coast and Geodetic Survey). Latitude and longitude are spherical coordinates, however, and therefore indicative of geodetic positions, and referencing of local surveys to them can be troublesome. Realizing this difficulty, the Coast and Geodetic Survey established in 1935 the *State Plane Coordinate Systems* (SPCS), which relate the horizontal locations of points in terms of plane coordinates. With these available, surveyors can tie in local surveys to the SPCS with the relatively simple performance of and computations involving plane surveying (i.e., assumption of a "flat earth") and thereby document forever the exact location of a local survey.

In the SPCS, each state has a separately established coordinate system, but all such state systems are tied together. The coordinate system for each state was established by projecting, mathematically, its surface onto the surface of a cone or cylinder (or, in some cases, both), which surface can then be developed into a plane. The projection onto a cone is known as the *Lambert conformal conic projection*. Distortions of this projection occur in the north–south direction; hence, it is used for states with relatively short north–south dimensions (such as Tennessee and North Carolina). The projection onto a cylinder is known as the *transverse Mercator projection*; it is used for states with relatively long north–south dimensions (e.g., Illinois and Mississippi). Some states utilize more than one projection of either kind, and some use both kinds of projection. For example, New York uses a Lambert projection for Long Island and three Mercator projections for the rest of the state. Florida uses a Lambert projection for the northwestern ("panhandle") part of the state and two Mercator projections for the lower ("peninsula") part of the state.

Some advantages of the SPCS, in addition to that of documenting forever exact locations of local surveys, are tying together separate surveys to a common system and providing checking capabilities to prevent excessive accumulation of errors of measurement in large-scale surveys.

B. Astronomical Observations

Astronomical observations can be used in surveying to establish directions by sighting on celestial bodies. Probably the most commonly performed observations are ones made to determine the direction of true north or of a line with reference to true north.

The true azimuth of a line can be determined by astronomical observation by setting up a transit over one end of the line, sighting on a particular celestial body, and reading the vertical and horizontal angles to the line in question. If the time of observation and location (latitude and longitude) of the observation point are known, the exact position of the celestial body can be determined from an almanac of celestial body positions (known as an *ephemeris*). From this information, the true azimuth of the line to the celestial body and then the true azimuth of the line in question can be calculated.

The celestial body most commonly used for determining the direction of true north is the North Star (Polaris). Fortuitously, it is located almost directly above the earth's geographic north pole; hence, a sighting on Polaris gives (approximately) the true north direction. In actuality, inasmuch as Polaris is not exactly directly above the geographic north pole, it (Polaris) appears to the viewer looking through a fixed telescope to move slightly back and forth (left to right to left, etc.) over a period of time. True north is, in effect, the average of the extreme left and right sightings of Polaris.

C. Photogrammetry

Various measuring techniques for gathering data to be used in preparing topographic maps have already been discussed. Although they are widely used and provide adequate results, they have at least two limitations, particularly when applied to surveys over large areas. One is the time (and cost) involved in obtaining required data;

the other is limitation on the actual amount of data that can be obtained. Both of these can be overcome by using *photogrammetry*.

In simple terms, photogrammetry consists of photographing the property in question from an airplane, viewing the aerial photographs, and preparing a topographic map therefrom. It should be clear that substantial savings in time are realized in gathering data by means of photogrammetry compared to more conventional means; and although the costs of operating an aircraft may be high, they are more than offset by savings in time and salaries paid. The fact that the camera "sees everything" ensures that a vast amount of data will be obtained and virtually no detail will be overlooked in the field.

In practice, a number of photographs are made as the airplane flies a prescribed course so that they will overlap one another. Maps are prepared by specialists who view the photographs through a stereoscope. Specialized plotting machines may also be used. As in other cases, adequate horizontal and vertical control systems must be employed to ensure sufficient accuracy.

Photogrammetry is used extensively nowadays by the U.S. Geological Survey in compiling its quadrangle maps. Significant improvements in cameras, film, plotting instruments, and other devices have made it possible to produce maps that meet high accuracy standards.

It is noteworthy that certain situations preclude the use of photogrammetry. For example, for small areas, it is generally not cost-effective. Also, in areas covered by sand (deserts or beaches) or snow, lack of change in color and texture of the ground makes interpretation of the ground surface difficult. Another problem is encountered in an area of extraordinarily dense growth (e.g., a rain forest), which, of course, makes it difficult or impossible to see the ground surface.

D. Recent Developments

Many recent developments in surveying, some of which have been mentioned, provide measurement with greater accuracy than could have even been anticipated a few years ago. Many developments occurred as a result of the space program, which required, on the one hand, very accurate determination of relative locations of space tracking stations around the world and, on the other hand, detailed mapping of the moon as well as planets and other stars. Whereas photogrammetry has been used for many years to "survey from above," nowadays this can be accomplished with earth-orbiting satellites.

Various modern devices utilizing lasers, infrared light, and so on provide extremely accurate determinations of elevations, distances, and directions. The computer revolution has affected surveying also. Microcomputers can automatically record data, make appropriate computations, and prepare maps.

Undoubtedly, efforts will continue into the future to perform surveying with even greater accuracy, in less time, and at a lower cost. Throughout it all, however, surveyors must be knowledgeable in the basic fundamentals of surveying and must constantly exercise good professional judgment in interpreting and utilizing results from the sophisticated instrumentation available.

SEE ALSO THE FOLLOWING ARTICLES

GEODESY • MINING ENGINEERING • NUMBER THEORY, ELEMENTARY • RAILWAY ENGINEERING

BIBLIOGRAPHY

Allan, A. L. (1997). "Practical Surveying and Computations," Butterworth–Heinemann, London.

Anderson, J., and Mikhail, E. (1997). "Surveying: Theory and Practice," 7th ed., McGraw–Hill Higher Education, New York.

Bannister, R. (1998). "Surveying," 6th ed., Addison–Wesley, Reading, MA.

Herubin, C. (1998). "Principles of Surveying," 4th ed., Prentice Hall, Englewood Cliffs, NJ.

Kavanagh, B. F. (2001). "Surveying with Construction Applications," 3rd ed., Prentice Hall, Englewood Cliffs, NJ.

McCormac, J. C. (1999). "Surveying," 4th ed., John Wiley & Sons, New York.

Moffit, F., and Bossler, J. (1998). "Surveying," 10th ed., Addison–Wesley, Reading, MA.



Ventilation, Industrial

Howard D. Goodfellow
University of Toronto and Stantec Global Technologies Ltd.

I. Ventilation Design Methodology
II. Ventilation Equipment
III. Design Equations for Industrial Ventilation
IV. Ventilation Modeling Using Computers
V. Ventilation Modeling Using Fluid Dynamics
VI. Solving Ventilation Problems for Existing Plants

GLOSSARY

Canopy Structure or enclosure located above a contaminant source to capture the rising contaminant into a ducted system.
CFD Computational Fluid Dynamics.
Dust Small particles created by the breaking of larger particles by mechanical action.
Face velocity Air velocity at the hood opening.
Froude number Dimensionless number that is the ratio of inertial forces to buoyancy forces.
Fume Small solid particles formed by the condensation of vapors of solid materials.
Hood Structure designed to enclose or partially enclose a contaminant-generating operation.
Industrial air technologies Air flow control technologies to control workplace indoor environment and emissions.
Mist Small droplet of materials that are ordinarily liquid at normal temperature and pressure.
Neutral zone or plane of neutral pressure Elevation within a building at which neither the outside air tends to move into the building nor the inside air tends to move out.
Reynolds number Dimensionless number that is the ratio of inertial to viscous forces.
Target levels Acceptable design levels for contaminants in an industrial environment.
Threshold limit values Airborne concentrations of substances for conditions under which it is believed that nearly all workers may be repeatedly exposed day after day without adverse effect.

THE FIELD OF INDUSTRIAL ventilation, or the more general, Industrial Air Technology (IAT) is a challenging field, which has been neglected by the scientific community until the last decade.

In all ventilation, the condition of the indoor environment called Indoor Air Quality (IAQ) and the exposures for the occupants are important. In industrial facilities, the contaminant emission rates may be 10–100 times higher than in nonindustrial facilities, but for many contaminants, the IAQ levels may be the same. From a design point of

view, the first priority is to consider the process, but other important issues, such as occupants, energy, environment, corporate image, etc., must also be considered.

The benefits of advanced industrial ventilation with improved IAQ are

- Improved health of workers and reduced absenteeism
- Improved work satisfaction, higher productivity, and reduced production failures
- Reduction in maintenance costs for the building fabrics, machinery, and products
- Reduction in energy consumption
- Opportunity to select new energy-efficient systems in ventilation design
- Environmental pollution is reduced by lower energy usage and lower emissions to the surroundings
- Embraces Clean Plant Design Concept
- Improve life cycle

Figure 1 is a schematic representation of the principles of industrial ventilation.

It is typical for industrial premises to have, in one space, zones with different target levels. The target levels may be determined for the whole area or locally. Often only a part of the space requires controlling of the indoor environment parameters. In addition to the main controlled zone, there may be one or more local controlled zones with different target levels than the main controlled zone.

Industrial air technology field is a more general description than industrial ventilation and includes measures to prevent harmful emissions from industrial processes to be discharged outdoors, conveying and cleaning technologies, and controlled discharge of exhaust air to outdoors. Industrial air technology systems include drying, process ventilation, and safety air systems.

The scope of the IAT field includes industrial process buildings, as well as hospitals, underground car parks, mining, railroad and vehicle tunnels, livestock buildings, and other similar premises and processes.

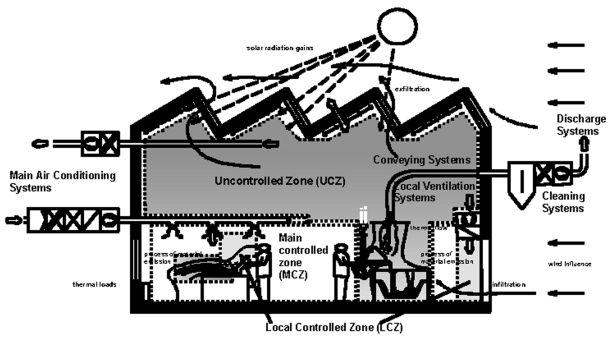


FIGURE 1 Zoning and IAT systems.

The more general IAT system category can be classified into three categories: (1) industrial ventilation, (2) process air technology, and (3) safety systems.

A more detailed breakdown of these systems is as follows.

INDUSTRIAL VENTILATION

- **Air conditioning systems**
  Air conditioning systems include control of air quality and thermal environment for both human occupancy and processes
- **General ventilation systems**
  In general ventilation systems, some indoor air parameters are controlled only partially. Target levels are usually lower than for air conditioning.
- **Local ventilation systems**
  These are used for local controlled zones. These systems are based on local capture of contaminants.
- **Process ventilation systems**
  In process ventilation, the target is to maintain defined conditions to ensure process performance, e.g., paper machine hoods.

PROCESS AIR TECHNOLOGY

- **Cleaning systems**
  Cleaning systems are used to remove contaminants, clean the resulting fluid flows and collect materials before discharging the exhaust air.
- **Pneumatic conveying systems**
  Conveying systems are used to transport captured pollutants from processes to a collection point.
- **Drying systems**
  Drying systems are used to remove moisture, gases, and vapors from the product

SAFETY AIR TECHNOLOGY SYSTEMS

- Designs to ensure safety from explosions

Typically, industrial premises will have, in one space, zones with different activities, which require different target levels for indoor environment and its control. These target levels may be determined for the whole area or locally. Also, often only a part of the space needs to be controlled. In addition to the main controlled zone, there may be one or more local controlled zones with different targets than in the main controlled zone. For example, machines equipped with electrical components require very clean and accurately controlled indoor environment, while the

unoccupied zone by the ceiling needs only a less stringent controlled protection against structural damages.

In industrial premises, the target levels of IAQ, as well as other targets, e.g., emissions, shall be specified zone by zone.

Controlled zone is a zone in which the thermal and air purity (quality) conditions are controlled to their specified levels. The two categories of controlled zones are

- Main controlled zone is normally a large area, which is often the same as occupied zone
- Local controlled zone, an area where the air is controlled locally, the control requirements may be for worker protection and comfort, for process control, or for production protection

Uncontrolled zone is a zone in which source emissions will be captured by source capturing system, and where the capture efficiency is determined and shall be maintained over the working period. From the pollutant concentration point of view, the capture zone is uncontrolled (e.g., workers shall not enter a capture zone without additional protection).

Room air conditioning systems are used for controlling the main controlled zone. Systems can be divided into subsystems:

- Air handling systems
- Air distribution systems (ductwork)
- Room air distribution systems
- Ventilation systems
- Room heating and cooling systems
- Main exhaust systems
- Discharge systems: stacks, environmental dispersion
  Note: Air distribution systems are not ventilation or air conditioning systems. For example, mixing air distribution or displacement air distribution are methods to bring the supply air to the treated space.

Discharge systems are used to discharge exhaust air to outdoors in such a way that harmful spreading of pollutants to environment and back to indoors is avoided.

Local ventilation systems are used for local controlled zones. These systems are based on local exhaust ventilation for local protection. Primarily local protection should be made using process methods such as encapsulation, process modification, source capturing, etc.

Local ventilation systems can be divided into the following subsystems:

- Local exhaust
- Local supply, including air curtains, i.e., control of air flow using jets
- Combined local supply and exhausts

Fans, ducts, and filters are subsystems of local exhaust ventilation.

Methods for Room Air Conditioning—Basic Strategies

Table I summarizes the strategies for room air conditioning. This table includes a description of the four strategies, main characteristics, and typical applications.

Industrial ventilation refers to the control of the environment with airflow as it applies to processing and manufacturing operations. The design methodology for ventilation systems is described on the basis of the development of ventilation system design as an integral part of the project planning and design activities. Procedures to be implemented for the development of technical specifications for louvers and roof exhausters are presented. The theoretical development of the design equations for heat release calculations, air set in motion, gravity or powered ventilators, and dilution ventilation is covered. Ventilation modeling is described for both computer applications and fluid dynamic techniques. Techniques for solving ventilation problems for existing plants are summarized.

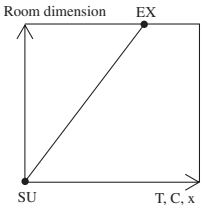
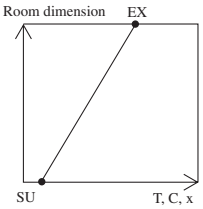
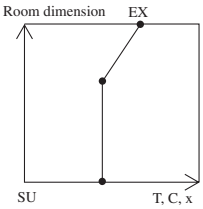
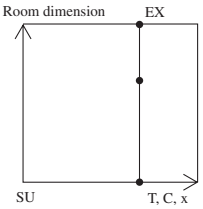
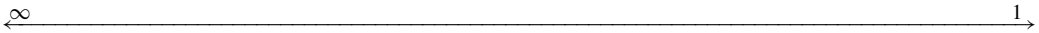
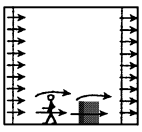
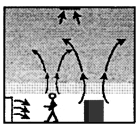
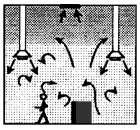
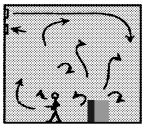
I. VENTILATION DESIGN METHODOLOGY

A. Background

Most industrial ventilation problems are complex. The development of cost-effective solutions requires an experienced and qualified ventilation engineer working with an industrial hygienist. Experience has shown that the four solutions to be considered for a ventilation problem, in order of priority, are as follows:

1. Process modifications to eliminate the contaminant problem can range from a total to a partial process change. For some operations, a change in the materials handling system may be required (e.g., type of system, frequency of operation, change in material composition or temperature).
2. If no acceptable process changes are possible, the application of local exhaust ventilation should be considered. Local exhaust ventilation usually requires ducted systems with hoods and covers. Figure 2 is a schematic of a local exhaust ventilation system. These systems are commonly referred to as dust control, fume control, and mist control systems. For local exhaust ventilation systems, the concentration of contaminant in the exhaust duct is significantly higher than in the general room area.

TABLE I Ideal Room Air Conditioning Strategies

Air conditioning strategy	Piston	Stratification	Zoning	Mixing
Description	To create unidirectional air flow field over the room area by supply air.	To support flow field created by density differences by replacing the airflow out from the room area with supply air	To control air conditions within the selected zone in the room by the supply air and allow stratification of heat and contaminants in the other room areas	To provide uniform conditions throughout the ventilated space
Heat, Humidity and Contaminant Distribution (Pictures) X-axis: °C, mg/m^3, %RH Y:axis: Room dim. (e.g. height) SU = supply, EX = exhaust				
Main characteristics:	Room air flow patterns controlled by low momentum unidirectional supply air flow, strong enough to overcome disturbances	Room air flow patterns controlled mainly by boyancy; supply air distribution with low momentum	Room air flow patterns controlled partly by supply and partly by boyancy	Room air flow patterns controlled typically by high momentum supply air flow
Ideal Contaminant- and Heat Removal Efficiency				
Typical application (An example of a general room air distribution method)				

3. For applications where local exhaust ventilation is not feasible, process building or general work area ventilation must be employed. For these systems, contaminated air is exhausted to the outdoors and large volumes of makeup air are introduced to dilute plant air contaminants to acceptable concentrations. For these industrial ventilation applications, the concentration of contaminant in the exhaust duct is not significantly higher than the contaminant concentration in the general room air. Figure 3 is a schematic of a general work area ventilation system.

For any specific industrial ventilation problem, the contaminated air can be exhausted by natural ventilation (taking advantage of buoyancy-driven forces arising from heat sources) or forced ventilation (using fans or roof exhausters). Design equations are developed in Section III for natural or forced ventilation systems for the control of heat, dust, and

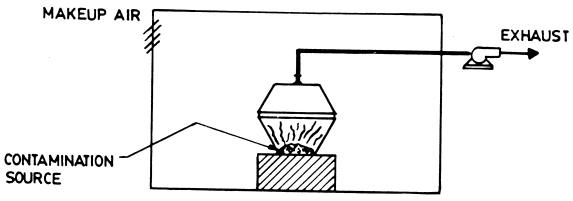


FIGURE 2 Schematic of local exhaust ventilation system.

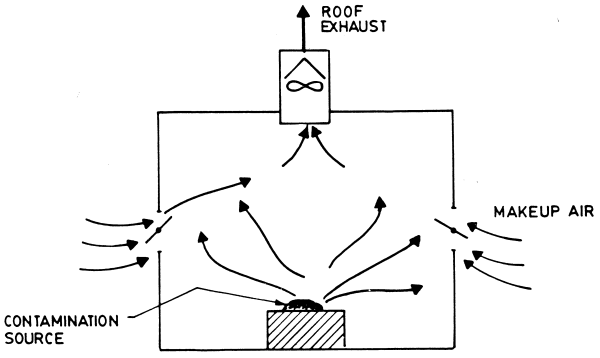


FIGURE 3 Schematic of general work area ventilation system.

fume for large industrial process buildings. Design equations are also developed for dilution ventilation applications as they apply to gases and volatile vapors.
4. Personal protection equipment, such as respirators, should only be considered for solving ventilation problems as an emergency or an interim maintenance requirement.

B. Design Methodology

On a global basis, significant progress has been made since 1990 in developing a systematic design procedure for industrial ventilation systems. The initial research and development effort was started in Finland in 1991 with the launch of the INVENT program and an investment of more than $20 million U.S. dollars. In 1995, a decision was made at a workshop in Zurich to start preparation of an international guidebook for industrial ventilation. An analysis of the current state-of-the-art identified the following issues:

1. No scientific basis for many design applications
2. No harmonization of design equations from different countries or researchers
3. Gaps in technical literature not defined and no roadmap identified for future technological developments
4. Many ventilation books are out of data
5. No longer acceptable to overdesign ventilation systems
6. No handbook in the industrial ventilation field
7. No accepted design methodology based on a rigorous scientific approach
8. Ventilation field fragmented on a global basis so need collaboration by a team of international experts
9. Excellent opportunity to collate worldwide research and development efforts into a single handbook
10. INVENT program started in Finland has generated momentum and a critical mass to make project successful

The design of a ventilation system must be incorporated into the plant design and layout at the earliest conceptual stage of the project. The ventilation engineer must work closely with the project design team at all stages of the project since the ventilation system can have a major impact on the type of process to be used, the plot plan, the building profile, and equipment layout within the plant.

In many applications, the ideal ventilation scheme may be in conflict with other environmental design criteria such as noise control and fume control. For example, ideal ventilation schemes may be based on the concept of a wide-open building to enhance natural ventilation: this is in conflict with noise control considerations that require the plant to be enclosed as much as possible to reduce neighborhood noise. Another example of this conflict is the use of canopies for remote fume capture. Canopies perform best if the building is totally enclosed to prevent cross drafts from disturbing the rising plume and causing it to miss the fume control system. Cost-effective ventilation systems can only be developed if the ventilation system design is an integral part of the project planning and design activities from the preliminary conceptual stage.

Figure 4 is an outline of the systematic design methodology procedure for industrial ventilation. This design methodology has been developed over many years of international collaboration and represents input from numerous engineers, researchers, scientists, and practitioners in the industrial ventilation field. A brief description of the different design steps follows below. (More details can be found in the "Handbook of Industrial Ventilation" edited by H.D. Goodfellow and E. Tahti, listed in the bibliography.)

C. Target Levels

It is difficult to achieve a good indoor environment in an industrial facility because of the lack of a scientific

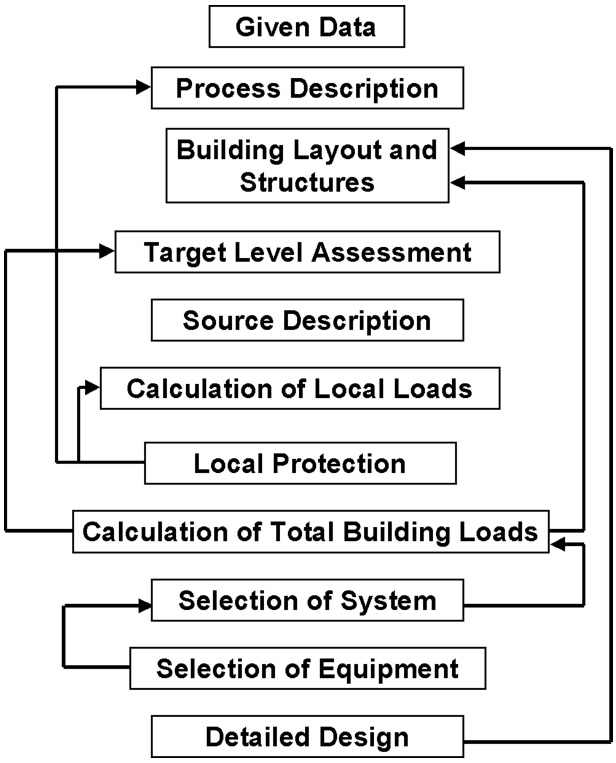


FIGURE 4 Outline of the systematic design methodology procedure for industrial ventilation.

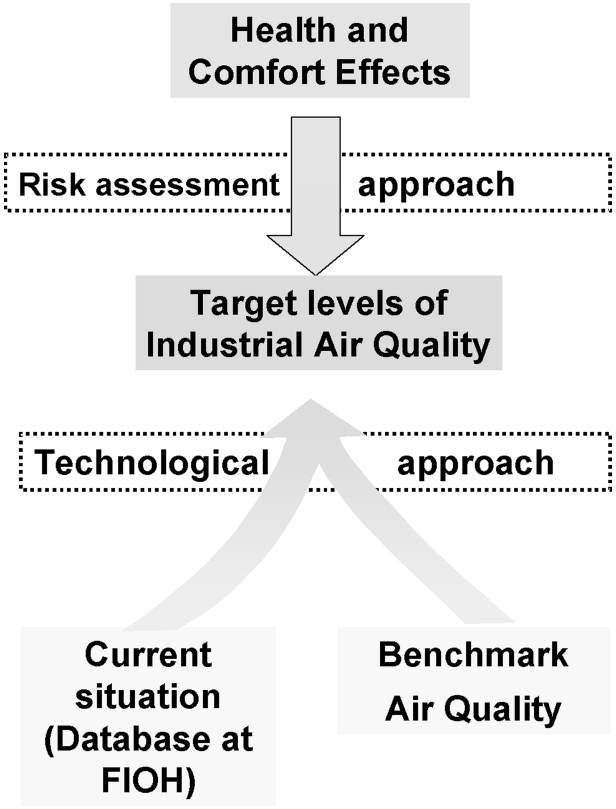


FIGURE 5 Design methodology for industrial ventilation systems.

basis for defining acceptable design levels for contaminants in an industrial setting. A detailed database for existing contaminant exposure is available from the Finnish Institute of Occupational Hygiene (FIOH). Figure 5 shows a recommended approach for assessing target levels of industrial air quality.

Target concentrations of air contaminants can be developed based on both human risk assessment and technology. Health risk estimates at low concentrations (e.g., below one-tenth of the current occupational exposure limits) are, in general, inaccurate and rather unreliable. Consequently, technology-based approaches for the assessment of target concentrations have been emphasized. The control technology approach is based on information on "current" concentrations achieved by standard practices and "benchmark" concentrations achieved by the advanced ventilation and production technologies.

Existing contaminant exposure data banks, such as available from FIOH and other international health and safety bodies, can be utilized for assessing concentration levels that are achievable. The benchmark air quality has been determined by measuring concentrations in factories equipped with the advanced technologies. This information on the benchmark air quality, the current concentration levels, and the health and comfort effects form the foundation for setting target levels for industrial air quality. This target level becomes the design input for establishing the industrial ventilation system design parameters. A detailed description of the use of target levels of air quality as a valuable design tool is presented in a "Handbook of Industrial Ventilation" by H.D. Goodfellow and E. Tahti. Details of a case history of the application of target level and design methodology for paper machine room ventilation are also included.

Step 1: Given data

- Identify and collect data that are site specific (i.e., climatic conditions, site elevation, etc.)
- Literature search to obtain published ventilation design data

Step 2: Process description

- Understand the industrial process and identify subprocesses
- Identify possible emission sources, occupational areas, effects of environmental parameters on production, needs for enclosure, and ventilation equipment
- Divide process in such parts that their inputs and outputs to the environment can be defined

Step 3: Building layout and structures

- Collect properties of building layout, structures and building envelope, and openings
- Complete zoning of building based on division of the process and building layout
- Identify layout requirements and structures required for ventilation systems
- Develop isometric of building showing all openings

Step 4: Target-level assessment

- Define target levels for indoor (zones) and outdoor (exhaust) conditions based on human risk assessment and technology
- Specific design conditions for which the target levels for ventilation system based on reliability, energy consumption, investment, and life cycle costs, etc.

Step 5: Source description

- Identify emission sources by type (chemical and physical properties)
- Complete typical data sheet to be used to develop a catalog of emission sources (Fig. 6) review requirements for heat load calculations

Emission Source Catalog

Description of Point of Emission	
Duration and frequency of Emission	
Coordinates of center of source (m)	
Geometry of point of emission (mm)	
Description of Pollutant: • Type of pollutant (particulate/gaseous) • Gas composition • Polluted Gas volume (standard) (NM^3/hr) • Estimated Temp at point of emission (°C) • Toxic or other special nature of pollutant	
If Pollutant is Particulate: • Chemical composition of particulate • Particle size : Typical (um) Maximum (um) • Average particulate loading (mg/Nm^3)	
Emission Control Philosophy	
Complete if the above emission is the Exhaust of Pollution Control Equipment • Type of dust collection and/or gas treatment equipment • Action to be taken for control failure • Estimate duration and frequency of failure	
Emission Rates and Emission Factors (Particulate/Gaseous) • Pollutant Generated at Source • Kg/min • Kg/hr • Kg/day • (Uncaptured) Fugitive Emission • kg/min • kg/hr • kg/day	
Additional Information	

FIGURE 6 Catalog of emission sources data sheet.

Step 6: Local protection

- Calculate heat loads from individual sources (see Section III)
- Use emission data or calculation models
- Calculate total volume of air set in motion
- Heat and flow balances

Step 7: Local protection

- Examine feasibility of local source control
- Calculate working conditions for different options
- Compare to target levels of local zone

Step 8: Calculation of total building loads

- Calculate total loads (heat, humidity, contaminants) from different subprocesses and building for specific zones of the building
- Account for time dependency of emissions
- Evaluate system performance and establish if target levels or building layout and structures need to be revised

Step 9: Selection of the system

- Select acceptable systems based on target levels
- Identify feasibility of different options selected
- Select the most effective ventilation system

Step 10: Selection of equipment

- Identify system specifications
- Select acceptable equipment based on performance characteristics
- Equipment recommendations
- Prepare a technical specification for selected equipment

Step 11: Detailed design

- Calculate dimensioning requirements for selected systems
- Detailed layout ventilation systems
- Design control system
- Consider special issues such as thermal insulation, condensation risks, fire protection, sound and vibration damping, maintenance, etc.
- Develop construction and commissioning plan

II. VENTILATION EQUIPMENT

A. Background

For industrial ventilation systems, the major equipment components are louvers and roof ventilators (gravity or powered). This equipment should only be purchased after the system design has been completed and proper technical specifications have been prepared by a qualified and experienced ventilation engineer. The technical specifications must clearly and concisely communicate the users' requirements or criteria to the bidders. Some important guidelines that will assist the engineer in the preparation of better technical specifications include the following.

1. Do your homework. The engineer preparing specifications must have an in-depth technical knowledge of the specified components.
2. Clearly define the scope of work. that is, what is included and excluded.
3. Strive for clarity and conciseness. Avoid ambiguous phrases, legalese, and repetition.
4. Discuss your requirements with suppliers first. What is available?
5. Define performance criteria and the required warranty. The basis of performance criteria must be well developed and must include details of testing procedures.
6. Aim for performance specifications. Avoid highly restrictive ones.
7. Do not reinvent the wheel. Modify standard specifications.
8. Supply pertinent process data, test results, information on operating conditions, and samples of contaminants.
9. Be specific. Minimize reference to general standards.
10. Attach a technical questionnaire to be completed by all bidders and used during bid analysis.

B. Louvers

In industrial buildings, the prime purpose of louvers is to control the location and flow of incoming air. Some design guidelines for industrial louvers follow.

1. The velocity of the incoming air must be low enough to prevent uncomfortable drafts on personnel ($\leq$0.5 m/sec).
2. Louvers must be located to account for wind patterns.
3. Any openings for louvers must minimize the penetration of rain or snow.
4. Airflow requirements must be reduced during the winter.
5. Louvers must be of rugged design and able to operate in hot, dusty environments as well as under freezing conditions.
6. Resistance to airflow must be minimized.
7. Louver designs must be architecturally pleasing.
8. A special screen design may be required to keep out insects or birds.
9. Acoustic treatment may be required.
10. Design must be flexible to accommodate the addition of heat to the incoming air.
11. Louver elevation must be high enough to avoid damage from snow removal or other equipment.
12. Louver elevation must be low enough to supply air without the occurrence of dead zones.

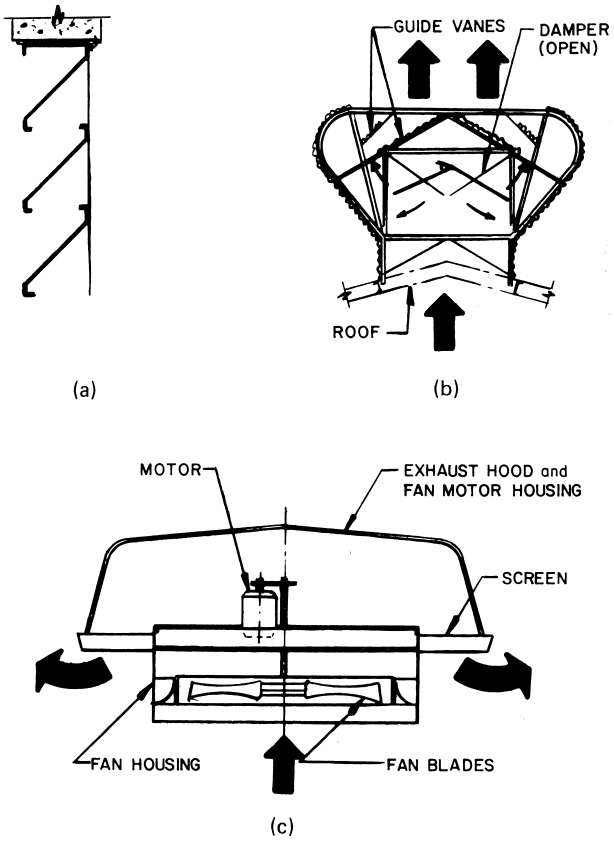


FIGURE 7 Typical ventilation equipment: (a) louvers, (b) roof gravity ventilators, (c) square or rectangular powered roof exhausters.

Louvers are available in a great variety of types, blade profiles, materials, and finishes. A typical industrial louver profile is shown in Fig. 7. Louver types can be fixed or operating, coordinated louvers, or custom special louvers. The operating types can be manual with a push bar spring and chain, gear box and shafting, pneumatic, or electric. Louvers come in a wide range of standard materials and finishes (e.g., galvanized, aluminum) as well as custom design louvers (e.g., stainless steel). Louvers may be located in a row around the perimeter of the building or may be stacked vertically in banks with proper support steel.

In the design of louvers for a ventilation system, the critical design features involve the air pressure drop at the specific velocity of the air (either expressed as a face velocity or velocity through the free area), as well as the weather penetration at this velocity. The specified airflow into the building through louvers is established on the basis of ventilation design calculations. For a specific application, the engineer, in consultation with the louver suppliers, can select the type of louver, material of construction, blade profile, and actual location of louvers. For guidance, selection should follow the design guidelines listed earlier.

With the known flow rate and a recommended louver face velocity, the louver size can be determined. For each type of louver, information is available from the louver supplier that relates pressure drop in millimeters of water to air velocities through the free area of the louver expressed in meters per second. The percent free area of the louver can be calculated from information supplied by the louver manufacturer. The actual velocity of air through the blades is calculated by dividing the flow rate by the louver free area. For this velocity through the blades, the pressure drop can be determined from a typical curve based on test work for the specific louver. The velocity usually selected is about 1.25 m/sec.

C. Roof Ventilators (Gravity or Powered)

Roof ventilators are widely used in industrial buildings and are the workhorses of the industry. Roof ventilators are devices designed either to introduce outdoor air into buildings or to exhaust air from them. Supply systems may bring outside air into specific work areas to control the in-plant environment, or they may supply makeup air to balance exhaust requirements. These supply units may be equipped with filters or humidifying sections to produce cleaned or conditioned air. Some supply units may be designed to recover heat or to add heat by means of hot water or steam coils to reduce the space-heating load. The primary function of a roof exhauster is the direct exhaust of heat, moisture, and contaminants, such as fume, smoke, noxious gases, or mists, from the workplace environment.

Ventilators may be gravity or powered. Gravity ventilators depend on thermal driving forces. with the hot air rising and being exhausted through the roof. Power ventilators or powered roof ventilators (PRVs) have motor-driven fans. They may be designed to exhaust air, to bring in outdoor air, or to operate in either mode as required.

Gravity ventilators are available in many different styles and materials of construction. Figure 7 shows a typical profile of a gravity ventilator. Gravity ventilators are sometimes referred to as monitors, streamline monitors, or guided-flow or sawtooth roof monitors.

Powered roof ventilators come in three basic configurations: round, dome, and square or rectangular. Figure 7 shows a schematic of a square or rectangular powered roof exhauster.

For almost all applications. both types of ventilators should be considered at the conceptual stage. A selection should only be made after proper technical and economic analyses have been completed. Some general guidelines for selecting either gravity or powered ventilators include the following.

1. The initial total cost will be much higher for gravity ventilators.
2. Operating costs for powered ventilators will be higher (power, maintenance).
3. Community noise is often a problem with powered roof exhausters.
4. Powered exhausters provide more flexible and better control. They can be located directly above the hot source, can operate in an on/off mode or with a two-speed fan, and can be turned off in winter.
5. A good feature of gravity ventilators is that they are self-regulating. The ventilation rate through a section will increase if the heat release from a source below increases.
6. Powered ventilators have higher discharge rates that help reduce ground-level concentrations and reduce the effect of wind.
7. Gravity ventilator designs can be used to complement the architectural features of industrial buildings.

The important design parameter for gravity ventilators is the throat area or flow area required for a given flow rate. Exhaust capacities must account for heat, height, and wind factors. It is also known that special building pressure conditions, air supply locations, equipment or structural obstructions, and the transient nature of heat releases all affect ventilation rates. For specific conditions, it may be necessary to perform modeling tests or a more rigorous calculation approach, as described in Section III.

Powered ventilators must be sized to handle the required flow at a static pressure by using tables available from the equipment suppliers. In many plants, poor flow conditions at either the inlet or outlet of the powered roof exhausters have a very significant effect on the ventilation rate. If a ventilation problem exists in a plant with powered roof exhausters, the first step in problem solving should always be to measure actual flow rates in the field.

III. DESIGN EQUATIONS FOR INDUSTRIAL VENTILATION

A. Heat Release Calculations

Detailed calculations can be carried out to establish heat releases from the different operations within the process building. Heat is released from surfaces by convective forces that cause thermal updrafts, thus making up the primary ventilation flows, and by radiant transfer. Radiant heat that is not ultimately converted into convective heat does not affect the primary ventilation flow. This would include radiant heat that is absorbed by the roof and walls and subsequently lost to the outside atmosphere. However,

radiant heat absorbed within the building and converted to convective heat does affect the primary ventilation flow rate. Both the steady and the intermittent heat releases are calculated. For design purposes, it is recommended that an average heat release be established to be somewhat higher than the steady heat release.

For convective heat losses, the heat release from equipment surfaces is calculated with the equation

$$q_c = h_c A(T_s - T_a)$$

where q_c is convective heat loss (kcal/sec), h_c the natural convective heat transfer coefficient in kilocalories (kcal/m^2 °C sec), A the surface area of hot source (m^2), T_s the surface temperature (°C), T_a the surrounding air temperature (°C). The convective heat transfer coefficient h_c can be readily estimated from well-established formulas that can be found in standard heat transfer reference books.

For radiant heat losses, the heat release from equipment surfaces is calculated by

$$q_r = \sigma \varepsilon F A\left(T_s^4 - T_a^4\right)$$

where q_r is radiant heat loss (kcal/sec), σ Stefan-Boltzmann constant (kcal/m^2 sec K^4), ε emissivity of substance (blackbody coefficient of the radiating surface), F view factor (usually assumed equal to 1), A surface area of hot source (m^2), T_s absolute surface temperature (K), T_a absolute surrounding air temperature (K). Emissivities for various surfaces are available in standard references.

The total heat release from the equipment surface is the sum of the convection and the radiation heat losses. The heat loss calculations provide the necessary data for the ventilation heat balance and total building heat balance.

B. Air Set-in-Motion Calculations

Air set-in-motion calculations account for major air and fume vertical flows that are produced by the total convective portion (may include some radiant heat loss that is converted to convective heat) of heat release from hot processes. This hot air is heated by contact with the surfaces of the hot equipment and rises due to thermal buoyancy. The size and velocity of this hot air column is a function of the heat release rate and the distance between the source and the roof level. The roof ventilators must accommodate the total flow of hot air delivered to them in order to ensure that the hot contaminated air does not recirculate to the lower working areas. It is imperative that the total volumes of air set in motion be calculated and compared to the exhaust rates of roof ventilators.

The heated air stream originating from the surfaces of hot bodies mixes turbulently with the surrounding air as it moves upward. An air stream at a hot body source may have a flow of 0.5 m^3/sec. After some vertical travel, this hot air stream may set in motion, and become mixed with (i.e., entrained in), an additional flow of 50 m^3/sec of surrounding air. To avoid recirculation, roof ventilators should, as a minimum, accommodate this total flow arriving at the roof level.

The following plume flow rate equations for point plumes (first equation) and line plumes (second) can be used to calculate air set in motion at different elevations:

$$Q = (6\pi\alpha/5)(18F\alpha/5\pi)^{1/3} z^{5/3}$$

$$Q = 2\alpha(F/\alpha)^{1/3} z$$

where Q is the plume flow rate at elevation z (m^3/sec) (for a line plume, Q is plume flow rate per unit length of source), α entrainment constant (point plume 0.093, line plume 0.156), z vertical distance from plume origin to roof truss in meters, and F buoyancy flux (m^4/s^3)

$$F = q g_c / C_p T_0 \rho_0$$

where q is source heat flux (kcal/sec), g_c acceleration due to gravity (9.81 m/sec^2), C_p ambient air specific heat (kcal/kg K), T_0 absolute ambient temperature (K), and ρ_0 ambient air density (kg/m^3).

C. Gravity or Powered Ventilation

In a classical technical paper in the ventilation field, published in 1926, Emswiler defined the concept of a neutral zone or plane of neutral pressure as the elevation within a building at which neither the outside air tends to move into the building nor the inside air tends to move out. Emswiler developed two theorems that apply to industrial ventilation and form the basis for design.

Theorem 1. The sum of flow rates into the building below the neutral zone must equal the sum of the flow rates out of the building above the neutral zone.

$$\sum_{i=1}^{n} A_i v_i + \sum_{j=1}^{n} Q_j = 0$$

where A_i is the area of opening (m^2), v_i the velocity through the opening (m/sec), Q_j the powered ventilation flow rates (m^3/sec).

Theorem 2. The driving force at each opening is related to the vertical distance between the opening and the neutral zone and the temperature difference between the opening and the neutral zone:

$$Q_N = A_N \sqrt{L_{nz} \Delta t_{nz} / (RT)}$$

where Q_N is the flow rate (m^3/sec), A_N the area of the opening (m^2), L_{nz} the distance from center line of the

opening to the neutral zone (m), Δt_{nz} the temperature difference between the opening and the neutral zone (K), R the resistance of the opening, and T the absolute temperature of the air at the opening (K).

The resistance of the opening R is given by

$$R = 1/2g_c C^2$$

where g_c is acceleration due to gravity (9.81 m/sec^2) and C the loss coefficient for the opening. Substituting for R in the preceding equation, the flow rate is given by

$$Q_N = 4.43CA_N\sqrt{L_{nz}\Delta t_{nz}/T} \quad (\text{m}^3/\text{sec})$$

For $C = 0.65$,

$$Q_N = (2.88/\sqrt{T})A_N\sqrt{L_{nz}\Delta t_{nz}} \quad (\text{m}^3/\text{sec})$$

Figure 8 shows the plane of neutral pressure for a hot process building. The elevation of the neutral zone is determined by trial and error using the preceding equations.

The temperature rise through a hot process building is a function of the amount of heat released in the building and the ventilation flow rate. The heat balance equation can be written as

$$\Delta t = q/\rho C_p Q$$

where Δt is the temperature rise through the building (°C), q the heat release rate (kcal/sec), ρ the air density (kg/m^3), C_p the heat capacity of air (kcal/kg °C), and Q the ventilation flow rate (m^3/sec).

Baturin has developed four equations that describe the ventilation process. These equations can be used to calculate air changes in single-span and multispan (commonly called multibay) process buildings.

The rate at which air flows through an opening of area A, with a pressure difference ΔP between the pressure indoors and outdoors, is given by

$$G = C_F A\sqrt{2g_c\rho\Delta P}$$

where G is the mass flow rate (kg/sec), C_F the discharge coefficient of flow (dimensionless), A the area of the opening (m^2), g_c acceleration due to gravity (9.81 m/sec^2), ρ the density of the air in the initial state (kg/m^3), and ΔP the difference between the inside and the outside pressure at a given opening (kg/m^2).

From the continuity equations for the steady state, the amount of air coming into the shop per unit of time (in kilograms per second) is equal to the amount of air leaving the shop in the same time period. Thus the air balance equation is given by

$$\sum G_{in} = \sum G_{out}$$

For the heat balance equation, the quantity of heat removed from the shop must equal the sum of the heat brought into the shop by the outdoor air plus the surplus heat given off in the shop in the same unit of time.

$$G = q/[C_p(t_{out} - t_{amb})]$$

where G is the mass flow rate of air leaving the shop (kg/sec), q the quantity of heat removed from the shop (kcal/sec), C_p the heat capacity of air (kcal/kg °C), t_{out} the outlet air temperature (°C), t_{amb} the ambient air temperature (°C).

For natural ventilation, wind has a significant effect on building ventilation rates. The effect of wind on a building manifests itself as an increase in pressure on its windward side and as suction on the leeward size. The pressure at any arbitrary plane on a building is a magnitude of pressure in

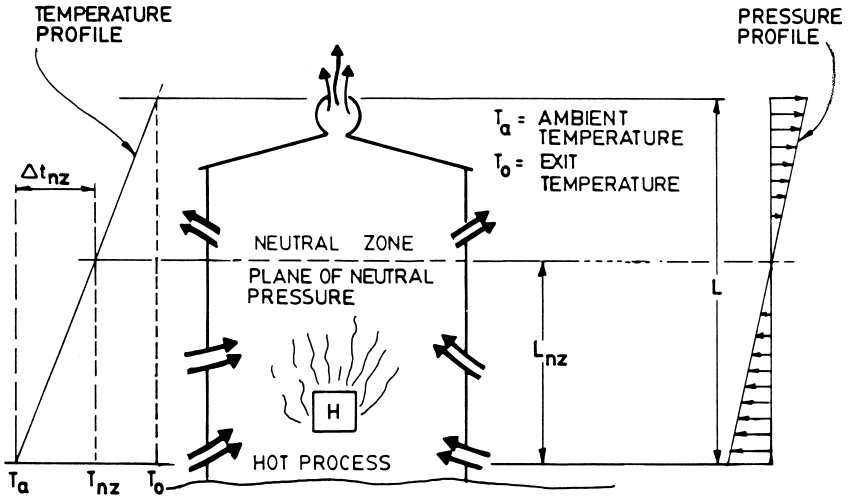


FIGURE 8 The plane of neutral pressure as defined by Emswiler.

excess of ambient pressure. This ratio is called the pressure coefficient and is dimensionless. Experiments have shown that pressure coefficients remain constant on geometrically similar buildings. Thus, pressure coefficients for buildings are found by wind-tunnel tests on geometrically similar models. The wind pressures are obtained by multiplying the pressure coefficient for the points concerned by the velocity pressure of the wind as

$$p = k_w(\rho v^2/2g_c)$$

where p is the wind pressure (kg/m^2), k_w the wind pressure coefficient (dimensionless), ρ the air density (kg/m^3), v the velocity (m/sec), and g_c the acceleration due to gravity (9.81 m/sec^2).

The preceding equations allow the amount of ventilation air required to meet allowable contaminant concentrations to be determined. The equations also allow the areas of the vents in the walls and roof to be calculated. For systems with mechanical ventilation as well as natural ventilation, the quantity of air that is supplied or removed by mechanical ventilation is written into the air balance equation on the inlet or outlet side.

D. Dilution Ventilation

Starting with a fundamental differential material balance, dilution ventilation requirements can be related to the generation and removal rates of a contaminant as

$$\underset{\text{accumulation}}{V\,dC} = \underset{\text{generation}}{G\,dt} - \underset{\text{removal}}{Q'C\,dt}$$

where V is the volume of the room or enclosure (m^3), C the concentration of gas or vapor at time t (ppm), G the rate of generation of contaminant (m^3/sec), t the time (sec), $Q' = Q/K$ the effective rate of ventilation corrected for incomplete mixing (m^3/sec), and K the design distribution constant or mixing factor, allowing for incomplete mixing.

Before solving the dilution ventilation equation for different cases, it is important to establish a procedure for estimating K, the mixing factor. Ventilation of contaminated air in a space would be simple if the outside air could enter the room in a laminar fashion without turbulence and remove the contaminated air in a piston fashion. For this condition, one air change would be required to completely remove the contaminated air ($K = 1$). In practice, however, the introduction of the fresh air results in turbulent mixing with the contaminated air, and after one air change, the room still contains a dilution mixture of fresh and contaminated air. For most industrial ventilation applications, K varies from 3 to 10. Factors that must be considered by an experienced industrial hygienist or ventilation engineer in selecting the K value include

1. Contaminant toxicity
2. Location and number of points of generation of contaminant in the room or work area
3. Location of air inlets and outlets
4. Duration of the process, operational cycle, and normal location of workers relative to sources of contamination
5. Geometry of enclosures or room
6. Reduction in operating efficiency of mechanical air moving devices
7. Seasonal changes in the amount of natural ventilation

The dilution ventilation equation will be solved for the following three cases: (1) case A, rate of contaminant concentration buildup, (2) case B, maintenance of acceptable concentrations at steady state, and (3) case C, rate of purging. Figure 9 is a graphical representation of these cases on a plot of concentration versus time.

1. Case A: Rate of Contaminant Concentration Buildup

Rearranging the dilution ventilation equation and integrating,

$$\int_{C_1}^{C_2} \frac{dC}{G - Q'C} = \frac{1}{V}\int_{t_1}^{t_2} dt$$

$$\ln\left(\frac{G - Q'C_2}{G - Q'C_1}\right) = -\frac{Q'}{V}(t_2 - t_1)$$

For the case of $C_1 = 0$ at $t_1 = 0$, this can be simplified to

$$t = -\frac{V}{Q'}\ln\left[\frac{(G - Q'C)}{G}\right]$$

2. Case B: Maintaining Acceptable Concentrations at Steady State

At steady state, $dC = 0$ and the preceding equation becomes

$$G\,dt = Q'C\,dt$$

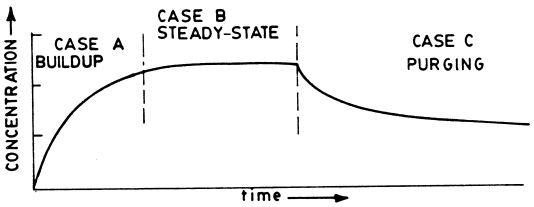


FIGURE 9 Contaminant concentration versus time for different dilution ventilation conditions.

For a uniform generation rate and a constant concentration

$$G\int_{t_1}^{t_2} dt = Q'C\int_{t_1}^{t_2} dt$$

$$G(t_2 - t_1) = Q'C(t_2 - t_1)$$

$$Q = \frac{KG}{C}$$

The preceding equation can be used to calculate the flow rate of uncontaminated dilution air required to reduce the ambient concentration of a hazardous material to an acceptable level. For liquid solvents, equations can be developed based on the steady state to determine the ventilation volume requirements for specific weights or volumes of solvent evaporated. In metric units, the equations are

cubic meters per liter of solvent evaporated

$$= \frac{24.1 \times \text{specific gravity of liquid} \times 10^6 \times \text{K}}{\text{molecular weight of liquid} \times \text{TLV}}$$

cubic meters per kilogram of solvent evaporated

$$= \frac{24.1 \times 10^6 \times \text{K}}{\text{molecular weight of liquid} \times \text{TLV}}$$

where TLV is the threshold limit value.

3. Case C: Rate of Purging

To calculate the rate of decrease of concentration of contaminant over a period of time for the case where a volume of air is contaminated and the contaminant generation process ceases (i.e., $G\,dt = 0$), the dilution ventilation equation is integrated as

$$\int_{C_1}^{C_2} \frac{dC}{C} = -\frac{Q'}{V}\int_{t_1}^{t_2} dt$$

$$\ln\frac{C_2}{C_1} = -\frac{Q'}{V}(t_2 - t_1)$$

or

$$t_2 = -\frac{V}{Q'}\ln\frac{C_2}{C_1}$$

For any specific contaminant, the design criteria may be based on regulated limits, comfort limits, or odor threshold concentrations. For many industrial solvents, the most stringent design criterion for a dilution ventilation system is based on the odor threshold concentration.

Ventilation by displacement is a concept developed and tested in laboratories in Norway and is based on ideal unidirectional flow (plug flow). For this situation, as shown in Fig. 10, no short circuiting takes place and the residence time for the room air is exactly the transit time (ideal case). A stratification takes place between the clean supply air and the contaminated air. Systems can be designed to utilize thermal or density stratification to create a tendency to unidirectional flow, that is, to create displacement flow.

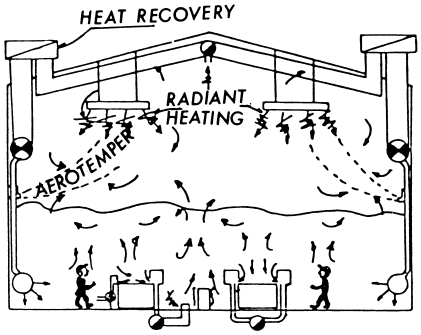


FIGURE 10 Displacement ventilation in workshops.

The benefits of displacement flow are that it improves the air renewal and contaminant removal speed and that it assists in maintaining favorable concentration gradients of the contaminants generated in the room. Calculation procedures based on mathematical two-zone models have been established.

IV. VENTILATION MODELING USING COMPUTERS

A mathematical model for building ventilation has been developed based on modifying Baturin's four ventilation equations. Except for the simplest cases, the solution procedure is iterative. For a building with many ventilation openings, the calculation procedure can become very time-consuming. Computer programs are available that can be applied for solving industrial ventilation problems. Figure 11, a flow diagram of the computer ventilation model, shows a typical input, the initialization and iteration loop, and the output. Further details are available in the book by Goodfellow, "Advanced Design of Ventilation Systems for Contaminant Control."

The computer model has been verified by using data from six independent surveys. Five of the surveys were carried out on three different days in a three-aisle process building approximately 275 m in length. The sixth survey was conducted in a multiple-aisle, high-production, BOF ingot teeming building with several hundred building openings. Results of the verification are shown in Table II.

With prevailing winds accounted for, the predicted and measured results are all within 20% of each other. Table II also gives the calculated flow rates for the no-wind condition. Results for calm conditions underestimate the total flow rate by 15–20%. The agreement between predicted model flows and measured flows for individual openings

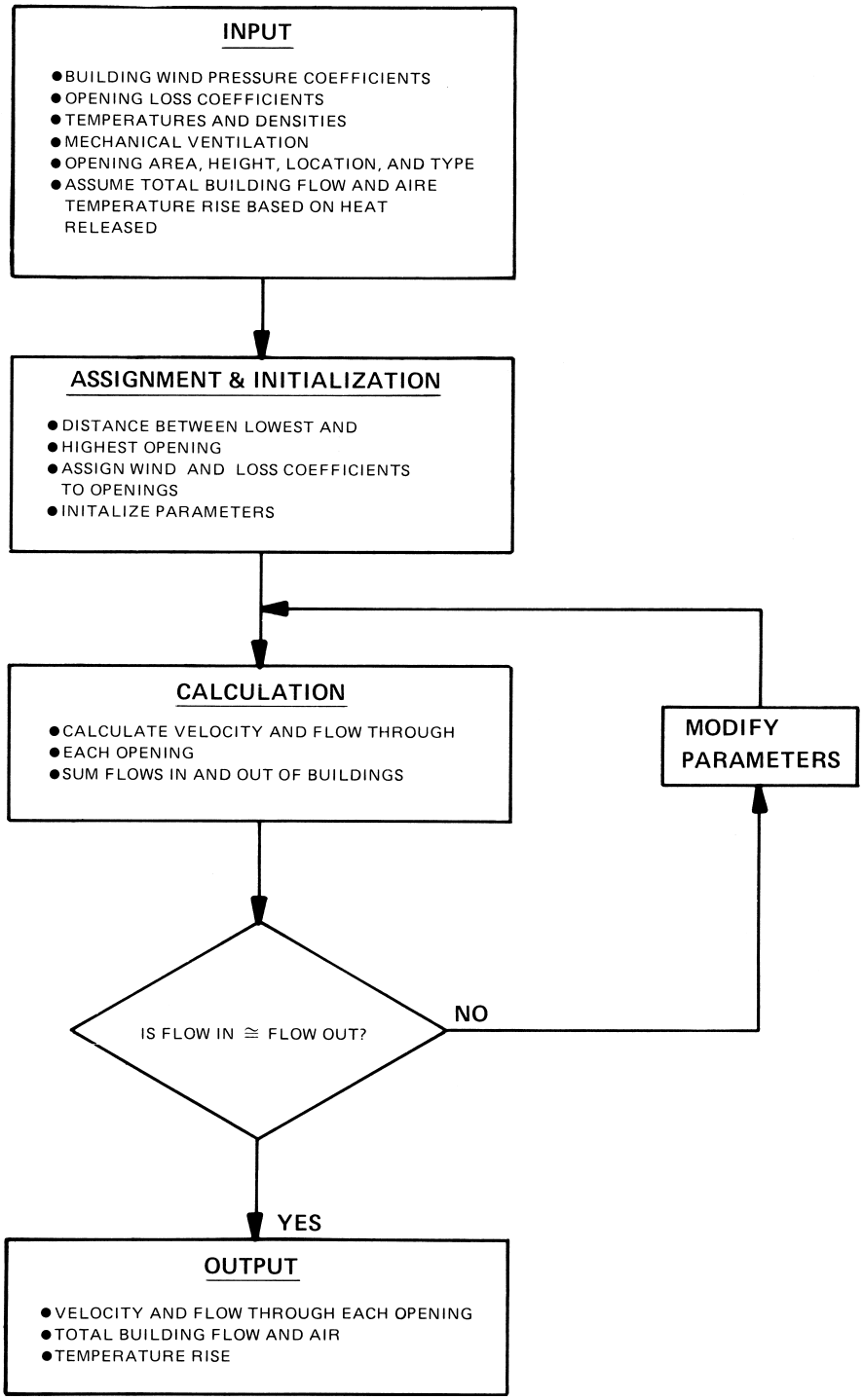


FIGURE 11 Computer ventilation model flow diagram.

has also been studied. Scatter of individual points does occur because of wind gustiness during measurement, nonuniform distribution of heat sources, and pressure disturbances due to local building features. However, the overall agreement is shown to be good.

Once correct modeling of observed ventilation flow rates is established, the computer program can be used to study the effects of different atmospheric conditions (summer/winter, wind direction and speed, etc.) on the ventilation characteristics for the shop. It is also a very

TABLE II Verification of Ventilation Computer Model

Survey number	Wind		Heat release (Mcal/sec)	Measured total flow (m^3/sec)	Model total flow	
	Direction	Speed (km/hr)			Wind (m^3/sec)	No wind (m^3/sec)
1	NNE	2.7	1.5	260	285	282
2	S	9.0	2.7	429	364	334
3	S	12.0	2.4	331	371	326
4	NNE	3.6	2.0	364	310	305
5	NNE	15.5	2.9	492	397	344
6	NE	16.2	20.0	5660	5190	4300

useful tool in the evaluation of proposed ventilation improvement schemes. Different ventilation schemes can be analyzed, and cost-effective solutions developed and implemented to solve the ventilation problems.

From the preceding discussions, it is apparent that the computer ventilation model can reliably predict the gross ventilation rates for complex process buildings. The use of high-speed computers provides the designer with the capability of examining the impact of architectural changes, wind conditions, or process changes on the performance of the proposed ventilation scheme. Problems such as contamination due to cross drafts or high temperatures in the work environment can be identified quickly and corrective measures taken.

The limitations of computer modeling are apparent as soon as predictions are required on the microenvironments inside the building. The steady-state and intermittent recycle flows caused by process or other heat releases convey contaminants from one area to another. These details of the internal flow fields are required by the designer to ensure an acceptable work environment for a new facility or to improve the work environment in an existing facility. Typical ventilation questions that may arise during the planning and design of ventilation systems include these.

1. What are the internal flow patterns under different layout and operating conditions?
2. What are the effects of intermittent peak heat releases on ventilation flow characteristics?
3. Can the flow fields be represented satisfactorily as a two-dimensional flow?
4. What happens to contaminated plume that misses a hood?
5. What effect do cross drafts have on the workplace environment?
6. Where does the fresh air enter the building?
7. What are the predicted contaminant concentrations in the breathing zone?
8. What is the source of contamination in a specific area?

For more complex ventilation and contaminant control, computational fluid dynamics (CFD) can be used. Governing equations similar to the ones presented earlier can be developed for the three-dimensional non-steady-state case. These equations represent the laws of conservation of mass, momentum, energy, and related entities, and their independent variables are allowed to vary in 1, 2, or 3 space dimensions and in time. The advent of high speed computers and the creation of general computer code systems allow the practical development of solutions to the type of internal flow questions listed above.

A number of items have been identified requiring further study in order to increase reliability of CFD simulations in large industrial premises. These are

- Identifying a turbulence model that provides a sufficient degree of reliability in the prediction of low Reynolds number regimes, near wall effects, and recirculating flows
- Understanding the relative importance of long and short wave radiation in an industrial setting to determine if a long-wave radiation can be omitted from simulations
- Identifying heat transfer correlations for natural, forced, and mixed convection that accurately predict the heat transfer off a surface
- Testing methods to provide for jets from diffusers in a ventilated room in order to accurately model the penetration length without resorting to fine grids
- Identifying the role of magnitude of the effect that potential errors in boundary condition specification might play on the final solution.

A technique that is well developed for quantifying internal flow fields is the use of fluid dynamic models. Using a small-scale replica of the building and the proper laws of similarity, the actual flow rates and velocities for the full-scale building can be predicted from measurements of the small-scale replica. The fluid dynamic modeling approach is described in Section V.

V. VENTILATION MODELING USING FLUID DYNAMICS

The technique of small-scale modeling for fluid flow problems (i.e., fluid dynamic modeling) has been used extensively for a wide variety of industrial ventilation applications. The application of fluid dynamic models has the advantage of enabling numerous different test conditions to be examined. The costs for performing modeling tests are usually much less than the cost for performing a full-scale field test. A typical range of costs for modeling studies is $20,000–$100,000, depending on the scope and complexity of the model and testing methodology. For new facilities, the designer does not have the option of field testing. Hence, fluid models can provide useful design data at the conceptual and preliminary engineering stages of a project.

The applications of fluid dynamic modeling techniques include the following:

1. External flow fields (e.g., to model pressure distribution, uncontrolled building emissions, stack discharge)
2. Internal flow fields (e.g., to model movement of large air masses far from enclosing walls, flow from one bay to another, flow around obstacles such as cranes)
3. Individual sources of heat and contaminants (e.g., to establish minimum capture volume for sources).

The general theory behind the use of scale models is well covered in numerous textbooks on fluid mechanics and in the technical literature. Therefore, only an outline of the approach used in fluid modeling for ventilation applications follows.

Data measured in the model flow may be related quantitatively to the full-scale prototype flow by establishing dynamic similarity between the model and the prototype. The two conditions necessary to establish dynamic similarity are (1) exact geometric similarity, which requires that the linear dimensions of the model are in the same proportion as the corresponding dimensions of the prototype, and (2) kinematic similarity which requires that the flow regimes be the same for model and prototype.

Kinematic similarity is achieved by matching governing dimensionless groups that describe the flow regime. For modeling ventilation systems, the governing dimensionless groups for scaling model (m) to prototype (p) are the Reynolds number and the Froude number. The Reynolds number similarity criterion is not a critical scaling parameter because prototype flows are almost always fully turbulent due to large Reynolds number. Flow patterns are similar in the different geometric scales as long as the important flow fields are fully turbulent. Therefore, the Reynolds number similarity criterion is met by ensuring that the flow in the model is fully turbulent. For processes involving hot gases (i.e., buoyancy-driving forces), the Froude number similarity criterion can be applied to the model and the prototype.

The flow required in the prototype is given by

$$Q_p = Q_m(S)^{5/3}(q_p/q_m)^{1/3}$$

where Q is the volumetric flow rate (cubic meters per second), q the heat flow rate (kilocalories per second), S the model scale (e.g., for a 1: 10 scale model, $S = 10$), p the subscript identifying prototype parameter, and m the subscript identifying model parameter.

Figure 12 is a typical flow chart of activities required for a ventilation modeling study. The first step is to define the contaminant and the source characteristics. Parameters to be defined are the size of the process building and details of the source flux (i.e., heat and contaminant release rates, etc.). Information is required on the major sources of heat in order to calculate heat balances and volumes of air set

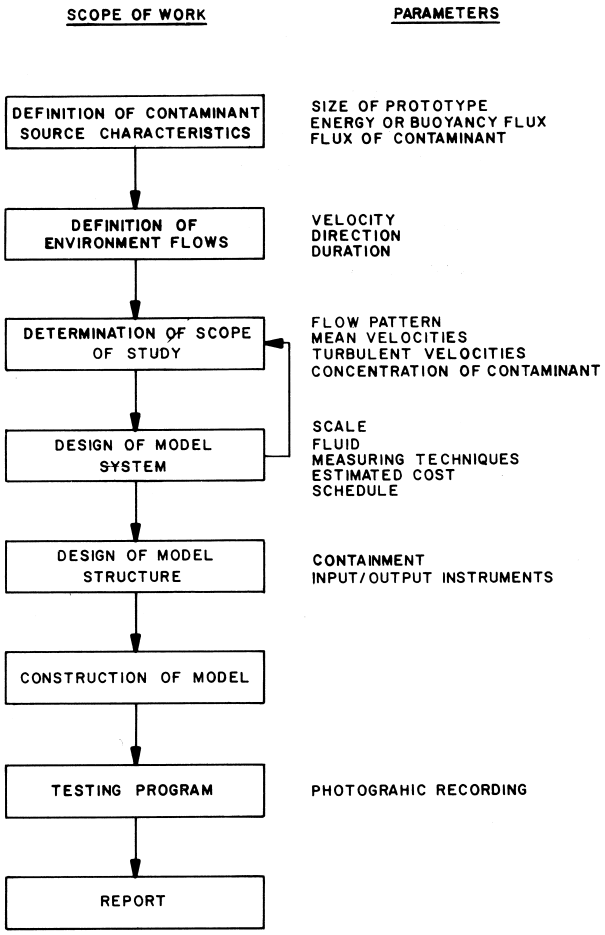


FIGURE 12 Typical flow chart for ventilation modeling study.

in motion. Data are required on the external and internal flow conditions for the prototype. Information may be required on site conditions such as wind, speed, direction, and frequency.

The design of the model system requires an examination of the scaling parameters and the type of fluid medium to be used. The common fluids used are air or water, but the working or buoyancy-driven fluids for models can vary widely. Fluid systems that have been used include air and heated air, water and saltwater, water and carbon tetrachloride, and mercury and carbon tetrachloride. Air models are usually the simplest and least expensive models to build and test. The use of air may require a large model to ensure fully turbulent flow. Because water has a smaller kinematic viscosity than air, a smaller model is required to ensure a high Reynolds number and turbulent flow. Flow visualization is also easier with water-based models because velocities are lower than in air. A detailed analysis of the scaling parameters and possible fluid systems is required in order to select the best modeling technique for the specific ventilation problem. The measuring technique for the contaminant of concern must also be considered since it may have a significant impact on the level of accuracy and the cost of the testing program.

Testing programs for many modeling studies are brief compared to the time required to design and construct the model system. For example, it is not uncommon to require 8–10 weeks to design and construct a model. 1–2 weeks to commission and calibrate the model, and 1–2 days to perform actual tests and develop solutions for the ventilation problem. For any model testing program, extensive use should be made of photographic and video recording equipment. The photographs and films will be invaluable in analyzing the test results and for subsequent presentation of solutions to management.

The fluid dynamic modeling technique is a cost-effective, flexible, and powerful tool for the design of ventilation systems for a greenfield plant and for the elimination of ventilation problems in existing plants. Some of the applications of scale modeling include

1. Finalizing building ventilation rates and schemes
2. Examining internal flow patterns and contaminant concentrations at any location
3. Examining external flow patterns, including quantitative measurements of downwash and transport of contaminants to other buildings
4. Establishing the effectiveness of source hoods.

For new process buildings, the recommended sequence is to use the mathematical models and equations to develop the overall ventilation concepts and architectural constraints. At this stage, the project design team can continue with the design of the structural steel while the small-scale model is being constructed and tested. The results of the small-scale modeling are used to refine the ventilation concept and to finalize all requirements.

For an existing plant, the fluid dynamic model is a valuable tool in modeling the problem and developing alternative cost-effective solutions.

VI. SOLVING VENTILATION PROBLEMS FOR EXISTING PLANTS

Ventilation flows and design parameters are unique for each specific process building. For most complex ventilation problems, it is recommended that both analytical calculations and a field measurement program be carried out

TABLE III Major Activities for a Ventilation Field Testing Program

Activity	Specific items
1. Information gathering	Obtain drawings, reports, operating procedures Define problem (summer, winter, heat, supply, exhaust) Review existing data and studies Visit site
2. Data on ventilation openings	Isometric drawings: plans, sections Schedule of openings
3. Plant questionnaire	Identifies factors specific for each plant Identifies significant heat sources in building Identifies typical operating practice Identifies gaps in data to be filled in by field testing program
4. Develop details of field testing program	General considerations Building ventilation In-plant flows Field data log sheets for ventilation measurements, weather conditions, plant operating records
5. Field testing program	Field measurements (volume, temperature, etc.)
6. Plant ventilation flow balance and in-plant flow patterns	Summary table of flows Sketches to show in-plant flows
7. Meltshop heat balance	Total plant heat balance calculation Ventilation heat balance calculation
8. Volumes of air set in motion	Based on surface temperature and convective heat release
9. Computer simulation of ventilation flows (natural ventilation)	Calibrate using field test data
10. Field testing report	Summarize test conditions and test results

in order to develop a sound, cost-effective solution to the ventilation problem. Experience has shown that the field sampling program must be developed in detail and must be tailormade for each specific ventilation problem. However, there are many elements of the field testing program that are common for any problem. Although there may be minor variations for any specific ventilation project, the 10 activities listed in Table III represent the major activities for a field testing program and the sequence for performing this work. A brief description of the scope of work required for each activity is included.

As a starting point, for the field testing program, obtain all pertinent reports and drawings on the plant for review and study. Specific information is required on the current operating practice, as well as information on the nature of the in-plant dust and heat stress problems. A visit to the site should include a walk-through ventilation survey using a data sheet as shown in Fig. 13.

It is necessary to develop the details of a field testing program and all necessary field data log sheets prior to the actual field testing. The development of a proper field testing program is the most important step in ensuring a successful ventilation study. This program should be reviewed and approved by plant operating personnel prior to commencement. Because the ventilation flow in large process buildings is usually complex, an inexperienced sampling team will often collect insufficient data for subsequent analytical calculations or will spend a considerable amount of effort collecting field data that are irrelevant. A good approach is to perform preliminary calculations on all sources and then to clearly identify where information gaps exist and what essential data must be collected.

STEP 1. *Site location*
- Name and location of the plant
- Date of survey
- Plant drawings (plans, layouts, elevations)

STEP 2. *Plant construction*

Wall material:
- Metal siding
- Metal siding plus insulation
- Brick

Concrete

Roof material:
- Metal siding
- Metal siding plus insulation
- Concrete
- Built-up, i.e. wood plus asphalt

Windows:
- Single glazed
- Double glazed

External access doors complete with airlocks

Louvres:
- High level
- Low level

Full height internal partitions

STEP 3. *Plant operations*
- Type of process
- Sources of heat emission
- Sources of fume emission
- Sources of dust emission

STEP 4. *Plant internal conditions*
- Fume apparent
- Dust apparent
- Hot
- Cold
- Noticeable air movement

STEP 5. *Type of ventilation equipment*
- Roof exhaust fans
- Roof supply fans
- Wall exhaust fans
- Wall supply fans
- Roof makeup air units
- Wall makeup air units
- Unit heaters

Ductwork:
- Supply
- Exhaust

Ductwork Construction:
- Sheet metal galvanized
- Fiberglass
- Stainless steel

Ductwork Condition:
- Corroded
- Plugged with material (exhaust)

STEP 5. *Type of ventilation equipment* (Cont'd)

Exhaust canopies located over:
- Process fume areas
- Process dust areas
- Process heat emission areas

Heating media used:
- Gas
- Steam
- Hot water
- Electricity

STEP 6. *Condition of ventilation equipment*

Age of equipment:
- 10–20 years and older
- 5–10 years
- 0–5 years

All equipment in operation

If some equipment de-energized:
- Bad maintenance
- Incorrect application
- Equipment failure

FIGURE 13 Walk-through ventilation survey data sheet.

The measurement of air velocities through all the openings must recognize the need to have a "representative" velocity for the survey. It is not acceptable to "be on the safe side" by rounding readings to a higher value. Gusty wind conditions will cause velocity readings to fluctuate. Sufficient time must be allowed to elapse to enable the individual monitoring the instrument to arrive at a representative velocity by "integrating" the observed values in his or her mind.

A temperature measurement is required for each velocity measurement. For air entering the building, this would generally be the ambient temperature. The ambient temperature should be recorded on an hourly basis. It will change during the course of the day and could be higher in the wake of the building. Do not expose the thermometer or temperature probe to sunlight or radiant heat from hot objects. The temperature of air leaving the meltshop will vary considerably. For heat balance calculations, good temperature readings are important.

Temperatures associated with in-plant flow patterns must also be recorded. Sufficient temperature data must be available to allow evaluation of the air density distribution within the shop.

Mean surface temperatures of hot surfaces must be recorded for subsequent heat release and air-set-in-motion calculations. The location of the equipment on the floor plan and equipment surface temperature should be recorded on a separate sheet.

Weather data should be measured at the site and obtained from the nearest airport or meteorological station as well. Data to be recorded include ambient temperature, relative humidity, and wind speed and direction. During the test period, this information should be recorded on at least an hourly basis.

A record of plant activities during the testing program is required. Data to be included are the status of operation of all major process and environmental equipment, as well as production levels.

An in-plant survey will be required to establish parameters such as dust levels and heat stress. These measurements would be concurrent with the ventilation survey. An industrial hygienist would work with the ventilation engineer to establish the scope and extent of the industrial hygiene sampling program.

For any ventilation field testing program, it is essential to prepare a proper engineering report, which includes all the field data, calculations, and test results. Using the results of this field testing program, an experienced ventilation engineer can develop cost-effective solutions for any plant ventilation problem.

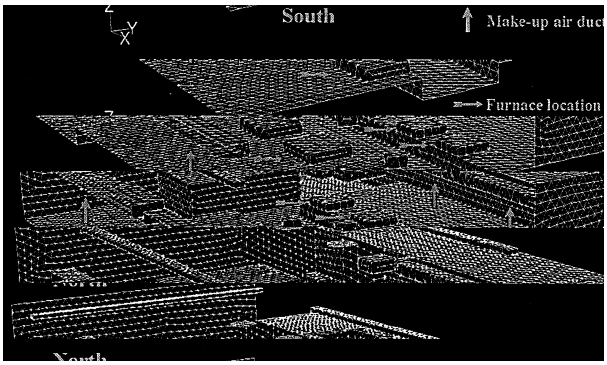


FIGURE 14

Computer models based on computational fluid dynamics (CFD) can be used to predict the velocity vectors and concentration of contaminants as a function of position and time. Figure 14 shows an industrial CFD case study with momentum sources. For this case, outputs from the CFD program can be used to establish airflow patterns for summer and winter conditions and for different process operating scenarios. Data from the computer models can be validated from the field measurement program.

SEE ALSO THE FOLLOWING ARTICLES

ENERGY FLOWS IN ECOLOGY AND IN THE ECONOMY • ENVIRONMENTAL TOXICOLOGY • FLUID DYNAMICS (CHEMICAL ENGINEERING) • HEAT TRANSFER • MACHINE DESIGN • POLLUTION, AIR

BIBLIOGRAPHY

American Conference of Government Industrial Hygienists (1998). "Industrial Ventilation: A Manual of Recommended Practice," 23rd ed., ACGIH, Cincinnati, Ohio.

Baturin, V. V. (1972). "Fundamentals of Industrial Ventilation," 3rd ed., Pergamon, London, U.K.

Goodfellow, H. D. (1985). "Advanced Design of Ventilation Systems for Contaminant Control," Elsevier, Amsterdam.

Heinsohn, R. J. (1990). "Industrial Ventilation: Engineering Principles," John Wiley & Sons, New York.

Goodfellow, H. D., and Tahti, E. (2000). "Handbook of Industrial Air Technology (DGB)," Academic Press, San Diego.