FUNDAMENTALS OF ELECTRONICS

1. From vacuum tubes to transistors. The *science of electronics* was founded on the ability of the electron tube (most importantly, the vacuum tube) to generate, amplify, and control an electric signal to accomplish a wide variety of functions. During its half-century dominance of the electronics art, the vacuum tube fostered the invention of radio, television, and radar, of factory automation and computers, and indeed of all the phenomena that we associate with electronics even today. Yet, with the invention of the transistor by Bell Laboratories in 1948, the tube was headed for extinction. Today, except for a few special functions, it no longer occupies an important place in electronic technology.

In place of the electron tube, the transistor has emerged as the cornerstone of modern electronics. Based on the theory of electron conduction in a solid crystalline material rather than in a vacuum or a gaseous environment, the transistor not only can perform virtually all the functions formerly associated with the tube, it can do them faster, more cheaply, and more reliably. Moreover, it occupies an infinitesimally small amount of space and, unlike the tube, requires no power-wasting filament that ties large equipment to the commercial power lines.

The dramatic effects of these advantages are most readily evident in the field of computers. The most advanced computer in the vacuum-tube era costs millions of preinflation dollars; consisted of an entire roomful of equipment utilizing tens of thousands of vacuum tubes; required a power plant just to heat the filaments and an air-conditioning plant to keep the computer cool; was difficult to keep in operation owing to vacuum-tube failures; and had less capability than today’s simple calculator, which fits into a vest pocket, requires only an occasional battery replacement, and is thrown away at the first sign of trouble because of its low replacement cost.

But while the transistor is the cornerstone of today’s electronics technology, it is not, in itself, the means for accomplishing the exponential expansion of electronic-capabilities...
2. **Integrated circuits.** There are two types of integrated circuits (IC) in use today: monolithic and hybrid.

A monolithic IC consists of a semiconductor substrate, called a *chip* or *die*, in which all necessary active and passive components are formed and interconnected to perform a desired circuit function (Fig. 6.1a). The chip is then encased in a suitable package, which may subsequently be interconnected on a printed-circuit board with other monolithic ICs and/or discrete components to form circuits of greater scope and complexity.

A hybrid IC (Fig. 6.1b) consists of an insulating substrate with deposited passive elements (resistors and capacitors) and metallized interconnected patterns, to which active elements are
added in uncased chip form either by wire bonding or by other die-attachment techniques. The active elements may be discrete devices or monolithic ICs. Hybrid circuits can be considerably smaller than circuits of equivalent complexity made by interconnecting the same components in encapsulated form on a more conventional printed-circuit board. However, they do not have most of the major attributes of monolithic ICs and are utilized principally where size reduction is required but where the required performance cannot be built into a monolithic structure.

3. Characteristics of monolithic integrated circuits

Size and Weight Reduction. Monolithic integrated circuits consist of a number of transistors and associated resistors fabricated and interconnected within the same monolithic piece of solid semiconductor material. Since these functional elements are fashioned by a slight change in the atomic structure of the material rather than by the addition of actual physical elements, the dimensions of these elements are so small that it is possible today to fabricate and interconnect tens of thousands of individual components within a piece of material no larger than the fingernail of a small child. And researchers are already anticipating circuits of million-transistor complexity within a single monolithic building block of similar dimensions.

Obviously, it is the size and weight reduction of integrated circuits, compared with equivalent vacuum-tube circuits (or even those made from individual discrete transistors), that has put the power of a full-scale computer into a pocket or purse. But size and weight reduction have long ceased to be the principal objectives on integrated circuits. Though vitally important in certain applications such as medical electronics and space apparatus, the physical characteristics of equipment are dictated more by the dimensions of the peripherals—the display and keyboard of calculators and computers, the loudspeakers and picture tubes of radio and TV, etc.—than they are by the associated electronics. Indeed, the size and weight advantages of integrated circuits would be of limited importance if they were not paced by corresponding reductions in cost and the improvements in component reliability and circuit performance.

Cost Reduction. The cost of the active elements associated with transistors, particularly transistors within complex integrated circuits, borders on the insignificant. This is so because semiconductor devices are made by batch processing, in which individual manufacturing processes operate not on a single device but simultaneously on a batch of devices ranging from hundreds of integrated circuits to thousands of discrete transistors. The basic cost of such devices therefore reflects primarily the packaging, testing, and marketing such devices rather than the raw materials or manufacture of the basic element. As a result, discrete small-signal transistors can be purchased for only pennies (in large quantities), and complex ICs housing up to thousands of fully interconnected transistors are available for only a few dollars. Each transistor serves the function of an erstwhile vacuum tube, whose individual cost equaled or exceeded that of many of today’s complex ICs. It is clear therefore that an attempt to duplicate the functions of today’s very-large-scale integrated circuits (VLSI circuits) before the era of semiconductor technology would have priced the resulting equipment out of the market.

Improved Reliability. Semiconductor electronics is based on the movement of free electrical particles (electrons and holes) within a loop of solid materials. Once such particles have been introduced into the material during manufacture, their movement can be controlled but the particles themselves cannot be destroyed. The basic semiconductor device therefore has no inherent wear-out mechanism. While the packaging of semiconductor devices does introduce some potential failure modes and while such devices can be damaged...
through overvoltage and excessive heat, modern manufacturing technology and circuit design techniques have reduced such failures almost to the vanishing point. It is this attribute of semiconductors that makes possible today’s highly complex electronic equipment utilizing millions of transistors.

**Improved Performance.** While transistors have replaced vacuum tubes in most applications, there are a few application areas in which tubes still dominate. Primarily, these are the areas of very high power and very high (microwave) frequencies. Since ICs are based on transistor technology, it is clear that they have not yet penetrated these applications to a significant degree. In all other applications, particularly those involving complex circuitry utilizing many transistors, the IC not only represents the most cost-effective component but also provides the best performance.

This performance advantage is the result of the close spacing between transistors within the monolithic substrate of the IC, which greatly reduces signal-propagation delay within a large system. Propagation-delay time is composed of two constituents: (1) the time required for a circuit stage (transistor) to perform its operation and (2) the time required for the signal to travel to the succeeding stage or stages. In the early days of transistors, the reaction time of the transistor was relatively slow so that interstage travel time was not a significant factor in total propagation delay. Today’s transistors, however, can operate at subnanosecond speeds. Thus, even if we assume that a signal travels along its conductive path between transistors with the speed of light (approximately 1 ft/ns), signal travel time through large systems composed of discrete transistors can become a significant part of the total delay. Within ICs, the spacing between transistors can be held to a few micrometers, eliminating travel time as a major factor of overall operating speed for most electronic applications.

**DISCRETE SOLID-STATE COMPONENTS**

Integrated circuits have penetrated all digital electronic equipment. They are still limited, however, in the amount of power they can produce, and their uses in linear applications have not yet been fully exploited. Discrete solid-state components therefore are still very much in evidence, and there is no indication that their use rate will diminish significantly in the immediate future.

Among the most prevalent discrete components are transistors themselves (both signal transistors and power transistors), as well as zener diodes and rectifiers, thyristors, and a growing family of transducers such as light-, temperature-, and pressure-activated devices. Most of the nontransistor components had counterparts in the electron-tube era, but they are subject to the same benefits of smaller size, improved performance, greater reliability, and lower cost when converted to the solid state.

| 4. Fundamentals of solid-state devices. | Solid-state devices, whether discrete or IC, are fabricated from semiconductor materials, so named because their conductivity lies somewhere between that of a conductive material and that of an insulator. While a variety of semiconductor materials (germanium, selenium, copper oxide, etc.) have been used to fabricate solid-state devices in the past and still others (notably gallium arsenide) are coming into limited use for special purposes, the most common material in current use is silicon. This material is used for both discrete components and integrated circuits. A basic understanding of its properties therefore is essential to an understanding of today’s solid-state devices. |
| **The Silicon Atom.** | From an electrical standpoint, the silicon atom (Fig. 6.2) consists of a nucleus containing 14 protons (each with an electrical charge of +1) and three orbital rings |

---

**SOLID-STATE DEVICES AND CIRCUITS**

**6.4 DIVISION SIX**

---
housing a total of 14 electrons (each with a charge of \(-1\)), which electrically neutralize the protons of the nucleus. The outermost ring contains four electrons. Since the electrons in the two inner rings are tightly bound to the nucleus, they contribute nothing to the electrical properties of the atom and therefore can be disregarded. The four outer-ring electrons, however, are loosely bound to their neutralizing protons and can easily be dislodged from their orbits by some external force.

**The Crystal Structure.** The atoms of silicon tend to arrange themselves into a crystal structure. With proper processing, the material can be of single-crystal form, in which all the atoms are arranged in a continuous, well-ordered lattice (Fig. 6.3), whereby each of the four outer-ring electrons associated with one atom forms a bond with one of the electrons

![Covalent bonds](image)
of a neighboring atom so that every electron of every atom is tightly bound within the crys-
tal lattice. Thus, while the four outer electrons of an individual silicon atom can easily be
removed by an external force, once they have been bound in the lattice, they are very dif-
cult to dislodge. Intrinsic (pure) single-crystal silicon therefore behaves like an insulator
at very low temperatures.

Electrons and Holes. As temperature is raised, however, thermal energy is imparted to the
lattice structure, causing it to vibrate. When this occurs, electrons at random points through-
out the crystal gain enough energy to break their bond and become free to wander through
the lattice. Wherever such a bond is broken, a hole is left in the lattice. And because a hole
is caused by an electron leaving the orbit of a particular atom, that atom becomes a positive
ion with an electrical charge of $+1$.

At any given temperature above absolute zero, there will be a given number of electrons
that gain enough energy to break the covalent bond. The higher the temperature, the greater
the number of broken bonds; hence the greater the number of electron-pole pairs scattered
throughout the crystal. These electrons and holes can be regarded as charge carriers which
are free to move through the crystal lattice. The electrons represent negative charges; the
holes, positive charges.

Once liberated from its covalent bond, an electron moves at random through the lattice
until it happens to land in the vicinity of a hole in the structure. Now the hole, having a pos-
itive charge, will have an attracting influence on the negative electron, which, when suffi-
ciently close, will jump into the hole. As a result, both the hole and the electron will
disappear as free charges within the crystal.

A hole, too, is free to move through the crystal lattice. When a hole is formed, it will
affect an attracting force on electrons in covalent bonds in the immediate vicinity
because of its positive charge. Since, in a vibrating lattice, these bonds are none too
tight, the energy gained by an electron in an adjacent bond, coupled with the attracting
force of the hole, will cause the electron to break its bond in order to fill the hole. Note
that the hole itself has not disappeared; it has merely moved to an adjacent atom. In this
way, the hole, too, can move at random throughout the crystal. Not until the hole meets
a free electron recombination will take place and cause both hole and electron to disap-
dear as free charges.

For every hole-electron pair that is generated at one point in the lattice, there will be
recombination of a free hole and electron at another point; and for every recombination,
another hole-electron pair will be generated elsewhere. At any given temperature, there-
fore, there will always be a predictable number of carriers, both positive and negative,
within the lattice, and the higher the temperature, the greater the number of free charges in
the material. But it is interesting to compare the number of free charges with the total num-
ber of atoms in the silicon material. Intrinsic silicon has $5 \times 10^{22}$ atoms per cubic centimeter.
At room temperature (25°C), approximately $1.5 \times 10^{10}$ of these atoms are ionized.
Therefore, there is only one hole and one free electron for approximately every 3 trillion
atoms in the crystal structure.

As the temperature is increased, the number of free carriers increases proportionally and
the silicon’s electrical resistivity is decreased proportionally.

5. Impurity concentrations in silicon. The only major electrical characteristic of
intrinsic silicon is a temperature-sensitive resistance which (except, perhaps, as a
temperature-sensing element) is undesirable in any actual use. To utilize the properties of
a semiconductor material, it is necessary to introduce additional charge carriers deliberately
into the intrinsic semiconductor lattice. This is done by introducing impurity atoms into the
structure. If the number of charge carriers introduced by these impurity atoms is much
larger than the number of charge carriers resulting from thermal agitation of the lattice, the
temperature sensitivity of the material is reduced to practical limits.

Two types of impurity atoms are of primary importance: those with three electrons in
their outer rings, such as boron; and those with five electrons in their outer rings, such as phosphorus.

**P-Type Material.** If an atom of boron is substituted for a silicon atom in the lattice struc-
ture, the electrons of the boron atom can form covalent bonds with only three adjacent
silicon atoms. Since the fourth silicon atom of the crystal structure is not symmetrically
bound in the lattice, a defect exists in the lattice structure. This defect takes on the charac-
teristics of a hole. This hole, by virtue of capturing an electron from another adjacent
silicon atom, will move through the lattice in the form of a positive charge.

Impurity atoms having three electrons in the outer ring are called acceptor atoms,
because such impurities will accept electrons from the silicon lattice. Silicon doped with
acceptor atoms is called p-type silicon, indicating that the resulting charge carriers are holes
having a positive charge.

Although the thermally generated holes in heavily doped p-type material have a negli-
gible effect on the characteristics of the material, their number being small in comparison
with the impurity-caused holes at normal temperatures, the thermally generated electrons
do play an important part in semiconductor devices. Because their number is small com-
pared with the number of holes, free electrons in p-type materials are called minority carriers.
The holes, for obvious reasons, are called majority carriers.

Now if an electrical force, e.g., a battery, is applied across a p-type material, the holes
will appear to drift toward the negative terminal of the battery and the free electrons will
drift toward the positive terminal. Because the number of majority carriers predominates,
conduction in doped semiconductor material is majority-carrier conduction or, in this case,
p-type conduction.

**N-Type Material.** If an intrinsic piece of silicon is doped with an impurity atom having
five electrons in its outer ring, only four of the impurity atom’s electrons can form covalent
bonds with the silicon atoms; the fifth, being very loosely bound in the lattice, is free to
detach itself from the impurity atom and wander through the lattice as a negative charge. In
this case, however, the free electron is not balanced by a corresponding hole because the
crystal lattice is unbroken. Under this condition, the material has an excess of free negative-
charge carriers and is called n-type material. Still, the material remains electrically neutral
because the detached electron leaves a positive phosphorus ion in its wake. Impurity atoms
that produce n-type material are called donor atoms.

The n-type material behaves very similarly to the p-type material just described. The
majority carriers in this case are, of course, electrons. The minority carriers are holes gen-
erated by thermal excitation.

6. **P-n junctions and their characteristics.** When two oppositely doped materials
are fused to form a p-n junction, free electrons from the n side and free holes from the
p side of the junction migrate across the junction to combine with the oppositely charged
carriers in the regions near the junction. Since both materials are originally electrically neu-
tral, this migration and subsequent combination create a small electrical potential across the
junction, the p side becoming slightly negative and the n side becoming slightly positive
(Fig. 6.4e). This “space charge” in the vicinity of the junction quickly reaches an equilib-
rium condition in which the acquired negative charge on the p side prohibits any further
migration of electrons from the n region and the acquired positive charge on the n side
repels any further incursion of holes from the p side.
When an external voltage is applied across the junction, the equilibrium condition is altered. With the voltage connected as shown by the battery in Fig. 6.4 (positive to the p side, negative to the n side), some of the excess electrons that originally migrated from n to p are drawn out of the material and the voltage barrier at the junction is reduced. As a result, majority carriers from both sides can move freely and continuously across the junction. Under this condition, the junction is forward-biased and represents little resistance to the flow of current.

Conversely, if the potential of the battery is reversed (negative to the p side, positive to the n side; Fig. 6.4c), electrons from the battery enter the p region and fill in more of the holes in that region. This increases the potential across the junction and prevents the migration of majority carriers across the junction. The junction is then reverse-biased and represents a very high resistance to the flow of current.

But while a reverse-biased junction is a high resistance to majority carriers, it is a very low resistance to minority carriers. These can cross the junction quite easily and form a leakage current. Since the number of minority carriers in doped semiconductor material is quite small, the leakage current is correspondingly small.

Rectification. From the foregoing, it is evident that a semiconductor junction has the properties of a rectifier, permitting a current flow when biased in one direction and prohibiting a
current flow (except for leakage) when biased in the other direction. The voltage-current characteristics of such a junction are shown in Fig. 6.5. Here it is seen that, when biased in the forward direction, the current rises quickly as the applied voltage is raised. The bend in the curve at low voltage is due to the fact that the externally applied voltage must neutralize the space charge before full current is permitted to flow. Thereafter, however, the current flow is directly proportional to the applied voltage, being limited only by the bulk resistivity of the material itself.

Biased in the reverse direction, the current is seen to be extremely small (leakage) until a breakdown point is reached. At that point, current flow increases greatly for any small increase in voltage. Unless current is limited by an external resistance, the junction can be destroyed.

**Capacitance Effect.** A capacitor is formed when two conductors are separated by an insulator (dielectric) across which a difference of potential can exist. For a p-n junction, with no applied external voltage, the region in the immediate vicinity of the junction is depleted of free current-carrying charges. This is so because any mobile electrons near the junction in the n region have migrated across the junction to combine with the mobile holes in the p region. With the junction area thus depleted of free-charge carriers, the region can be considered to be an insulator. This region, however, is bounded at each end by a conductive region of p- and n-type material in which there are numerous free charges (holes and electrons, respectively), so that the structure fits the definition of a capacitor.

The value of the capacitance is a function of the area of the junction (die size), the resistivity of the material (determined primarily by the doping level of the high-resistivity side of the structure), and the voltage applied across the junction. For a reverse-biased junction, this capacitance decreases as the reverse voltage is increased. This is so because an increase in reverse bias increases the width of the depletion area near the junction, thereby widening the dielectric region between the two conductive regions of the structure.

The capacitance associated with a p-n junction represents a limiting factor on the frequency response of semiconductor devices. However, in some devices, such as varactor diodes, it is utilized for constructive purposes.
7. Practical applications for p-n junctions. A simple p-n junction forms the nucleus of a number of practical semiconductor components. Among the most prevalent are signal and rectifier diodes, zener and reference diodes, and varactor diodes.

Rectifier Diodes. The characteristics of a p-n junction are utilized most extensively in the half-wave rectifier diode to convert an ac voltage (or current) into pulsating direct current (Fig. 6.6). This characteristic is used in power supplies, demodulator circuits of AM radios, signal clippers, square-wave generators, and a variety of other applications. Many of these functions, particularly those associated with small-signal applications, are currently being combined with other circuit functions in integrated circuits, so that the use of such diodes in discrete form is diminishing. The primary exception is the rectifier diode for power supplies, which is required to pass relatively high currents and is not yet replaceable with ICs. The selection of such rectifiers, in both capabilities and packaging, continues to expand.

The use of silicon rectifiers in the most popular power supply configurations, as well as the more important circuit characteristics, is shown in Fig. 6.7.

FIGURE 6.6 Schematic diagrams and waveforms of diodes used as (a) rectifiers and (b) clippers.

FIGURE 6.7 Practical power supply circuits and their characteristics.
**Schottky Rectifiers.** When a metal is brought into contact with a highly doped (low-resistivity) semiconductor material, an ohmic contact is formed. The resulting junction is highly conductive, exhibiting linear voltage-current relationships with little series resistance. Such junctions are used for the connection of metal leads to semiconductor devices. However, when a metal is brought into contact with a lightly doped (high-resistivity) semiconductor material, the resulting junction exhibits rectifying properties similar to those of a p-n junction. Alternatively known as Schottky rectifiers (diodes), surface-barrier diodes, and hot-carrier diodes, such structures are commonly employed as VHF-UHF mixers, as which they offer faster response time and lower noise than other types of devices and as high-speed rectifiers in power supply circuits, for which they have a number of advantages over standard p-n junction rectifiers. Among these advantages are:

1. Virtually nonexistent offset voltage at low current and substantially lower forward voltage drop than comparable silicon rectifiers at high currents (Fig. 6.8)
2. Substantial improvement in rectification efficiency
3. Significantly higher switching speed owing to the absence of storage time as a speed-limiting parameter

This combination of features makes Schottky rectifiers highly useful in low-voltage, high-current power supplies, in switching power supplies, and in switching circuits associated with integrated circuits.

![FIGURE 6.8](https://www.digitalengineeringlibrary.com) Typical offset voltage and forward voltage drop of standard and Schottky rectifiers with similar current ratings.
The primary disadvantage of a Schottky rectifier is its substantially higher reverse-bias leakage current, which may be of an order of magnitude greater than that of p-n junctions. Total power dissipation with symmetrical signals, however, favors the Schottky diode because of its low voltage drop in the on condition.

The lower turn-on potential of Schottky diodes is used to increase the switching speed of conventional saturated-logic circuits by keeping the switching transistor out of deep saturation while in the on condition. This circumstance increases the turn-off speed by eliminating storage time as a factor of turn-off time. This is accomplished by connecting a reverse-biased Schottky diode across the collector-base junction of the transistor. As a positive-going base signal drives the transistor toward saturation, resulting in a collector voltage lower than the base voltage (tending to forward-bias the collector-base junction), the lower turn-on potential of the Schottky diode causes the diode to conduct before the normally reverse-biased collector-base junction of the transistor goes into conduction. Thus, the collector-base forward voltage is clamped by the low on voltage of the diode to a value that is less than required to cause reverse conduction in the collector-base junction.

**Zener Diodes.** Operated as a rectifier, a p-n junction utilizes the forward-conduction–reverse-blocking action of the structure. As a zener diode, it employs the “breakdown” region of the diode.

Figure 6.9 shows that, for an applied reverse voltage, the diode current is very small (leakage current) until the breakdown point is reached, at which point the reverse current increases very rapidly. This breakdown voltage is usually called zener voltage ($V_z$). A zener diode differs from a rectifier diode in that the breakdown-voltage point $V_z$ is very precisely controlled during design and manufacture.

When used as a zener diode, the device is operated in the breakdown region, where it acts as a constant-voltage device and can be utilized as a voltage regulator in power supplies. The simple circuit in Fig. 6.10 illustrates this application. Here, when the applied unregulated dc potential is low (below the breakdown point of the diode), the diode current is very small and the device acts essentially as an open circuit. Therefore, its effect on the load circuit is virtually zero. If the applied dc voltage exceeds $V_z$, however, the junction breaks down and its voltage drop remains essentially constant. Therefore, the voltage available to the load remains approximately constant and equal to the zener voltage regardless of variations in source voltage or load-current requirements. Of course, if the applied voltage should drop below the point where diode breakdown is maintained or if the load resistance $R_L$ were to decrease to the point where it could no longer sustain the voltage required for zener breakdown, the diode would come out of conduction and voltage regulation would be lost. With proper design, the value of the applied dc voltage and
the value of the series resistance \( R_s \) are chosen so that the voltage drop across \( R_L \) in the absence of the zener diode would be greater than \( V_z \) under any anticipated load-resistance variations.

Zener diodes are also used for a variety of other applications, including clipper circuits (by replacing the batteries utilized in the circuits of Fig. 6.6), overvoltage-protection circuits (similar to the voltage-regulator circuit of Fig. 6.10), and surge-protection circuits for all sorts of electrical equipment as exemplified by the circuit of Fig. 6.11, where the zener is used to protect against excessive fuse burnout due to prolonged periods of operation near the melting point of the fuse. The zener provides a sudden burst of current at voltages beyond \( V_z \) so that a more tolerant fuse may be selected.

Commercial zener diodes are available with zener voltages ranging from about 2.4 to 200 V and with power ratings from milliwatts to 50 W or more so that a great many power supply regulation requirements can be satisfied. Manufacturers’ data sheets usually provide a minimum current (\( I_{ZK} \)), which will assure reasonable regulation, as well as a maximum current (\( I_{ZM} \)) beyond which the diode is subject to damage.
Reference Diodes. The forward and reverse characteristics of a typical zener diode are affected by temperature. This is shown in the graph of Fig. 6.12, where it is seen that a change in temperature from 25 to 100°C causes a decrease in voltage of approximately 150 mV in the forward-biased region of the diode. This corresponds to a negative temperature coefficient (TC). In the zener region, the voltage increases as temperature is increased, resulting in a positive TC.

For many applications this change in zener voltage with temperature can be tolerated. Critical applications, however, require temperature-compensated diodes, or reference diodes, whose breakdown voltage remains constant under varying temperature conditions. A reference diode can be formed by combining one or more forward-biased diodes with a reverse-biased diode, as shown in Fig. 6.13. Here it is seen that the increase in voltage across the reverse-biased zener junction is counterbalanced by the decrease in voltage across the forward-biased junction at 7.5 mA. By judicious selection of diode junctions and proper operation, excellent temperature compensation can be achieved. It is important, however, to operate the TC device at a current level specified in the data sheet. Otherwise, voltage compensation is not obtained because the $\Delta V$'s at different current levels are not necessarily equal.

The Transistor. The transistor is the most versatile component in the semiconductor family. Its most important characteristic is current and/or voltage amplification. Because of this capability it is the heart of most electronic circuits involving signal amplification and switching.
There are two basic kinds of transistors: unipolar transistors, usually referred to as field-effect transistors (FETs), and bipolar transistors. The FET was first described in the early 1930s but was not exploited commercially at that time. The bipolar transistor was invented in the late 1940s and rapidly became a commercial product. The FET was finally developed as a commercial product early in the 1960s, long after the bipolar transistor had established itself. FETs, especially MOSFETs, are now widely used in the front ends of low-noise radio and TV equipment, and they are the dominant devices in digital integrated circuits.
8. Bipolar transistors. The basic element of any bipolar transistor is the $p$-$n$ junction. When the junction (diode) is forward-biased ($p$ layer positive, $n$ layer negative), current through the device increases rapidly as the voltage is increased. Under reverse bias, only a very small leakage current can flow until the reverse voltage becomes high enough to cause breakdown.

A bipolar transistor is formed by sandwiching a very thin layer of $n$-doped material between two layers of $p$-type material ($p$-$n$-$p$) or a thin layer of $p$-type material between two $n$-doped layers ($n$-$p$-$n$). The characteristics of such a structure are varied by varying the geometries and resistivities of the three layers.

The three layers of a bipolar transistor (Fig. 6.14) are the emitter, the base, and the collector. The emitter represents the current source, where the current carriers originate. The base is the control element, and the collector is the element through which the current carriers are transferred to an external circuit.

![FIGURE 6.14 Physical representation of bipolar transistors.](image)

The main difference between a $p$-$n$-$p$ and an $n$-$p$-$n$ transistor is that the former operates with a negative voltage on the collector while the latter operates with a positive collector voltage (with respect to the emitter). This makes it possible to have complementary circuits that often provide improved performance over single-ended circuits, which can be implemented with only one type of transistor.

Transistor action can be understood by analyzing the current flow through an $n$-$p$-$n$ transistor under the influence of externally applied voltages (Fig. 6.15). For the normal case,
the external voltages are applied so that the emitter junction is forward-biased and the collector junction is reverse-biased. By examining first the collector junction with the emitter-open (a), it is obvious that only minority carriers from the base and collector can cross this junction. If the number of minority carriers in these regions is assumed to be relatively small, the total current across this junction is correspondingly small and is referred to as leakage current ($I_{CBO}$).

If the collector circuit is opened and a forward bias is applied to the emitter-base junction (b), majority carriers from the emitter and the base (electrons and holes, respectively) find it quite easy to cross this junction, causing a heavy base current.

With both junctions properly biased (c), if the base region were quite thick, the electrons from the emitter would combine with abundantly available holes in the base and cause an equivalent and compensating flow of current out of the base lead. In that case, the current in the collector would not be greatly affected by the emitter-injected carriers and would remain at its leakage-current value while the base current would be quite high.

However, for a very thin base region, the lifetime of the electrons injected into the base (before recombination takes place) is long enough so that most of them can traverse the entire base region toward the collector. Since these electrons represent minority carriers in the base region, the reverse-biased collector-base junction is actually forward-biased for the minority carriers, and the electrons will be swept across the junction into the collector region and from there into the collector power source. Therefore, the collector current equals the leakage current plus most of the injected current, while the base current is very small.

The above action gives rise to the most commonly used expressions for transistor current gain, beta (β), which is the ratio $I_C/I_B$ and is called the common-emitter current gain. Beta can be quite high since, for a very thin, high-resistivity base region, almost all of the injected carriers diffuse into the collector, with very few exiting via the base lead.

**Circuit configurations.** There are three basic transistor circuit configurations: common-emitter, common-base, and common-collector. They differ principally in the manner in which the signal is applied to the transistor and where the load is attached. Figure 6.16 shows these basic circuits. Since the common-emitter circuit is by far the most prevalent, data sheets normally characterize the transistor in terms of this circuit.

![Figure 6.16](image)

**FIGURE 6.16** Basic transistor circuit configurations and their relative characteristics.

<table>
<thead>
<tr>
<th>Circuit</th>
<th>Current gain</th>
<th>Voltage gain</th>
<th>Input resistance</th>
<th>Output resistance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common emitter</td>
<td>High</td>
<td>High</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Common base</td>
<td>&lt;1</td>
<td>High</td>
<td>Very low</td>
<td>Very high</td>
</tr>
<tr>
<td>Common collector</td>
<td>High</td>
<td>&lt;1</td>
<td>High</td>
<td>Low</td>
</tr>
</tbody>
</table>
**Bipolar-Transistor Characteristics.** Figure 6.17 shows the characteristics of a typical small-signal transistor in a common-emitter circuit. The input circuit, curve a, shows that a base-emitter voltage $V_{BE}$ of less than approximately 0.5 V (for silicon) causes virtually no emitter current $I_E$ to flow, thereby keeping the transistor cut off. Above 0.5 V, $I_E$ rises sharply, limited only by the ohmic resistance of the emitter region. Since the latter is very small, a very small rise in $V_{BE}$, beyond the threshold voltage $V_T$, causes a large injection of emitter current into the base and from there into the collector, with a small portion flowing out of the base lead.

A typical plot of the division of emitter current between base and collector is shown in Fig. 6.17b. The numbers vary considerably for different transistors, but the shape of the curve remains similar. This shows that the $\beta$ of the transistor can vary considerably over the range of the curve and that, for large signals, distortion can be quite high.

The collector curves in Fig. 6.17c are the most significant. They can be used to determine the relationship between the input and output signals of a transistor.

When the transistor is cut off ($I_B = 0$), a residual current $I_{CES}$ flows in the collector circuit. This leakage current can be reduced somewhat (but not to zero) by applying reverse bias to the emitter junction. The limiting value, with reverse bias applied to the emitter-base junction, is $I_{CBO}$. This is equivalent to the collector-base leakage current when the emitter is open-circuited.

Collector current increases rapidly as the base is energized. The maximum $I_C$, $I_{C(max)}$, is that which would damage the internal transistor structure. This value of collector current is given as a maximum rating on data sheets. Thus, the output current could range from $I_{C(max)}$ to $I_{CES}$ (or $I_{CBO}$ in the event of reverse bias), but in practice $I_C$ is limited to a value far less than $I_{C(max)}$.

**Power Dissipation.** Power dissipation $P_D$ is the product of $V_{CE}$ and $I_C$, and it causes the collector junction to heat up. Beyond a critical junction temperature $T_{J(max)}$, the device could be damaged. Thus, the $P_{D(max)}$ rating of a transistor limits the maximum $V_{CE}$ and $I_C$ that can be applied simultaneously. The locus of a $P_{D(max)}$ rating is the parabolic curve on the $V_{CE}/I_C$ plane.
Switching Action. From the above it is evident that when the voltage applied between base and emitter is less than approximately 0.5 V, there is only a very small current flow through the transistor. Between collector and emitter, therefore, the transistor represents a very high resistance, or open switch. When $V_{BE}$ is raised significantly above the threshold voltage, the internal resistance of the transistor drops to a very small value so that the device approximates a closed switch. As shown in Fig. 6.17, it takes a change of only 7 mA of $I_b$ along the load line $XQY$ to cause $V_{CE}$ to cause the voltage across this particular transistor to change from $V_{CC}$ to near zero, that is, from an open switch to a closed switch. The switch action is accomplished by a $V_{BE}$ of only a few tenths of a volt.

Biasing. When operated as an amplifier, the transistor must first be biased to some quiescent value of collector current, so that both positive- and negative-going input-voltage excursions will cause corresponding changes in output voltage and current. The ideal bias point is represented by $Q$ on the load line because this permits approximately equal excursions in $I_c$ and $V_{CE}$ without signal clipping. The bias point is established by a quiescent base current that results in a dc collector voltage of approximately $V_{CC}/2$.

Several circuits are used for establishing the bias point. Among the most familiar are those in Fig. 6.18. The basic performance difference is in the bias-point stability.

\[
\begin{align*}
I_c & = \frac{V_{CC}}{R_B} \\
\beta & = \frac{V_{CC}}{R_c + R_B/\beta} \\
R_2 & = \frac{V_{CC}}{R_c + R_2/\beta} \\
S & = \frac{1}{1 + \beta R_c/R_B} \\
& \text{for } R_B = \beta R_c, \quad S = 0.5 \\
& \text{or } R_B = \frac{R_c R_2}{R_1 + R_2}, \quad S = \frac{1}{1 + \frac{\beta R_c}{R_1 + R_2}} \\
I_c & = \frac{I_{C(eq)}}{2} \\
R_B & = 2\beta R_c \\
R_b & = \beta R_c \\
R_2 & = 1 + \frac{R_1}{R_2} \\
\end{align*}
\]

To make

At point $Q$ on the load line in Fig. 6.17, the transistor has a beta ($I_c/I_b$) of approximately 20. If a transistor with a $\beta$ of 40 were substituted (simulated by dividing all $I_b$ values by 2) and if $I_b$ were held by the bias circuit to 2.5 mA as before, the operating point would move up the load line to point $Q'$, a much higher value of $I_c$. As a result, considerable distortion would occur for high-value input signals.

The bias-point stability factor $S$ is defined as the percent change in $I_c$ for a percent change in $\beta$, or $\Delta I_c/I_c = S \Delta \beta/\beta$. If a percent change of $\beta$ causes a corresponding percent change in $I_c$ (the least desirable condition), then $S = 1$. If $I_c$ is independent of $\beta$ (corresponding to a zero change in $I_b$ when $\beta$ is varied), then $S = 0$. The

\[\text{FIGURE 6.18} \quad \text{Conventional common-emitter bias circuits. The following table gives approximate characteristic expressions.}\]
formulas accompanying Fig. 6.18 give $I_c$ and $S$ as functions of $\beta$ and assign values for $S$ under specific operating conditions. The bias arrangement in Fig. 6.18c, using emitter degeneration, is preferred because, by proper choice of resistor values, the effect of $\beta$ on $I_c$ can be made almost negligible.

9. Field-effect transistors (FETs). There are two types of field-effect transistors: the junction field-effect transistor (JFET) and the metal-oxide–semiconductor field-effect transistor (MOSFET). Of these, the MOSFET is by far the more popular and forms the basis for most digital large-scale and very-large-scale integrated circuits (LSI and VLSI). JFETs, because of their low-noise advantages, are utilized extensively in TV and radio front ends as well as in input circuitry for integrated operational amplifiers, etc.

**Junction Field-Effect Transistors.** In its simplest form, a JFET starts with a bar of $n$- or $p$-doped silicon which behaves as a resistor (Fig. 6.19a). By convention, the terminal into which current is injected is called the source; the other terminal is the drain. Current flow between source and drain is dependent on the drain-source voltage and on the resistivity of the material.

In Fig. 6.19b, $p$-type regions have been diffused into an $n$-type substrate, leaving an $n$ channel between source and drain. (A complementary $p$ channel is made by reversing the dopants in all regions of the structure.) The $p$ regions control the current flow between source and drain and are called gates.

When the gates are connected to the source and a positive voltage is applied between drain and source (Fig. 6.19c), current flow in the channel creates a voltage gradient along the length of the channel. This voltage reverse-biases the gate-substrate junctions and creates the usual depletion layer in the vicinity of the junction. The depletion layer, being devoid of mobile charges, is in effect an insulating region which cannot be penetrated by the channel current. This reduces the channel width, thereby increasing its resistance.

The overall effect is shown in Fig. 6.20a, where an increase in $V_{DS}$ causes a corresponding increase in drain current $I_D$, but the rate of change of current tapers off as the increasing voltage drop in the channel increases the spread of the depletion region into the channel. At $V_P$ (pinch-off voltage) an equilibrium condition is reached whereby an increase in $V_{DS}$ generates a neutralizing increase in channel resistance, and drain current remains essentially constant until the breakdown condition is reached.

By applying a reverse-bias voltage between the gates and the substrate (Fig. 6.20b), the channel width is reduced by the gate voltage and the maximum drain current is lowered. Further increases in gate voltage correspondingly reduce the channel width and produce successive reductions in maximum drain current.

It will be noted that current flow in an FET is made of majority carriers only and that, unlike in a bipolar transistor, the current does not cross a junction. Hence, an FET is sometimes called a **unipolar transistor**.

![FIGURE 6.19 Development of junction field-effect transistors.](image-url)
A more practical structure for a junction FET is illustrated in Fig. 6.21. In this single-ended geometry, the p-type substrate takes the place of gate 2 in Fig. 6.20. An n-type layer is epitaxially grown on top of this substrate, and a p-type gate diffusion into the epitaxial layer forms the second gate and creates the channel.

**MOS Field-Effect Transistors.** The MOSFET operates on a somewhat different control mechanism. Figure 6.22 shows the development. Two separate low-resistivity n-type regions (source and drain) are diffused into a high-resistivity substrate, as shown in Fig. 6.22a. Next, the surface of the structure is covered with an insulating oxide layer (Fig. 6.22b), and holes are cut into the layer, allowing metallic contact to source and drain. Then, a metal area is overlaid on the oxide, covering the entire region between source and drain, and simultaneously metal contacts to drain and source are made as shown in Fig. 6.22c. The metal area between source and drain is the gate terminal. Note that there is no physical penetration of the gate metal through the oxide into the substrate. Since drain and source are isolated by the opposite-polarity substrate, the structure is analogous to two diodes connected back to back (Fig. 6.22d). If a voltage were to be applied between source and drain (in the absence of a gate voltage), the resulting current flow would be very small because one of the back-to-back diodes would be reverse biased regardless of the polarity of the applied voltage.

The metal gate area, in conjunction with the insulating oxide layer and the semiconductor layer underneath, forms a capacitor. The metal area is the top plate; the substrate material, the bottom plate. When a positive potential is applied to the gate metal, the positive charge at the metal side of the capacitor induces a corresponding negative charge at the semiconductor side. As the positive charge at the gate is increased, the negative charge “induced” in the semiconductor increases until the region beneath the oxide is changed from p-type to n-type (Fig. 6.23), and current can flow between drain and source through the induced channel. In other words, drain-current flow is “enhanced” by the gate potential. Thus, drain-current flow can be modulated by the gate voltage; i.e., channel resistance is inversely proportional to gate voltage. (The n-channel structure may be changed to a p-channel device by reversing conductivity of the semiconductor regions.)
The FET just described is called an enhancement-type MOSFET. A depletion-type MOSFET can be made by diffusing a low-resistivity $n$ channel into the space between source and drain so that considerable drain current will flow when the gate potential is at zero volts. In this manner, the MOSFET can be made to exhibit depletion characteristics by applying a negative voltage to the gate.

For a moderate-resistivity induced channel, the FET will exhibit both enhancement and depletion characteristics. Typical performance curves for such structures are illustrated in Fig. 6.24.

Comparison of Small-Signal Bipolar and MOS Transistors. The MOSFET enjoys several clear-cut advantages over its bipolar counterpart. First, it has an extremely high input resistance, making it more adaptable to many circuit applications. Second, it is a low-noise
device and is often preferred for front-end communications circuits. Third, it is extremely compatible with integrated-circuit processing, particularly for LSI and VLSI circuits, for which its smaller geometry, lower power consumption, and simpler processing make it the preferred basic device for many applications.

On the other hand, bipolar transistors are capable of operating at higher speeds and higher power. In discrete form, they are generally available with a wider range of specifications and often at considerably lower cost for applications for which both types of devices are suitable.

**Power MOSFETs.** While MOSFETs for small-signal applications have been widely used for some years, their performance in power circuits has been limited. The limiting factor of the conventional small-signal MOSFET structure has been the lateral channel, which has an inherently low current-carrying capacity and a relatively high on resistance coupled with a low reverse breakdown voltage.

The disadvantages of the lateral FET from a power standpoint were overcome with the development of a structure that permits the control of a vertical current flow by means of a gate field. The structure utilizes thousands of “source” sites interconnected in parallel on a single die (Fig. 6.25). A common drain at the bottom of the die minimizes the geometry of the structure to make such a multiecell architecture practical. Moreover, it permits very short channel lengths to reduce the on resistance of each cell to a minimum. The overall result is a device which allows current and voltage ratings to meet virtually any desired electronic application while maintaining the higher speed and circuit-simplification advantages of MOSFETs in comparison with bipolar transistors.

The power MOSFET has made great strides, but emphasis on cost reduction is continuing. Improvements of present devices over the early double-diffused MOS structure have been significant, and prices are now approaching those associated with bipolar power transistors for medium voltage and current applications.

10. **Thyristors.** Thyristors and their trigger devices can take numerous forms, but they share these characteristics:

1. They are “open circuits,” capable of withstanding rated voltage until triggered.
2. They become low-impedance current paths when triggered and remain so, even after the trigger source has been removed, until current through that path is interrupted or is reduced below a minimum “holding” level.

The regenerative action which holds a thyristor in the on state is due to multiple layers of opposite p- and n-doped silicon which result in the two-transistor-equivalent circuit
structure shown in Fig. 6.26. Here, part of the current through transistor 2 is injected back into transistor 1 to supplement the trigger current and sustain conduction when the trigger is removed. This characteristic, coupled with the thyristor’s low on resistance, makes it possible to control a portion of each cycle of an ac power waveform into a load, for low-dissipation “dimming” or motor-speed–control applications, to switch capacitive discharge currents precisely in electronic pilot ignition systems, and to fulfill innumerable other control purposes.

There are a number of semiconductor devices under the title of thyristor; the most important of these are silicon-controlled rectifiers (SCRs) and triacs.

**SCRs.** As a rectifier, the SCR conducts current in only one direction. But unlike a two-layer rectifier, which begins conduction when its anode becomes slightly positive with respect to its cathode, the SCR will remain nonconductive even in the forward direction until the anode voltage exceeds a certain minimum value called the *forward breakover voltage* ($V_{BRM}$). Moreover, the value of $V_{BRM}$ can be varied through the injection of a small current into a third, or gate, element, which governs the amplitude of the anode voltage needed to cause conduction, or firing.

The current-voltage relationship of an SCR is shown in Fig. 6.27. With the gate terminal open or shorted to the cathode and an external voltage applied only between anode and cathode, the reverse-bias characteristics (anode negative with respect to cathode) are identical to those of a conventional $p$-$n$ junction, or rectifier. Under forward-bias conditions, however, current flow does not begin (except for a small leakage current) until $V_B$ is reached. At that point current flow begins to increase very rapidly until, at the *forward-breakover-current point*, a switchback effect takes place and the voltage across the device suddenly drops to a very low value. At that point, the voltage-current relationship becomes similar to that of a conventional forward-biased diode.

It is important to recognize the difference between the reverse-bias breakdown and the forward-bias-breakover effect. In the former, the voltage across the junction remains at its breakdown value so that power dissipation in the device is high and care must be taken to keep the maximum current at a relatively low value to prevent device damage. In the forward-breakover region, owing to the switchback effect, the voltage across the device is very
low, and it can carry large amounts of current without exceeding the power-dissipation rating of the device. Hence, the device with its gate open (or shorted to the cathode) acts like a voltage-operated switch. The switch is turned on when the applied anode voltage exceeds the breakover voltage and is turned off when the anode current through the device is reduced to the holding current $I_H$, at which point the device returns to its high-resistance state and the switch is turned off.

The switching action of the SCR is greatly enhanced by employing its gate characteristics. When a current is injected into the gate, the breakover voltage of the device is reduced. As this current is increased, the breakover point becomes less and less until the barrier to forward conduction is removed entirely and the device acts as a forward-biased rectifier. The amount of gate current required to reduce breakover voltage from its open-gate value (open switch) to virtually zero (closed switch) is extremely small, so that a small amount of gate current can control a large amount of anode current. In this respect, the SCR acts very much like a sensitive relay, in which a small current through the relay coil can control a much larger current through the relay contact circuit.

A simplified SCR circuit is shown in Fig. 6.28, where the SCR is connected in series with the load $R_L$ and an applied voltage $V_{ac}$. $V_{ac}$ is less than the breakover voltage of the SCR, so that the SCR is an open circuit until a gate signal is applied. If a gate-signal pulse is applied at the beginning of the positive-going cycle of $V_{ac}$, the SCR is turned on and the current through the load follows the positive half cycle of the applied voltage until that voltage approaches the point $A$ where it can no longer sustain a current above the holding-current value. The SCR is then turned off.

It must be noted that the SCR can be turned on only during the positive-going portion of the applied voltage. No amount of gate current can turn it on during the negative portion of the cycle, so that a single SCR acts like a triggered half-wave rectifier.

Various techniques can be used to permit SCRs to function during both halves of an applied ac voltage. This can be accomplished by connecting two SCRs in an inverse-parallel configuration so that one device is turned on during the positive-going half cycle and the other during the negative portion (Fig. 6.29a). Another is to utilize a triac in place of an SCR (Fig. 6.29b).

**Triacs.** A triac is a device which operates exactly like an SCR, except that it can be triggered on during both the positive-going and the negative-going half cycles of an applied voltage. Thus, it acts like a triggered full-wave rectifier or, as indicated by its schematic symbol, like a pair of inverse-parallel-connected SCRs. In application, the triac offers circuit...
simplification and a reduction in component requirements where full-wave power control is required.

A triac may be triggered into conduction by either a positive- or a negative-going gate signal, but sensitivity varies considerably. The recommendations given in manufacturers' data sheets should be followed for best performance.

**Gate-Turnoff Thyristors (GTOs).** A gate-turnoff thyristor (GTO) combines the most desirable characteristics of SCRs and bipolar power transistors. Like a transistor, the GTO can be switched on and off by a low-power gate drive signal; like an SCR, it can pass high forward currents when turned on and block high forward voltages when turned off. But unlike the power transistor, which requires a constant drive source to keep it turned on, the GTO is turned on by a momentary pulse and remains in the on condition until the anode voltage is reduced below the level that sustains a holding current or until it is turned off by a reverse-polarity gate-turnoff pulse. In this way, it differs, too, from more conventional
thyristors (SCRs, triacs), which cannot be turned off by a gate signal. Present devices have voltage ratings up to 1400 V and current ratings on the order of 20 A. Gate current trigger requirements are 300 mA. Devices will withstand surge currents as high as 200 A.

**Triggers for Thyristors.** The most common form of power control using SCRs (or triacs) is phase control of alternating current. Power control is achieved by controlling the timing of the gate trigger pulse. If the trigger pulse is applied at the beginning of the positive-going anode cycle, the thyristor is turned on for virtually the entire positive half of the cycle, allowing the maximum load current to flow. As the trigger pulse is delayed, correspondingly less of the positive-going anode voltage waveform is applied to the load. Figure 6.28 explains visually how this phase control of alternating current is achieved.

There are some instances when a thyristor can be adequately triggered by the slowly rising voltage or current from the 60-Hz line without a separate trigger device, e.g., the simple motor-speed-control circuit in Fig. 6.30. Here the setting of $R_3$ determines the point of the applied-voltage cycle at which the gate voltage will be high enough to trigger the SCR. The circuit offers control over almost the entire half of the positive portion of the applied voltage. Much better control can be achieved, however, by employing a separate trigger device to provide reliable turn-on pulses timed to trigger the SCR at any desired point during the cycle.

The most common trigger circuit is a relaxation oscillator (Fig. 6.31) using one of a variety of semiconductor switching devices. As shown, when the voltage on the capacitor rises to a point where the electronic switch turns on, the discharge of the capacitor produces a trigger pulse to the load. The charge time of the capacitor can be varied by the setting of $R_T$, so that pulse spacing can be varied at will.

A number of trigger devices serve the purpose of the electronic switch. Among them are unijunction transistors (UJTs), three- and four-layer diodes, and sidacs.

**Unijunction transistors.** The UJT is a three-terminal device having only a single semiconductor junction. Its basic structure and equivalent circuit are shown in Fig. 6.32a and b.

In the physical structure, when a positive voltage is applied between base 2 ($B_2$) and base 1 ($B_1$) and the emitter is shorted to $B_1$, the current through the device sets up an internal voltage drop. The voltage at the p-n junction, being slightly positive, reverse-biases the diode junction, and there is a small reverse-bias (leakage) current through the diode. The diode represents a high resistance in parallel with $R_{em}$ and has no effect on the internal voltage distribution.
If a positive-going voltage is applied between the emitter and $B_1$, a point will be reached where the voltage overcomes the internal reverse-bias voltage on the p-n junction and the junction becomes forward-biased, representing a very small resistance in parallel with $R_{b1}$. At that point, the voltage at point $A$ drops to that of a forward-biased junction.

The emitter $I-V$ characteristic curve of a UJT is shown in Fig. 6.33. To an external circuit, it appears that the UJT has a negative resistance to applied emitter voltage between $V_p$, the diode turn-on point (peak-point voltage), and $V_V$, the point where the diode becomes fully conductive and its voltage drops to its forward-biased value (valley point).

A relaxation oscillator utilizing a UJT (Fig. 6.34) provides an output across $R_L$ after capacitor $C_t$ has charged up through $R_t$ to reach the peak-point voltage of the UJT. Subsequently, $C_t$ discharges through $R_L$ and the UJT to generate a positive pulse across $R_L$. The frequency of the pulsed output can be varied by varying the resistance of $R_t$, which permits triggering the thyristor load at any desired point of the positive-going anode cycle.

Unijunction transistors are highly stable devices for general-purpose trigger applications and as pulse generators and timing circuits at frequencies ranging from 1 Hz to 1 MHz.

A number of other trigger devices are available. All take advantage of the phenomenon of negative-resistance characteristics to produce switching action. The $V-I$ characteristic curves of some of the more popular devices are illustrated in Fig. 6.35.
FIGURE 6.35 Voltage-current characteristics of various popular thyristor devices.
Programmable unijunction transistors (PUT). Programmable unijunction transistors are similar to UJTs except that their basic characteristics are programmable (adjustable) by means of external voltage dividers. This ability stabilizes circuit performance for variations in device parameters. The general operating frequency range is from 0.01 Hz to 10 kHz, making PUTs suitable for long-duration timer circuits.

Bilateral triggers: DIACs. Specifically designed as low-cost triggers in line-operated triac control circuits such as light dimmers, motor controls, and temperature controls, these devices exhibit symmetrical bidirectional characteristics with defined switching voltage and current points. The negative-resistance region extends over the full range of operating current, so that the concept of holding current is not applicable. Switching voltages on the order of 35 V or greater are achievable.

Silicon bidirectional switches (SBS). These devices are actually integrated circuits as opposed to the four layer discrete structures of other thyristors. They offer greater accuracy than conventional triggers, the difference between positive and negative breakover voltage being less than 0.5 V. A third lead, designated the gate, permits variations in characteristics such as breakover voltage. Switching voltages up to 10 V are achievable.

Sidacs. These components extend trigger capabilities to significantly higher voltages and currents than are achievable with other devices. With voltage ratings up to 300 V and a steady-state rms current rating of 1 A, they can replace triacs in numerous low-power applications.

11. Optoelectronics devices. The field of optoelectronics deals mainly with the phenomena of converting light into electric current and, conversely, turning electric current into light. Phototubes, followed by solid-state cadmium sulfide photocells, have been early and successful examples of light-to-current conversion, while tungsten-filament and gas-filled bulbs have been serving current-to-light conversion for many years. Their applications, however, have been restricted to direct current and low frequencies. Semiconductor light sources (LEDs and lasers) and detectors have brought about significant changes in optoelectronic applications.

Light sources basically consist of injection laser diodes (ILDs) and light-emitting diodes (LEDs). Both are semiconductor chips made of gallium arsenide (GaAs) or related III-V compounds.

Detectors consist primarily of silicon p-n junctions that perform the conversion function, but these may be combined (on the same chip) with other semiconductor devices such as transistors or integrated circuits to improve sensitivity and resulting output.

Light Sources. One basic characteristic of semiconductor p-n junctions is their sensitivity to light. Specifically, in the cutoff region the leakage current of such a diode increases with light intensity, while in the avalanche breakdown region it actually emits light. A diode therefore can be operated both as a light detector and as a light source.

The important characteristics of light sources are spectral width, the portion of the spectrum covered by the emitted light; peak wavelength, the wavelength at which maximum output is achieved; emission pattern, the "spread" of the light beam as it leaves the emission point; power output, the amount of light emitted by the device at its peak wavelength; speed, usually determined by the time it takes for the output to rise from 10 to 90 percent of peak power in response to an applied signal pulse; and reliability and lifetime, the anticipated mean time between failures of a component.

In electronic applications, the ILD is usually considered for long-distance communications in the fiber-optic transmission of CATV and telephone signals. It has a very narrow optical beam (Fig. 6.36) and can launch a greater amount of power into small-diameter fiber cores than an LED. And, because of its significantly higher speed, it has greater frequency-modulation
rates. The LED is normally preferred for short-distance communications and for other electronic functions such as optocoupling (isolation) and light-control applications. This preference is due to its substantially lower cost, its spectral match with a wide variety of detectors, its comparatively long life, and its general compatibility with most short-distance communications requirements.

The spectral response of an LED depends on the material used in its manufacture. Gallium arsenide is used primarily for LEDs with outputs in the infrared region of the spectrum between 900 and 1000 nm. Gallium arsenide phosphide produces a visible red light at about 660 nm, and gallium phosphide yields green at about 560 nm.

Visible-light diodes are widely used in panel lights, circuit-condition indicators, light modulators, displays, and the like. Infrared emitters are employed in fiber-optics communications, in card and tape readers, for shaft and position encoders, and in other applications requiring the use of photodetectors in place of the human eye because, as shown in Fig. 6.37, the detector response is much greater to infrared than to the visible-light spectrum.
Packaging has a substantial effect on the characteristics of an LED. It affects not only the power output but the spatial response as well. Figure 6.38 shows variation of output power and light intensity to be linear functions of current for typical LEDs. One limiting value is the current limit set by the package in which the die is housed. Spatial response is a function of the placement of the die within the package (with respect to the lens) and of the shape of the lens.

Spatial distribution is also a function of the basic chip design. A simple p-n junction yields a lambertian emission pattern which radiates in all directions. While this is desirable for some applications, it is undesirable in others. A number of techniques have been developed to control the emission pattern of the junction. Some of the prevalent chip structures and packaging methods are illustrated in Fig. 6.39.

**FIGURE 6.38** Typical characteristics of commonly available LEDs. Graph, upper left, shows typical light output as a function of drive current. Spatial-response curves are functions of package and lens designs.
FIGURE 6.39 Chip design (a) and packaging methods (b) affect the radiation pattern of LEDs.
**Photodetectors.** All semiconductor materials are light-sensitive. An increase of light within a specific range of frequencies, like an increase in temperature, imparts energy to the crystal lattice, causing some of the electrons to break their covalent bonds and become free. In a homogeneous semiconductor material or in a forward-biased p-n junction, the number of light-generated carriers is relatively small compared with the normally available free carriers and add little to the overall current flow. In a reverse-biased junction, however, the increase in current due to light-induced carriers becomes significant.

**Photodiodes.** If a near-intrinsic layer of silicon is interspersed between oppositely doped layers, a PIN diode is formed. The I (intrinsic) layer, being nearly devoid of free carriers, adds significantly to the width of the depletion region, making the structure far more light-sensitive than the ordinary p-n junction.

If the reverse bias on a p-n junction is increased to a point near the reverse breakdown point, the electric field within the depletion region will be high enough to give any light-liberated carrier enough energy to liberate additional carriers through collision with other atoms in the lattice. This avalanche effect actually multiplies the current generated by the impinging light and greatly increases the sensitivity of the diode.

**Phototransistors.** When one of the junctions of a transistor, e.g., the normally reverse-biased collector-base junction, is exposed to light, a phototransistor is created. Such a device adds gain to a photodiode and therefore increases sensitivity considerably. It pays for this, however, with a decrease in response time. Whereas the response time of diodes is measured in nanoseconds, that of phototransistors is specified in microseconds. Even greater gain and correspondingly slower response time are achieved with photodarlington devices.

Virtually any semiconductor device can be made photosensitive. Thus, manufacturers are now providing phototriacs and photo SCRs in order to reduce the component count in light-sensitive equipment requiring such devices.

With the advent of integrated circuits, optodetectors are being combined with complete preamplifiers to provide an amplified low-impedance output that is far less noise-sensitive than that of a discrete detecting device.

**Optocouples.** An excellent example of optoutilization to replace conventional electronic components is the optical coupler. Consisting of an infrared-emitting diode coupled to a phototransistor in a single package, the device advantageously replaces such components as interstage transformers and relays as well as coupling and feedback networks.

The diagrams in Fig. 6.40 show such a unit being used as a linear-signal coupler and as a pulse coupler. In the former mode a constant current supplied to the emitter biases this diode, and since the output (infrared) is directly proportional to the diode current, any increase or decrease of diode bias current resulting from an applied modulation input causes corresponding variations in light output. These are coupled to the detector, which provides an equivalent linear current output. Devices with current transfer ratios ($I_e/I_c$) ranging from 2 to 1000 percent (depending on detector gain) are currently available.

As pulse couplers, the devices are equally interesting. The principal operating difference is that in the switching mode they require no bias current. The detector is either off (with a dark current of typically less than 20 nA) or on (with a maximum continuous forward current of around 50 mA).

As an electronic relay, the coupler is fast (much faster than a mechanical relay), and it has no contacts to bounce, pit, or corrode. It is small and insensitive to vibration, and, in contrast to other forms of electronic relays, its output is completely isolated from its input.

Couplers of this type could be produced with an almost infinite variety of gain, sensitivity, and output current. Currently they are available with transistor output, Darlington
output, SCR output, triac driver output, and Schmitt trigger output. They are even being combined with on-chip amplifiers. This is a case of matching design to a required application, and the couplers clearly lend themselves ideally to custom fabrication when end-use quantities are high enough to permit economical production.

12. Varactor diodes. Normally, a p-n junction (diode) is operated in the forward conduction region, as a rectifier, or in the reverse breakdown (avalanche) region, as a zener. There is, however, a third region, the region between forward conduction and reverse breakdown (Fig. 6.41), in which the junction is nonconductive and acts as a capacitor. This junction capacitance is an undesirable parasitic in most applications but becomes the mechanism whereby the junction can act as a variable-capacitance (tuning) diode or frequency multiplier.
An unbiased semiconductor junction simulates a slightly charged capacitor, with the depletion region representing the dielectric and the n- and p-type regions adjacent to the junction representing the two conductive plates. If an external voltage is connected across the diode so as to reinforce the contact potential (reverse bias), the depletion-layer width increases, resulting in a decrease in capacitance. Conversely, if an external forward bias is applied, the depletion region narrows and capacitance increases. However, if the forward voltage is made large enough to overcome the contact potential, forward conduction occurs and the capacitance effect is destroyed. By applying a variable dc voltage of a magnitude between points A and B of Fig. 6.41, the capacitance of the junction can be varied to cover the tuning requirements of radio and television receivers.

Three parameters are of particular importance with varactor diodes: nominal capacitance, capacitance ratio, and $Q$. The nominal capacitance $C_n$ is normally specified on data sheets at a particular value of reverse-bias voltage. This serves as a reference point for comparing the capacitance values of various varactor diodes. Capacitance ratio is the ratio $C_{\text{max}}/C_{\text{min}}$ for a specified reverse-bias voltage range. $Q$ is the figure of merit that defines the quality of the device as a capacitor. As a reference, mechanical tuning capacitors that are being largely replaced by varactor tuning diodes often have $Q$'s on the order of 1000 or greater. Tuning-diode $Q$'s are generally considerably lower but still are adequate to achieve the desired performance.

The capacitance change over a specific voltage range for a commercial line of tuning diodes is shown in Fig. 6.42.

Varactors can be applied to most sections of a receiver or transmitter where variable capacitors are required. Usable frequencies range to several thousand megahertz, and nominal capacitance values range from about 6 to 500 pF or greater. Capacitance ratios run from approximately 15, for AM broadcast-band tuning diodes, to as low as 2.5 for diodes used for general tuning and control purposes.

Varactor diodes are useful for a variety of radio-frequency applications, including electronic tuning, harmonic generation, and parametric amplification.

13. **Transducers (sensors).** Transducers are representative of a class of device that converts one form of energy into another. Sensors are often defined as the basic element within a transducer in which the actual conversion takes place. This may be followed by...
additional circuitry that provides signal conditioning and amplification to accomplish an end result.

Semiconductor devices, owing to their low cost and high reliability, are being adopted in increasing numbers to accomplish the basic sensing functions. Optoelectronic devices in the form of light emitters and detectors (Sec. 11) are examples of semiconductor applications for sensing purposes. Other categories include temperature sensors and pressure sensors.

**Pressure Sensors.** One form of pressure sensor employs the piezoresistive effect in silicon to convert a change in pressure into a change in electric current. A commercial structure of this type is illustrated in Fig. 6.43. Here a cavity, or chamber, is etched into one side of a bar of silicon, leaving a thin, flexible bridge of silicon that acts as a diaphragm. A resistive element is diffused into this diaphragm at a point where it is most susceptible to stress when the diaphragm is flexed. A pair of contacts is placed at a point of the resistor to measure the transverse voltage at that point when a current flows through the resistor.

![Pressure Sensor Diagram](image-url)
The resistance value is subject to change as the resistor is stressed as a result of pressure being applied to the diaphragm (piezoresistive effect). This causes a change in voltage at the points of contact that is available for measurement or control purposes by means of the package pins.

The device can measure absolute pressure with respect to a vacuum or differential pressure with respect either to atmospheric pressure or to two different pressures applied to opposite sides of the diaphragm. For absolute-pressure measurements, the chamber is sealed off in an evacuated atmosphere to create a vacuum on the underside of the diaphragm. This flexes the diaphragm and generates a reference offset voltage equal to 14.5 psi (1 atm). When an external pressure is applied to the opposite side of the diaphragm, the output voltage varies linearly above or below the reference point, depending on the polarity of the external pressure, whether positive or negative (Fig. 6.44).

For differential-pressure measurements, the chamber is left unsealed so that an external positive (downward) pressure applied to the “pressure” side of the diaphragm acts against existing atmospheric pressure at the underside to create a differential voltage. Alternatively, the difference between two external pressures can be measured by applying these simultaneously to opposite sides of the diaphragm, through pressure ports associated with the package (Fig. 6.45).

While piezoresistive techniques in silicon provide a very accurate and linear output, they are temperature-sensitive and generate an initial offset voltage. Both effects can readily be compensated for with external circuitry, but integrated-circuit technology has been
employed to put temperature compensating and calibrating directly on the chip during the fabrication process. The compensating thin-film resistor network is laser-trimmed during the computer-controlled manufacturing process to provide good temperature stability and to eliminate the offset (Fig. 6.46). The more advanced devices feature on-chip amplification to increase the output-voltage span from millivolts to volts.

**INTEGRATED CIRCUITS**

The evolution of integrated circuits has initiated tremendous innovations in the conception, design, and manufacture of electronic equipment. The individual components such as resistors, capacitors, and inductors, which once were the building blocks of all electronic equipment, are being relegated to relatively minor peripheral functions. Circuit designs are being implemented without the benefit of capacitors and inductors, which are difficult to combine with active components on a practically sized silicon chip. Even the resistor is losing its battle against active devices, such as transistors, which can often serve the resistive function more compatibly and in much less space. Digital techniques are replacing analog functions in communications circuitry. The computer is replacing the time-honored “breadboard” for developing, testing, and evaluating new circuit and system designs; and in the factory the microscope is replacing the soldering iron as the basic manufacturing tool in many instances.

But there still remains a large number of applications for which ICs are not yet adaptable. Integrated circuits are not able to produce high power output; they are uncomfortable in VHF-UHF applications, and they are not particularly well suited to very simple electronic equipment that might have special requirements outside the realm of standard off-the-shelf ICs. Indeed, while ICs are expanding into ever larger subsystem functions, they will continue to need discrete-component supplements for some time to come.

---

**FIGURE 6.46** Schematic diagram of a fully compensated monolithic pressure sensor (a) and output characteristics of compensated and uncompensated devices (b).
All integrated circuits have a common goal of reducing equipment size and cost and improving reliability and performance. There are many different paths toward achieving these goals. The more important of these are listed in Table 1.

### Table 1: Integrated-Circuit Classifications

<table>
<thead>
<tr>
<th>Technology</th>
<th>Applications</th>
<th>Complexity</th>
<th>Selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monolithic</td>
<td>Digital</td>
<td>SSI</td>
<td>Standard</td>
</tr>
<tr>
<td>Bipolar</td>
<td>Linear</td>
<td>MSI</td>
<td>Semicustom</td>
</tr>
<tr>
<td>MOS</td>
<td>LSI</td>
<td></td>
<td>Custom</td>
</tr>
<tr>
<td>Hybrid</td>
<td>VLSI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thin film</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thick film</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

14. **Integrated-Circuit Technology**

**Monolithic Circuits.** Monolithic circuits are those in which all associated components are fabricated and interconnected on or within a single chip of silicon. Processes employed are basically separated into bipolar, using the two-junction transistor as its basic element, and MOS, using the unipolar field-effect structure as its building block.

Bipolar processing generally yields greater speed and permits the design of high-voltage and higher-power circuits. MOS processing, including the widely implemented complementary MOS (CMOS) structure, offers greater component density and much lower internal power dissipation. Therefore, bipolar circuits are widely used in linear circuits, in small-scale-integration (SSI) and medium-scale-integration (MSI) logic circuits, where they are so well entrenched that they cannot be economically replaced, and in circuits where state-of-the-art speed is required.

MOS circuits, particularly CMOS circuits, are utilized primarily in digital LSI and VLSI designs, where their small geometries and extremely low power requirements permit implementation of circuit complexities that cannot readily be achieved by other means.

Progress is being made in combining bipolar and MOS technologies, each to its own advantage, within a single chip, and some such circuits are being marketed. Generally, however, the two technologies are being used independently.

**Hybrid Circuits.** Hybrid circuits utilize a combination of monolithic integrated circuits and discrete components attached to and interconnected by a thin- or thick-film conductive pattern deposited on an insulating glass or ceramic substrate (Fig. 6.47). They are used basically for size and weight reduction of complex circuitry whose component complement does not permit monolithic fabrication. High-power circuits and radio-frequency circuits which require capacitors and inductors are typical candidates for thin-film construction techniques.

The advantages of thin-film circuits over conventionally wired circuits using printed-circuit boards lie in the fact that most of the passive elements of a complete circuit, such as resistors and capacitors, can be fabricated directly and simultaneously on the insulating substrate. Similarly, all interconnections can be made at one time, often in conjunction with one of the deposition steps of the passive elements. Then the active components (or monolithic ICs) are added separately and interconnected with the previously deposited pattern on the substrate. To save space, active elements are often utilized in chip form and wire-bond
techniques are employed for subsequent interconnections. This practice, however, introduces manufacturing and testing complications. The recent proliferation of surface-mounted packages for both discrete active components and ICs (Fig. 6.48) greatly simplifies hybrid-circuit manufacturing with little penalty on space utilization.

Digital Circuits. Digital circuits are those dealing exclusively with the technology of ones and zeros. Their active elements, predominantly transistors, are used as switches which are either on or off. These circuits are used as digital logic elements in computer and control applications. They are composed primarily of basic building blocks, called gates, which are interconnected in a variety of ways to form more complex functions such as flip-flops, counters, and a myriad of other circuit entities. In all instances, however, the basic
A circuit is driven by a signal whose voltage level is either high or low (one or zero), causing the circuit to be turned on or off and the output voltage to be either high or low, or true or false, depending on the terminology associated with a particular application. Generally, digital circuits, no matter how complex, are operated at low voltages, normally ranging between 3 and 10 V.

Linear Circuits. Linear circuits are not as clear-cut in definition. Generally, they encompass analog functions, such as amplifiers, whose output is proportional to the level of the input signal. Also included in the linear category are analog-to-digital (A/D) and digital-to-analog (D/A) converters, comparators, voltage regulators, and even interface circuits whose functions are more digital than analog in nature.

In recent years, the proliferation of circuits and the increasing complexity which permits the combination of digital and linear functions in a single chip have made these classifications virtually obsolete. The present trend is toward definitions that classify circuits in terms of their primary intended applications. Categories such as logic circuits, microcomputer circuits, voice-data circuits, analog circuits, power-conversion circuits, etc., are becoming more commonplace. While such definitions also have limitations and while there is as yet no standardization of such classifications, they serve as the basic subsections for the following discussions. These discussions, moreover, will be limited primarily to standard circuits, available off the shelf from IC manufacturers and distributors as the basic building blocks for system designs. A brief discussion of the growing trend toward semi-custom and custom approaches is included.

15. Integrated-circuit complexity. Integrated circuits in general and logic circuits in particular are classified within four basic categories: small-scale integration (SSI), medium-scale integration (MSI), large-scale integration (LSI), and very-large-scale integration (VLSI). Initially, each of these encompassed a degree of chip complexity involving a specified number of equivalent gate circuits (gate equivalents). The rapid progress in IC technology has rendered this practice virtually obsolete, since the VLSI circuit of only a decade ago would hardly be classified as more than MSI today. Nevertheless, the terms still apply on a relative basis, although the judgment of where one classification ends and the other begins is largely a matter of personal interpretation.

16. Integrated-Circuit Selection

Standard versus Custom Circuits. From the very beginning of the technology, the development and production of state-of-the-art ICs has been a time-consuming and expensive endeavor. For IC manufacturers, each new circuit represented a gamble on whether the device would sell in large enough quantities to permit recovery of the development cost and, subsequently, to produce a profit. In the earliest days, the odds were good because there were enough universally required circuits to assure large-scale acceptance. Gates and flip-flops for digital equipment, operational amplifiers, and voltage regulators in the linear field were certain to find ready markets. IC manufacturers found it easy to identify such circuits and rapidly built up inventories of a large variety of “standard” off-the-shelf devices that were sold to all comers.

Even then, however, some equipment manufacturers opted to have special proprietary circuits designed on a custom basis if their volume requirements were large enough to amortize development costs. In many instances, these custom circuits, after an agreed-upon time, became standard circuits offered on the open market.
As the technology advanced, it became more difficult to identify increasingly complex circuits that would find a large general market. Accordingly, the development of standard circuits has become a greater gamble for the manufacturer. And as the demand for ever-increasing device complexity has increased, the number of original-equipment manufacturers (OEMs) able to afford the cost of designing such proprietary VLSI circuits is dwindling. As a result, an intermediary technology has emerged that permits the development of very complex semicustom circuits, utilizing a library of predefined basic circuits.

Called applications-specific integrated circuits (ASICs), these devices sacrifice some custom-design flexibility, by being constrained to the basic-circuit library, but are much less costly and less time-consuming to produce. ASIC technology is applicable largely to circuits so complex that they can be implemented only through computer-aided-design (CAD) techniques. Currently, two types of ASIC techniques have reached production status: gate arrays and standard cells.

**Gate Arrays.** The most complex digital logic circuits can be designed through the interconnection of basic gate circuits in various unique arrangements and sequences. IC manufacturers take advantage of this fact by producing complete wafers full of identical basic gate circuits. Then they provide a computer-stored program for interconnecting the gates into circuits of increasing complexity. During the design cycle, the system designer formulates a complex design by using the library of subcircuits issued by the IC manufacturer. The computer-stored program is then utilized to develop an interconnect system that converts the network of gates on the wafer into an exact duplication of the complex design. Use of prefabricated gate arrays and a proven CAD program saves months of time compared with custom design of VLSI circuits.

**Macrocell Arrays.** Macrocell arrays are similar to gate arrays, except that the prefabricated wafers do not contain prewired basic gates but rather a pattern of islands each of which contains a number of prediffused transistors and resistors (Fig. 6.49). The IC manufacturer issues a library of predefined functions called *macros* that the computer program can produce from the discrete parts within each island and subsequently interconnect into a custom IC. In theory, the macrocells—flip-flops, counters, registers, and a wide variety of additional digital building blocks—can be designed more efficiently from discrete components than from basic gates, yielding smaller geometries (equivalent to lower cost) and improved performance. As with gate arrays, the designer develops VLSI circuits with the macros contained in the library, and the computer develops the necessary mask patterns that convert the prediffused macrocell wafers into functional VLSI circuits.

**Commercial Implementation.** The efficiency of the gate-array–macrocell-array concepts depends upon the availability of arrays that closely match the needs of the designer. For example, a prediffused array with fewer on-chip gates (or islands) than required for the envisioned VLSI circuit would increase the number of IC packages; an array with too many on-chip gates would result in unused wafer space. IC manufacturers therefore offer a variety of array sizes (Table 2) that permit a close match between need and availability. Moreover, such arrays are available for a variety of technologies (bipolar and MOS) to offer a number of performance options.

**Standard Cells.** Standard-cell concepts go one step further toward true custom designs than gate or macrocell arrays do. These too start with a computer-stored catalog of basic circuits (standard cells) that form the building blocks for VLSI designs. Rather than using prediffused wafers of basic cells, with each cell requiring an equal
amount of chip space, a standard cell is designed to provide maximum efficiency with a layout requiring a minimum amount of space. The computer then provides the necessary mask patterns for the fabrication of the complete circuit, not just the interconnect patterns for the prediffused arrays. The use of custom-designed standard cells, compared with equivalent cells made from array components, can save a significant amount of chip space (Fig. 6.50) and produce end results that closely rival those of true custom circuits. Yet, because these end circuits are fashioned from pre-defined building blocks with the computer doing much of the design, development time is substantially reduced.
### TABLE 2  Macrocell Arrays*

<table>
<thead>
<tr>
<th>Technology:</th>
<th>ECL</th>
<th>TTL</th>
<th>ECL-TTL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bipolar array number:</td>
<td>MCA 600 ECL</td>
<td>MCA 1200 ECL</td>
<td>MCA 2500 ECL</td>
</tr>
<tr>
<td>Gate equivalent</td>
<td>652</td>
<td>1192</td>
<td>2472</td>
</tr>
<tr>
<td>Major or primary cells</td>
<td>24</td>
<td>48</td>
<td>110</td>
</tr>
<tr>
<td>Macrocell components</td>
<td>67</td>
<td>106</td>
<td>178</td>
</tr>
<tr>
<td>I/O ports</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Input only</td>
<td>28</td>
<td>34</td>
<td>52</td>
</tr>
<tr>
<td>Output only</td>
<td>. . .</td>
<td>. . .</td>
<td>. . .</td>
</tr>
<tr>
<td>Uncommitted</td>
<td>18</td>
<td>26</td>
<td>68</td>
</tr>
<tr>
<td>Maximum gate delay, ns</td>
<td>1.2</td>
<td>1.2</td>
<td>0.5</td>
</tr>
<tr>
<td>Maximum toggle frequency, MHz</td>
<td>160</td>
<td>160</td>
<td>400</td>
</tr>
<tr>
<td>Power dissipation, W (typical)</td>
<td>2.2</td>
<td>4.0</td>
<td>6.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Technology</th>
<th>3-μm silicon-gate HCMOS</th>
<th>2-μm silicon-gate HCMOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMOS array number</td>
<td>HCA 6306</td>
<td>HCA 6312</td>
</tr>
<tr>
<td>Gate equivalent</td>
<td>648</td>
<td>1200</td>
</tr>
<tr>
<td>Major or primary cells</td>
<td>216</td>
<td>400</td>
</tr>
<tr>
<td>Macrocell components</td>
<td>104</td>
<td>104</td>
</tr>
<tr>
<td>I/O ports</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Input only</td>
<td>2 + 1 or 0 + 1</td>
<td>17 + 1</td>
</tr>
<tr>
<td>Output only</td>
<td>. . .</td>
<td>. . .</td>
</tr>
<tr>
<td>Uncommitted</td>
<td>35</td>
<td>42</td>
</tr>
<tr>
<td>Maximum gate delay, ns</td>
<td>2.5 typical</td>
<td></td>
</tr>
<tr>
<td>Maximum toggle frequency, MHz</td>
<td>70 typical</td>
<td></td>
</tr>
<tr>
<td>Power dissipation, W (typical)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*These arrays run the gamut of technologies and offer a variety of cell configurations. Comparative speed-power specifications are given. [Motorola Inc.]
Logic circuits are implemented with either bipolar or MOS technologies. Bipolar logic circuits encompass a variety of different basic “families,” each with its unique advantages and limitations. These various families are divided into two generic categories: saturated logic and nonsaturated logic. In saturated-logic designs, the transistor (acting as a switch) is driven from the cutoff condition to the saturated condition by the input signal. In nonsaturated logic, the transistor is not switched from full off to full on but swings above and below a specified reference level. Thus, nonsaturated logic avoids some of the inherent transistor time delays associated with saturated-mode switching but pays for the resulting improvement in switching speed with increased power consumption.

Integrated logic circuits have gone through a number of iterations. The early forms, carryovers from discrete designs, were bipolar resistor-transistor-logic (RTL) and diode-transistor-logic (DTL) circuit designs, all utilizing saturated-logic forms. These have been largely replaced with a variety of transistor-transistor-logic (TTL) designs utilizing bipolar technology, with bipolar nonsaturated emitter-coupled-logic (ECL) designs where maximum speed is required, and with CMOS (complementary MOS) designs using MOS technology where low power consumption is a basic requirement. In general, SSI and MSI circuits are being implemented with bipolar and CMOS technologies, both of which offer a wide selection of standard circuits as building blocks. LSI and VLSI designs utilize CMOS technology to an ever-increasing degree.

17. Transistor switching characteristics. While the performance of a logic family is judged on a number of characteristics, the most important of these is switching speed. Switching speed defines the time it takes for a transistor, or gate, to change its output from one state to another in response to a change in the input signal.

Figure 6.51 shows a typical switching transistor and its associated input and output waveforms. In a “perfect” switch, the output current would follow the input-voltage waveform exactly. The practical circuit involves a number of delays. Note that collector current...
does not begin to flow until time $t_1$, some time after $t_0$, where the base voltage rises to its maximum value. The difference between $t_0$ and $t_1$ is the turn-on delay time ($t_{on}$), caused by the need to charge or discharge the transistor interelectrode capacitances ($C_B$ and $C_C$) before current flow can begin.

At $t_1$, collector current begins to flow but cannot rise instantly to its maximum value because of the finite time required for charge carriers to traverse the base region of the transistor between the emitter and the collector. The time required for the current to rise from 10 to 90 percent of its maximum value (from $t_1$ to $t_2$) is called rise time ($t_{rise}$).

When the transistor is driven into saturation, both the emitter and the collector junctions are forward-biased and dump minority carriers (electrons, in the case of an $n$-$p$-$n$ transistor) into the base region. When the base voltage suddenly drops to zero, at $t_3$, these carriers must be ejected through the collector circuit. Hence, current continues to flow until all these carriers have been pulled out of the base region. The time difference between $t_1$ (the drop of the base signal to zero) and the point where collector current begins to drop at $t_4$ is called storage time ($t_s$). Even then, collector current cannot drop instantly to zero because the interelectrode capacitances cannot discharge instantly. The gradual decay of collector current between $t_4$ and $t_5$ is the fall time ($t_f$).

From the above, it is clear that a practical transistor switch involves both a turn-on delay time ($t_{on}$) and a turn-off delay time ($t_{off}$). Of these, $t_{off}$ is by far the greater and is significantly influenced by $t_s$, which in turn results from driving the transistor into saturation.

**The Gate Function.**  The basic circuit of all logic elements is the gate circuit. Virtually all building blocks associated with digital equipment, flip-flops, counters, shift registers, etc., can be made by interconnecting a series of gate circuits. Hence, the complexity of any given logic circuit is specified by the number of equivalent individual gate functions it encompasses. The electrical characteristics of a basic gate circuit within a given logic family (TTL, ECL, CMOS, etc.) are often used for comparing the capabilities of various logic families.

A gate is simply an electrical or electronic switching circuit, with two or more inputs, through which power is applied to a load. The mechanical equivalent (Fig. 6.52a) shows an OR gate consisting of two switches (relays) connected in such a way that power is applied to the lamp (load) when either switch $A$ or switch $B$ is closed. Figure 6.52b shows an AND gate which applies power to the load only when switch $A$ and switch $B$ are closed.

In Fig. 6.52c, an additional relay is connected to the OR circuit so that power is applied to the load only when switch $A$ and switch $B$ are not closed. This is a NOT OR, or NOR, gate.

Similarly, an AND circuit connected in such a way that the load is powered only when switch $A$ and switch $B$ are not closed simultaneously (Fig. 6.52d) is a NAND gate.

The additional relays perform the function of an inverter; i.e., they invert a normally high (or 1) output of an OR or AND gate into a low (or 0) output, and vice versa. Hence, an OR or AND gate can be converted to the NOR or NAND function by running the outputs through an inverter.

In Fig. 6.53, the switching relays of Fig. 6.52 have been replaced by diodes which serve similar functions. In the OR gate (a), the output is high (or 1) when the input to diode $A$ or diode $B$ is high. In the two-input AND gate (b), the output is high only when both diode inputs are high. If either input were low, the corresponding diode would conduct and the output would be equal to the very low (0) voltage drop of the conducting diode.
FIGURE 6.52 Basic mechanical gate functions. (a) Mechanical equivalent of a two-input OR gate and its representative functional truth table which illustrates the status of the output $o$ as a function of the status of inputs $A$ and $B$. (b) Mechanical AND gate and its accompanying truth table. (c) NOR gate and its truth table. (d) NAND gate and its truth table.
When a common-emitter transistor circuit is added to the above circuits (Fig. 6.54), the transistor automatically converts a low-level input into a high-level output (and a high-level input into a low-level output), thereby acting as an inverter. Thus, a transistor added to a diode AND gate results in a NAND gate, and a transistor added to a diode OR gate results in a NOR gate.

18. Saturated-logic gates. In many switching-circuit families the transistors are driven from deep-in cutoff, where the collector voltage is equal approximately to the supply voltage, to a high level of conduction (saturation), where the collector junction becomes forward-biased, resulting in a very low collector voltage, $V_{CE(sat)}$, that remains relatively constant even if the base-drive voltage is further increased. In this latter condition, virtually all the power supply voltage is dropped across the collector load resistor—a limiting condition that prevents any further increase in collector current even if the charge injected into the base is increased by a further rise in base voltage. This is called saturated-mode switching.

Diode-Transistor Logic (DTL). The basic DTL gate circuit is the DTL NAND gate shown in Fig. 6.55. Note that it uses the NAND circuit of Fig. 6.54 plus an offset-voltage diode circuit. The offset-voltage diode circuit, being returned to a negative bias voltage through resistor $R_3$, improves noise immunity and transistor turn-off time.
Transistor-Transistor Logic (TTL). The DTL circuit just described was the basis of a more advanced logic line currently in use. Known as transistor-transistor logic (TTL), this logic is considerably faster than other forms of saturated logic.

The basic diagram of a typical TTL gate circuit is shown in Fig. 6.56. The principal difference between DTL and TTL is that the latter uses the collector-base junctions of a multiple-emitter transistor \( Q_1 \) as the input diodes and the collector-base junction as the offset diode. This, however, has a considerable effect on the operation, as follows:

When the input voltage to one or more of the \( Q_1 \) emitters is very low, say, zero, the base-emitter junction is forward-biased, through \( R_1 \), to a high positive voltage. Current flow through this junction, therefore, establishes approximately 0.7 V at the base. The collector of \( Q_1 \) is returned to ground through the base-emitter junctions of \( Q_2 \) and \( Q_3 \). Thus, the
collector-base junction is also forward-biased, and it would appear that there should be a current flow through the emitter-base junctions of Q2 and Q3. This cannot occur, however, because it would require a Q1 base voltage of at least three junction drops \([V_{BEQ1} + V_{BEQ2} + V_{CBQ1}]\), or 2.1 V, to turn both Q2 and Q3 fully on. Therefore, Q2 and Q3 remain cut off and the output voltage \(V_{out}\) is equal to \(V^+\) (minus a small leakage drop).

Now if the input voltage is increased to approximately 0.6 V, the Q1 base voltage rises to 1.3 V, and the collector junction of Q1 and the base junction of Q2 will be brought to the threshold of conduction through \(R_e\). As input voltage increases still further to about 1.3 V, Q1 base voltage rises to 2.0 V, the collector and Q2 base junctions become fully conductive, and the voltage across \(R_e\) rises to 0.6 V. This brings Q1 to the threshold of conduction, so that any additional increase in input voltage causes Q1 to turn on and the output voltage drops to \(V_{CE(sat)}\).

For most of the operating cycle the multiple-emitter input transistor acts as though it were a pair of back-to-back diodes. During one small part of the turn-off operation it does act like a transistor, thereby improving circuit speed. Specifically, when the input is in the high state and both Q2 and Q3 are conducting, the voltage at the base of Q1 is 2.1 V while the base voltage of Q2 is 1.4 V. Thus, Q2 is in saturation, and a heavy charge is built up in its base region. Now the input suddenly drops to zero and the Q2 base voltage drops to 0.7 V (the forward voltage drop of the input diode). If the Q2 collector-base junction were an ordinary diode, it would be reverse-biased and Q2 would be suddenly cut off. As a result, the stored charge in the Q2 base would have to be dissipated by recombination, and Q2 would continue to conduct until equilibrium was restored. Storage time, therefore, would be relatively long.

With the transistor, however, for the conditions existing at Q2 turn-off, the emitter of Q1 quickly drops to nearly zero volts. The base drops to 0.7 V, but the collector remains near 1.4 V because of the charge in the base of Q2. These are the proper biases for transistor action and cause a sudden surge of Q1 collector current that quickly clears Q2 of its stored charge. Thus, storage time of Q2 is drastically reduced.

19. **Nonsaturated-logic gates.** See Secs. 20, 21, and 22.

20. **Schottky TTL.** The basic TTL family has been substantially improved by a number of variations, involving the use of Schottky diodes to keep the transistors out of saturation during turn-on. Since the switching time delay associated with transistor storage time is the result of driving the device into saturation, nonsaturated logic can considerably improve switching speed.

Schottky TTL simply involves the use of a Schottky diode between collector and base of a transistor, as shown in Fig. 6.57. Since this configuration can be achieved during the fabrication of the transistor without the use of a separate diode, it adds little to the device cost, but it has a great effect on performance.

Specifically, since the turn-on voltage for the diode is considerably less than that for a conventional junction diode (in this case, the collector-base junction of the transistor), diode conduction occurs before the \(V_{CE(sat)}\) region of the transistor is reached. Therefore, the collector voltage will be clamped at a value that prevents the collector-base junction from becoming forward-biased, and storage-time delay during turn-off is avoided (see Secs. 17 to 23).

![FIGURE 6.57 A representative Schottky-clamped transistor (a) is normally indicated by the schematic symbol in (b).](image-url)
By raising the minimum achievable collector voltage somewhat, the maximum output voltage of the transistor is correspondingly reduced. This lowers the noise margin slightly, but not enough to be seriously detrimental for most applications.

Noise margin can be simply described with the aid of the basic RTL inverter circuit and transfer characteristics in Fig. 6.58. From the transfer characteristics it is seen that when the input to the transistor is low (say, 0 to 0.6 V in this instance), the output voltage is high and current is delivered to the load. As the input voltage rises to approximately 0.6 V, the transistor begins to conduct and its output rapidly drops toward $V_{CE(sat)}$ (approximately 0.1 V). For any input voltage greater than 0.9 V, the output remains at $V_{CE(sat)}$. The output voltage swing therefore is from 0.1 to 1.65 V.

Under low-input conditions, if a noise pulse of 0.5 V or higher were to appear at the input of the gate, the noise voltage would add to the 0.1-V input voltage and the gate would tend to change its output from high to low. Therefore, the gate is immune to noise signals up to 0.5 V, or it has a high-state noise margin of 0.5 V.

The RTL circuit is used in the above example because it permits a simple discussion of noise immunity (it is no longer a popular logic format). It also offers a simple demonstration of the effect of fanout (FO) on noise immunity. In this instance, if the output voltage were applied simultaneously to five succeeding stages, the high-state voltage would drop to approximately 1 V and the low-state noise immunity would approach zero.

Schottky TTL logic families now include a variety of circuit innovations. All are based on the familiar 54/74 TTL series, which for years has been the dominant logic family for general-purpose applications. Only within the last few years has the Schottky LSTTL series made significant inroads, replacing the original 7400 series with lower power requirements and higher speed. The latest in this progressive logic series are the advanced low-power Schottky (ALSTTL) family and the FAST Schottky TTL series.

The low-power Schottky (LSTTL) family combines an improvement in current and power reduction over the standard 7400 TTL by a factor of 5. This is accomplished by advanced processing and by using Schottky diode clamping to prevent saturation.

The advanced low-power Schottky TTL family (ALSTTL) provides a 50 percent power reduction compared with the standard 54/74 LSTTL and yet offers improved circuit performance over the standard LS owing to a state-of-the-art oxide-isolated process (MOSAIC).
ALS also differs from LS in that p-n-p transistors on the input stage are utilized to lower input currents and raise thresholds.

The FAST Schottky TTL family provides a 75 to 80 percent power reduction compared with the standard Schottky 54/74S TTL and yet offers a 20 to 40 percent improvement in circuit performance over the standard Schottky owing to the MOSAIC process. Also, FAST circuits contain additional circuitry to provide a flatter power-frequency curve. The input configuration of FAST uses a lower input current, which translates into higher fanout. Speed-power comparisons of these three families are indicated in Table 3.

### Table 3

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Symbol</th>
<th>LS</th>
<th>ALS</th>
<th>FAST</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quiescent supply current or gate</td>
<td>$I_{G}$</td>
<td>0.4</td>
<td>0.2</td>
<td>1.1</td>
<td>mA</td>
</tr>
<tr>
<td>Power or gate (quiescent)</td>
<td>$P_{G}$</td>
<td>2.0</td>
<td>1.0</td>
<td>5.5</td>
<td>mW</td>
</tr>
<tr>
<td>Propagation delay</td>
<td>$t_{P}$</td>
<td>9.0</td>
<td>5.0</td>
<td>3.7</td>
<td>ns</td>
</tr>
<tr>
<td>Speed-power product</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clock frequency (D-FF)</td>
<td>$f_{max}$</td>
<td>33</td>
<td>35</td>
<td>125</td>
<td>MHz</td>
</tr>
<tr>
<td>Clock frequency (counter)</td>
<td>$f_{max}$</td>
<td>40</td>
<td>45</td>
<td>125</td>
<td>MHz</td>
</tr>
</tbody>
</table>

*All typical ratings.

21. **Emitter-coupled logic (ECL).** Emitter-coupled logic is the fastest form of logic currently available. It is nonsaturated logic which is extremely flexible, with both OR and NOR logic outputs available from the same basic gate configuration.

ECL utilizes a pair of input transistors, one of which is in a conductive state while the other is nonconductive. Switching is accomplished by means of a signal appearing across a common-emitter resistor. The basic gate circuit of an emitter-coupled logic circuit is shown in Fig. 6.59. It consists principally of a switching circuit, followed by an emitter-follower output circuit. The switching circuit consists of transistors $Q_1$ and $Q_2$, which are connected in a differential amplifier configuration. Operation is as follows:

Assume for a moment that all input transistors $Q_1$ are cut off owing to logical 0 being applied to their gates. $Q_2$ has its base connected to a fixed stable bias source $V_{BB}$ of −1.15 V, which causes $Q_2$ to conduct heavily since its emitter is connected, through $R_E$, to −5.2 V. Under these conditions the voltage drop across the base-emitter junction of $Q_2$ is approximately 0.75 V, and the current through $Q_2$ is

\[ I_{Q2} = \left[ V_{EE} + V_{BB(Q2)} + V_{BE(Q2)} \right] / R_E \]

\[ = \left( -5.2 + 0.75 + 1.15 \right) / 1.24 \]

\[ = -2.66 \text{ mA} \]

and $V_{C2} = I_{Q2}R_{C2} = -0.8 \text{ V}$

In the on condition, therefore, $Q_2$ is not in the saturated mode because the voltage across the collector-base junction is still positive ($V_C - V_B$), keeping this junction reverse-biased.

Transistor $Q_3$, seeing −0.8 V on its base compared with −5.2 V as an emitter source, goes into heavy conduction, yielding the OR output of

\[ V_{OR} = V_{C(Q2)} - V_{BE(Q2)} = -0.8 + 0.75 = -1.55 \text{ V} \]
which represents the output-low condition.

At the same time, the base of \( Q_4 \), being connected to the collector of \( Q_1 \) (which is nonconductive), sees a voltage of zero, causing this transistor to conduct even more heavily. Its output voltage is

\[
V_{\text{NOR}} = V_{C(Q1)} - V_{BE(Q4)} = -0.75 \text{ V}
\]

which represents the output-high condition indicative of the NOR function.

If a positive-going signal is now applied to one or more of the \( Q_1 \) transistors, the transistors begin to conduct, tending to increase the emitter voltage of \( Q_2/Q_3 \). In turn, this reduces the forward bias on \( Q_2 \), causing a nearly equivalent decrease in \( Q_2 \) collector current. Thus, it is clear that emitter current remains essentially constant but is switched from \( Q_2 \) to \( Q_1 \). At the same time, the output-voltage levels of \( Q_3 \) and \( Q_4 \) reverse from their
previous states, which again is consistent with their respective OR/NOR functions. The
transfer characteristics illustrated in the graph of Fig. 6.59b indicate the following:

1. When an ECL gate is driven by another ECL gate, it provides two output signals: one
   identical to the input, the other equal to the inverted input.
2. The fixed logic levels are –0.75 (logical 1) and 1.55 (logical 0) V, resulting in a differ-
   ence of 800 mV between the on state and the off state.
3. The transistors are never driven into saturation; therefore, storage time is eliminated.
4. The noise margin is approximately 250 mV.
5. The total current flow in the circuit remains relatively constant, thereby maintaining a rela-
   tively constant drain on the power supply and preventing internally generated noise spikes.

   The circuit, furthermore, has a very high input impedance and a very low output imped-
   ance, which makes high fanout possible. Thus, ECL not only is the highest speed logic
   available but provides other features that make it highly desirable for advanced systems. Its
   principal detriment is its limited (800-mV) logic swing, which makes it somewhat more
   sensitive to externally generated noise than some other logic forms. In addition, power con-
   sumption is considerably higher than for other logic forms.

   Like TTL, emitter-coupled logic also comprises a number of families with progressively
   improved performance. Three families are in popular use: ECL 10K, ECL 10KH, and ECL III.

   The ECL 10K series has become the industry standard for high-speed applications. To
   make the circuits comparatively easy to use, edge speed (rise and fall times) are deliber-
   ately slowed to 2.0 ns while the important propagation delay is held to 2.0 ns. The slow
   edge speed permits use of wire-wrap and standard printed-circuit lines; however, the cir-
   cuits are specified to drive transmission lines for optimum performance.

   The newer ECL 10KH family features 100 percent improvement in propagation delay
   and clock speeds while maintaining a power supply current equal to that of the ECL 10K.
   This ECL family is voltage-compensated, which allows guaranteed dc and switching para-
   meters over a ±5 percent power supply range. Noise margins of ECL 10KH are 75 percent
   better than those of the ECL 10K series. ECL 10KH is compatible with ECL 10K and ECL III,
   a key element in allowing users to enhance existing systems by increasing the speed in crit-
   ical timing areas.

   ECL III, with its 1-ns gate propagation delays and greater than 1-GHz flip-flop toggle
   rates, is the industry speed leader. The 1-ns rise and fall times require a transmission-line
   environment for all but the smallest systems. For this reason, all circuit outputs are designed
   to drive transmission lines, and all output logic levels are specified when driving 50-Ω loads.
   Because of ECL III’s fast edge speeds, multilayer boards are recommended above 200 MHz.
   ECL III’s popularity is with high-speed test and communications equipment.

   Speed-power comparisons for Motorola ECL families are given in Table 4.

   **TABLE 4**  ECL Family Comparisons

<table>
<thead>
<tr>
<th>Parameter</th>
<th>MECL 10K</th>
<th>MECL 10KH</th>
<th>10,100 series</th>
<th>10,200 series</th>
<th>10,500 series</th>
<th>10,600 series</th>
<th>MECL III</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gate propagation delay</td>
<td>1.0 ns</td>
<td>2.0 ns</td>
<td>1.5 ns</td>
<td>1.0 ns</td>
<td>2.0 ns</td>
<td>1.0 ns</td>
<td>1.0 ns</td>
</tr>
<tr>
<td>Output edge speed</td>
<td>1.2 ns</td>
<td>2.2 ns</td>
<td>2.0 ns</td>
<td>1.0 ns</td>
<td>2.0 ns</td>
<td>1.0 ns</td>
<td>1.0 ns</td>
</tr>
<tr>
<td>Flip-flop toggle speed</td>
<td>250 MHz min</td>
<td>125 MHz min</td>
<td>200 MHz min</td>
<td>300–500 MHz min</td>
<td>300–500 MHz min</td>
<td>300–500 MHz min</td>
<td>300–500 MHz min</td>
</tr>
<tr>
<td>Gate power</td>
<td>25 mW</td>
<td>25 mW</td>
<td>25 mW</td>
<td>25 mW</td>
<td>25 mW</td>
<td>25 mW</td>
<td>25 mW</td>
</tr>
</tbody>
</table>
CMOS logic. While individual p-channel and n-channel MOS devices have certain performance advantages over bipolar structures and while a p-channel logic line did gain some popularity for a short time, these advantages were not sufficient to displace bipolar logic. Not until processing technology permitted complementary MOS cells to become space-competitive with individual bipolar cells did MOS become a serious challenge. Now, CMOS logic families promise to become the dominant logic form over the next few years, particularly in the areas of general-purpose logic, where speed is not an overriding consideration, and in VLSI circuitry, where high chip density demands the lowest possible dissipation.

Some of the features of CMOS are attributable directly to the basic MOS transistor structure, while others are the result of (or are enhanced by) the complementary symmetry configuration. For example, the MOS transistor inherently has a very high input resistance, thereby eliminating dc fanout restrictions and providing low power dissipation. The CMOS configuration reduces power dissipation even more and increases speed, noise immunity, and logic swing.

A simple p-channel MOSFET inverter and its transfer characteristics are shown in Fig. 6.60a. (The use of a second [fixed-bias] MOSFET in place of a conventional load resistor is particularly beneficial since a resistor would require far more chip area. Thus, MOS is cost-effective by increasing the permissible number of circuits on a wafer.)

As indicated by the transfer characteristic curve, for an input voltage \( V_{in} \) between ground and \( V_t \), transistor \( Q_s \) is off, so the output voltage \( V_o \) approaches the \( V^- \) state. As the input voltage approaches threshold voltage \( V_t \), \( Q_s \) begins to conduct, and for further increases of \( V_{in} \) the output voltage is reduced. Note, however, that for a fixed MOSFET load resistance \( V_o \) can never reach zero because the resistance of \( Q_o \) never reduces to zero regardless of the value of \( V_{in} \). In fact, its on resistance is substantially higher than that of bipolar transistors. Therefore, the total output-voltage swing is always less than the supply voltage \( V^- \).

The complementary configuration (Fig. 6.60b) results in far more satisfactory performance. In this connection, the signal is applied simultaneously and in phase to both transistors so that when the signal value is zero, \( Q_1 \) is off and \( Q_2 \) is on. Under this condition, the output voltage is very nearly the full supply voltage. When the gate voltage goes high (positive), transistor \( Q_1 \) is turned on while \( Q_2 \) is turned off. This causes \( V_o \) to go virtually to zero because the current flowing through \( Q_1 \) is the leakage current of \( Q_2 \), which is very, very low. (The resistance of an MOSFET in cutoff is approximately 5000 M\( \Omega \), resulting in a leakage current of less than 1 nA.)

![Inverter circuits and their response characteristics. (a) Typical PMOS inverter circuit using a fixed-bias MOSFET as a load. Transfer characteristics compare response with a “perfect” switch. (b) Typical CMOS inverter circuit with p- and n-channel MOSFETs. Characteristics closely approach an “ideal” switch.](image-url)
The transfer characteristic curve for the complementary circuit is shown in comparison with that of a single-ended circuit. Observe that the slope of the complementary circuit curve is much steeper in the transition region, thus providing much greater noise immunity. This is caused by the input signal acting on both transistors in opposition, turning one device on and the other off.

In addition, it is evident that the circuit conducts current only during a very short time after turn-off (Fig. 6.61), as required to charge capacitor \( C_o \) through the load (the load is assumed to be another CMOS inverter circuit). This reduces power dissipation of the structures to such a low value that thousands of them can be fabricated within a tiny chip of silicon without special heat-sinking techniques.

Progressive improvements in technology have resulted in the current availability of two CMOS logic families: a standard metal-gate family and a high-speed silicon-gate family. In silicon-gate CMOS, the metal (aluminum) gate associated with standard CMOS is replaced by a heavily doped silicon layer which serves a similar function but has the following advantages:

1. It reduces the threshold voltage required to turn the transistor on.
2. It provides self-alignment of the gate electrode, thereby simplifying manufacture and improving fabrication accuracy.
3. It reduces gate-to-drain capacitance, thereby increasing circuit speed.
4. It permits improved space utilization, thereby increasing packing density.

General comparisons of speed and power requirements for the two CMOS families are indicated in Table 5.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Symbol</th>
<th>Standard CMOS</th>
<th>High-speed CMOS</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quiescent supply current or gate</td>
<td>( I_g )</td>
<td>0.0001</td>
<td>0.0003</td>
<td>mA</td>
</tr>
<tr>
<td>Power or gate (quiescent)</td>
<td>( P_g )</td>
<td>0.0006</td>
<td>0.001</td>
<td>mW</td>
</tr>
<tr>
<td>Propagation delay</td>
<td>( t_p )</td>
<td>125</td>
<td>8.0</td>
<td>ns</td>
</tr>
<tr>
<td>Speed-power product</td>
<td></td>
<td>0.075</td>
<td>0.01</td>
<td>pJ</td>
</tr>
<tr>
<td>Clock frequency (D-F/F)</td>
<td>( f_{\text{max}} )</td>
<td>4.0</td>
<td>40</td>
<td>MHz</td>
</tr>
<tr>
<td>Clock frequency (counter)</td>
<td>( f_{\text{max}} )</td>
<td>5.0</td>
<td>40</td>
<td>MHz</td>
</tr>
</tbody>
</table>

TABLE 5  Speed-Power Characteristics for CMOS Logic*

23. Comparison of logic-family characteristics. The variety of established and newly introduced logic families challenges the system designer with choosing the best available technology for each design. Each family offers distinct advantages and limitations.

The three most often used characteristics for determining family selection are propagation delay, operating frequency, and power consumption. For the logic families described here, these characteristics are displayed and compared in the graphs of Fig. 6.62. Graph \( a \) illustrates the tradeoffs between power dissipation and propagation delay at low operating frequencies. As frequency increases, the changes in power-dissipation characteristics of the various families vary considerably, as shown in graph \( b \). These factors, plus the ever-changing impact of economics, must be considered in the selection of the most suitable logic line.
Linear-Circuit Families. Linear (analog) integrated circuits are as pervasive and as unique as the markets they serve. They do not lend themselves readily to classification by building-block families, as do digital circuits for which standardized basic circuits can be used repetitively in various combinations to build highly complex end systems. Linear circuits tend to be complete functional circuits applicable to a wide variety of specific unique functions. And as circuit and chip complexity increases, this situation will become even more prevalent.

There are, however, a number of linear classifications that do encompass the family approach. Chief among these are operational amplifiers and voltage regulators. These represent basic circuits utilized in many different pieces of equipment, each of which requires some variation in performance specifications.

24. Operational amplifiers. Throughout history, there has been a continuous goal to develop the “ideal” device for a particular class of application: one that would satisfy all necessary requirements, thereby eliminating the need for any other product. In the field of signal amplifiers, the operational amplifier comes close to approaching that ideal.

Originally, the operational amplifier was designed principally for performing arithmetic operations. In such applications, high amplifier gain was not of critical importance, but absolute stability and accuracy were. These were achieved through the design of a high-gain amplifier that permitted the application of a large amount of negative feedback to provide the necessary accuracy.

This is still the basic design criterion for operational amplifiers. But while virtually any operational amplifier will serve a tremendous number of amplifier functions, the ideal device still remains to be invented. Hence, while the basic operational amplifier exhibits tremendous versatility, the operational-amplifier classification probably contains more type numbers than any other single applications category.

Basically, the operational amplifier is a multistage low-level amplifier with an open-loop gain \( A_{\text{VIL}} \) (before the application of negative feedback), ranging from a low of, say, 25,000 (25 V/mV) to 1 million or more. With such high gain levels readily available, further gain increases are apt to be by-products of improvements in other parameters rather than specific design goals. This is evident from Fig. 6.63, which shows the percent of amplifier gain error as a function of open-loop gain. It is clear that for a reasonable closed-loop–gain requirement, say, 100 or less, an open-loop gain of \( 1 \times 10^6 \) will permit sufficient feedback to hold gain error to negligible levels.

Since integrated circuits are intolerant of capacitors, analog functions are achieved with dc amplifier configurations. This usually involves differential amplifier input circuits which provide stable, direct-coupled amplification with very high circuit gain.
The block diagram for a typical operational amplifier is shown in Fig. 6.64a. Most integrated-circuit operational amplifiers follow this format. The primary differences are those of the circuits within each of the blocks that give the amplifier its ultimate performance specifications.

**FIGURE 6.63** When operational amplifiers are used at modest closed-loop gains, closed-loop-gain error can be held to negligible levels.

The block diagram for a typical operational amplifier is shown in Fig. 6.64a. Most integrated-circuit operational amplifiers follow this format. The primary differences are those of the circuits within each of the blocks that give the amplifier its ultimate performance specifications.

**FIGURE 6.64** A block diagram of a typical operational amplifier (a) and schematic diagram of an MC1709 circuit (b).
The first stage of an operational amplifier is a differential amplifier that provides most of the circuit gain. It is desirable to have high gain in this section so that any imperfections in succeeding stages (offset voltage, etc.) have little or no effect on the output. The first stage, too, should employ a current source at the common-emitter node for good common-mode rejection.

The second stage does not require a current source in the emitter because common-mode rejection and other matching-dependent characteristics of the total amplifier are determined primarily by the specifications of the first stage. It is needed primarily to provide additional gain. Its input resistance should be relatively high to prevent excessive loading of the first stage. Therefore, an emitter-follower or Darlington-amplifier stage is often employed.

Since a single-ended output is normally employed for the second stage, it follows that a dc voltage is present at its output. In a direct-coupled system, this dc level is propagated through the amplifier chain so that the amplifier output voltage would have a dc component in addition to a desired ac output signal. Therefore, some means of level translation is employed between the second and final stages. By eliminating the dc level at the final stage, the output voltage will vary about a zero reference level, thus preventing any undesired dc current in the load and also increasing the permissible output-voltage swing.

**Differential Amplifier.** The differential amplifier is the ideal circuit for direct coupling because of its versatility, its excellent stability, and its high immunity to interfering signals.

From the standpoint of stability, the circuit can be made virtually insensitive to temperature changes, which often cause excessive drift in other configurations. It is versatile in that it may be adapted for applications requiring floating inputs and outputs, as in the case of some sense amplifiers, or for applications in which grounded inputs and/or outputs are more desirable. In both cases, it exhibits the same drift-free interference-rejection capabilities.

To illustrate these capabilities, examine the schematic diagram of a simple differential amplifier shown in Fig. 6.65. Note that the two transistors with their respective collector resistors form a bridge which, if the transistor and resistor characteristics are identical, is perfectly balanced. Thus, the voltage across the output terminals is zero. If we now apply a differential-mode input signal and if $R_{i1}$ equals $R_{i2}$, so that the input voltages are equal in amplitude but opposite in phase, there will be a difference in voltage between the two output terminals which is proportional to the gain of the transistors.

If a common-mode input signal is applied (caused by line pickup or other interference), the input signals to each transistor will be equal in amplitude and in phase. The bridge will remain balanced, and the voltage between the output terminals will remain zero. Thus, the circuit provides high gain for differential-mode signals and no output at all for common-mode signals.

In applications in which output must be taken between one output terminal and ground, the common-mode signal will produce some output voltage, although there will be a very substantial reduction of gain for the common-mode signal as compared with the gain for the differential-mode input.

If, for example, a common-mode signal causes the current through the transistors to rise, the voltage drop across the common-emitter resistor also rises. This represents...
degeneration, and the gain of the transistors is very low. For differential-mode signals, the current through one transistor rises while that through the other transistor drops by an equal amount. Thus, the current through the emitter resistor remains constant and no degeneration occurs. The gain for the differential-mode signal, therefore, with matched conditions assumed, is unaffected whereas that for the common-mode signal is substantially reduced.

A modified version of the basic differential amplifier, using a constant-current source circuit in place of an emitter resistor (Fig. 6.66), represents the most common basic amplifier not only for operational-amplifier input stages but for virtually all low-power amplification purposes.

**Darlington Circuit.** In many applications for which an amplifier is used with a high-impedance signal source, it is desirable to have a higher input resistance than is available with conventional bipolar transistor circuits. In such cases, today’s technology offers field-effect transistors and Darlington (bipolar) transistor configurations. The Darlington circuit, used as the intermediate amplifier stage in Fig. 6.64, is shown in its simplistic form in Fig. 6.67.

The input resistance of a transistor is a function of its emitter resistance \( r_e \) and its \( \beta \) such that \( R_{in} = \beta r_e \). In the Darlington circuit shown, the input resistance of \( Q_2 \) becomes the emitter resistance (approximately) of \( Q_1 \), and the input resistance of the circuit is

\[
R_{in} = \beta r_e(Q1)
\]

but \( r_e(Q1) = \beta r_e(Q2) \)
Thus the input resistance of the Darlington transistor pair is extremely high, and its loading effect on the driving source is negligible. The total $\beta$ of the Darlington is also the square of the individual betas so that extremely high current gain can be obtained.

On the other hand, the ac input resistance of the pair is dependent on the input capacitance, which now shunts a much higher input resistance. Therefore, the frequency response curve of a Darlington drops off very rapidly as frequency is increased. In operational amplifiers, which are primarily intended for low-frequency applications, this is not a serious limitation, particularly since the high open-loop gain of such devices permits large amounts of negative feedback to compensate.

**Operational-Amplifier Specifications.** Operational-amplifier specifications, as listed on manufacturers’ data sheets, cover a wide range of characteristics. Among the most important are the following:

*Open-Loop Voltage Gain ($A_{\text{voll}}$).* This is normally specified in volts (output) per millivolts (input). This specification defines the maximum available voltage gain, without feedback, and gives an indication of the eventual accuracy that can be obtained when feedback is applied (see Fig. 6.63).

*Common-Mode Rejection Ratio (CMRR).* CMRR is defined as the ratio of common-mode input voltage to differential-mode input voltage that will yield the same differential-mode output voltage. With well-designed monolithic integrated circuits, common-mode rejection ratios can be as high as 100 dB.

*Input Offset Voltage ($V_{\text{io}}$).* Although IC technology permits close matching of adjacent transistors, perfect matching is not achievable. A slight mismatch in gain between the two differential-amplifier transistors will result in an undesirable dc voltage across the differential output terminals. $V_{\text{io}}$ is defined as the offsetting base-emitter voltage required for equal emitter currents in the two transistors.

*Input Offset Current ($I_{\text{io}}$).* $I_{\text{io}}$ is the difference in base currents required to produce equal emitter currents in differential-amplifier transistors.

*Input Bias Current ($I_B$).* $I_B$ refers to the amount of current flowing in input terminals under no-signal conditions. When an amplifier is required to respond to extremely small input signals, the input bias current limits the response. Today’s standard operational amplifiers have normal $I_B$’s on the order of only a few nanoamperes, and even this can be substantially reduced through the use of FET input devices to the picoampere range.

*Slew Rate (SR).* Slew rate refers to the maximum time rate of change of closed-loop–amplifier output voltage. It is determined by applying a step-function input signal and measuring the slope of the output pulse, as shown in Fig. 6.68a. Obviously, if an applied signal voltage varies more rapidly than the slew rate of the amplifier, the amplifier cannot respond exactly and the output signal is distorted. This can occur on a sine-wave signal, as illustrated in Fig. 6.68b, as well as on a pulsed signal.

Slew rate is specified in terms of the maximum signal-voltage change per microsecond that can be tolerated without signal distortion.

*Other Parameters.* Among other parameters normally specified for operational amplifiers are power supply voltage, which indicates the maximum output voltage obtainable; bandwidth, which indicates frequency response; and temperature coefficient of the input offset voltage, which describes the effect of temperature on operational-amplifier operation.
25. **Power supply circuits.** The power supply, which converts ac to dc operating voltages, is the one circuit that is universally required for all ac line-operated electronic equipment. In addition, the sensitivity of active components to changes in operating voltage and the potentially large variations in applied ac source voltage and load currents demand that power supplies for most equipment be well regulated in order to maintain stable operation under these variable conditions. Voltage-regulator circuits therefore are important functions for most power supplies.

Voltage regulators can be designed with discrete components to match the specific requirements of any equipment. However, the prevailing need for such circuits has resulted in a large variety of monolithic IC regulators to approximate most ideal requirements closely at a small fraction of the cost of equivalent discrete circuits.

26. **Voltage regulators.** A voltage regulator operates on the principle that a variable resistance in series or in shunt with the load resistance of an unregulated power supply can be used to compensate for any variation in output voltage or load resistance to keep the voltage across the load at a “constant” value.

The simplest shunt voltage regulator (Fig. 6.69a) is that of a zener diode in series with a voltage-dropping resistor \( R_s \). Operated in the breakdown region, the zener acts much like

---

**FIGURE 6.68** The effect of slew-rate limiting on a pulse input (a) and on a sine wave (b).
a voltage-activated variable resistor, in that the voltage across its terminals tends to remain constant while the current through the device varies in accordance with an applied voltage. Thus, in the simple circuit of Fig. 6.69a, if the input voltage to the circuit $V_i$ were to increase (for a constant load resistance) or if the load resistance were to increase (for a constant input voltage), the output voltage $V_o$ would tend to increase. As a result, the current through the zener would increase, and this increased current, flowing through $R_s$, counteracts the initial increase in $V_o$, thereby keeping this voltage close to its original value.

An improved regulator circuit is the series regulator of Fig. 6.69b. Here, the series resistance $R_s$ of the previous circuit is replaced with the collector-emitter resistance of a power transistor. The collector-base feedback resistor, in conjunction with the zener diode, varies the bias on the transistor, thereby causing a change in its internal resistance to compensate for any change in output voltage.

A further improvement in performance results from the circuit configuration in Fig. 6.69c, in which a differential amplifier (or operational amplifier) is used to control the resistance of the series-connected transistor. The differential amplifier continuously samples the output voltage, compares it with a fixed reference voltage, and produces an “error voltage” to control the bias on the series transistor. Because the differential amplifier has considerable gain, even very small changes in output voltage will be detected and compensated for. The circuit in Fig. 6.69c forms the basis of most IC voltage regulators in use today.
IC voltage regulators are available in a variety of configurations to meet the specific needs of the system designer. These configurations include the following:

1. Three-terminal fixed-output regulators with either positive or negative output
2. Three-terminal adjustable-output regulators with positive or negative output
3. Floating regulators for increased voltage range
4. Tracking regulators with a combination positive and negative output voltage
5. Switching regulators for use in switching power supplies

**Three-Terminal Fixed-Output Regulators.** Such regulators are available with a variety of fixed-output voltages in a range from approximately 2 to 40 V. Permissible output currents run the gamut from 100 mA to approximately 3 A, and both positive and negative output regulators are readily obtainable. The diagram in Fig. 6.70 illustrates the simple application and the range of voltage and current available for one commercial line of regulators, but even this wide variety of capabilities is expandable through the addition of external components. Paralleling the regulator with an external series-pass transistor (Fig. 6.71) can increase current-handling capacity; adding an external preregulator (Fig. 6.72) permits operation with higher input voltage than the maximum of 35 to 40 V for which this series of regulators was designed.

**FIGURE 6.70** Basic circuit configurations for positive fixed-output three-terminal regulators. \( C_{in} \): Required if regulator is located more than a few (about 2 to 4) inches away from input supply capacitor; for long input leads to regulator up to 1 \( \mu \)F may be needed for \( C_{in} \). \( C_{in} \) should be a high-frequency type of capacitor. \( C_0 \) improves transient response. XX: these two digits of the type number indicate nominal output voltage. Available voltages are 5, 6, 8, 12, 15, 18, 20, and 24 V.

**FIGURE 6.71** A current-boost configuration for positive three-terminal regulators. \( R \): used to divert IC-regulator bias current; determines at what output-current level \( Q_1 \) begins conducting. Values shown are for a 5-V, 5-A regulator using an MC7805CK on a 2.5°C/W heat sink and \( Q_1 \) on a 1°C/W heat sink for \( T_A \) up to 70°C. [Motorola Inc.]
The simplicity of obtaining the exact output voltage required with an adjustable three-terminal output-voltage regulator is shown in Fig. 6.73. As in the previously used series, maximum output voltage is approximately 40 V, but any value from 1.2 V to the maximum limit can be obtained through the adjustment of $R_2$. Current values from 100 mA to 3 A are available.

Floating Regulators. When wide flexibility in output voltage and current is required, a floating regulator can be employed. In the configuration illustrated in Fig. 6.74 the regulator is isolated from the main power supply by a dedicated power

![Diagram](https://via.placeholder.com/150)

**FIGURE 6.72** A preregulator for input voltages above the specified maximum input voltage of the regulator. Values shown are for $V_{in} = 60$ V. $Q_1$ should be mounted on a 2°C/W heat sink for operation at $T_A$ up to $70°C$. $IC_1$ should be appropriately heat-sunk for the package type used. [Motorola Inc.]

![Diagram](https://via.placeholder.com/150)

**FIGURE 6.73** An adjustable three-terminal regulator circuit.

![Diagram](https://via.placeholder.com/150)

**FIGURE 6.74** An MC1566, MC1466 floating-regulator configuration. For constant-voltage operation, output voltage $V_o$ is given by $V_o = (I_{ref})(R_2)$, where $R_2$ is the resistance from pin 8 to ground and $I_{ref}$ is the output current of pin 3. The recommended value of $I_{ref}$ is 1.0 mA dc. Resistor $R_1$ sets the value of $I_{ref}$ at $8.5/2R_1$, where $R_1$ is the resistance between pins 2 and 12. For constant-current operation, (a) select $RS$ for a 250-mV drop at the maximum desired regulated output current $I_{max}$; and (b) adjust potentiometer $R_2$ to set constant-current output at the desired value between zero and $I_{max}$. Values shown are for a 0- to 250-V, 100-mA regulator using an MCI486L with $Q_1$ and $Q_2$ mounted on a 1°C/W heat sink for $T_A = 70°C$. [Motorola Inc.]
source of its own. Its output-voltage and -current capabilities are limited only by the choice of external series-pass transistors.

**Tracking Regulators.** Applications requiring a dual polarity (+ and −) power supply, such as operational-amplifier circuits, can be served with tracking regulators. These consist of two regulators, interconnected so that the positive and negative regulators have the same output-voltage levels, which track each other in the event of parameter changes.

The circuit of Fig. 6.75 uses an MC1568 or MC1468 monolithic dual regulator for this purpose. Its outputs are set internally for ±15 V, but an external adjustment can change both outputs simultaneously from 8.0 to 20 V through the use of two balancing resistors.

Alternatively, tracking regulators can be developed by interconnecting two separate regulators, one positive and the other negative, in such a way that the output of one drives the input of the other.

**Switching Regulators.** Switching power supplies, or “switchers,” are making rapid inroads into the power supply market. Compared with linear supplies, they have significant advantages in efficiency, size, and weight. They are, however, more complex and in the past have been considered principally for high-power applications. With the advent of IC regulators and improvement and cost reductions of other solid-state components, the cost-performance tradeoffs are no longer severe even for medium- and low-power requirements. As anticipated, therefore, switching power supplies are rapidly increasing their market share.

The basic circuit configuration of a typical flyback switcher is shown in Fig. 6.76. In this off-line circuit, the ac line voltage is rectified by a bridge rectifier and filtered by capacitor $C_1$. The resulting dc voltage is then “chopped” by high-frequency square-wave pulses applied to the transistor in series with the primary-transformer winding. The resulting square wave is then passed through a high-frequency transformer. This provides the required step up or step down to produce the desired dc output after a second rectification and filtering process.

The output voltage is regulated by varying the width of the pulses applied to the chopper transistor. This is done through a pulse-width-modulator control circuit consisting of a triangular-wave generator, a pulse-width modulator, a dc reference source, and a comparator. The comparator senses the dc output voltage of the supply and compares it with the reference source. Any deviation of the output voltage from its design value will cause a control voltage from the comparator to provide a compensating pulse-width compression or expansion by the pulse-width modulator.
The control circuits are quite complex, as indicated by the block diagram of a typical unit in Fig. 6.77. Monolithic IC processing, however, has reduced the cost to a point at which switching power supplies are cost-competitive with series-regulated supplies for many applications. This is particularly true in view of the cost savings accruing from lower-cost transformer and filtering requirements.

**MICROCOMPUTERS**

Computer concepts in the 1940s were credited with revolutionizing the scientific and engineering fields by replacing people power with electronic power for computational purposes. In the 1950s, by using vacuum tubes, large computer installations performed rather primitive...
routines and proved the value of these concepts in actual practice. In the early 1960s, wide-
spread use of the transistor and its attending improvement in reliability, performance, and size
expanded the influence of the computer to include business and process control applications.
In the 1970s, integrated-circuit technology, with its attending cost reductions, spawned the
microprocessor, which broadened computer applications to affect virtually every human
endeavor and practice. From the laboratory to the factory floor, from the office to the ware-
house, from the kitchen to the automobile, the invasion of the microprocessor has created dra-
matic changes in equipment architecture and its use.

Computers, whether mainframe, mini, or micro, all have essentially the same basic
organization; that is, they all consist of a central processing unit (CPU) which performs the
basic arithmetic and logic operations in response to a sequential series of instructions
(program), a memory section which stores the program, and an input/output section (I/O)
which communicates with external equipment. The principal difference between the three
classes of computers is in the word length that can be processed and in the memory capacity,
being largest for mainframe and smallest for microcomputers. The larger the word size and
the internal the memory, the faster the operation of the computer and the larger and more
powerful the program which can be sorted and executed.

Additionally, mainframe computers and minicomputers are usually made with bipolar
components (TTL or ECL), which are inherently faster than the MOS technology employed
for microprocessor units (MPUs), the CPUs associated with microcomputers. But the
distinction between the various computer classifications is blurring as the latest microcom-
puters exceed the capabilities of today’s minicomputers and approach those of existing main-
frames. That is, microcomputers utilizing the latest 32-bit MPUs can serve applications
formerly in the domain of the more powerful minis and mainframes, depending on the amount
of memory and other peripheral circuitry used in association. It is likely that in the future the
distinction between computers will be based more on the amount of circuitry crammed into a
computer’s architecture than on the inherent capability of the technology employed.

The pervasiveness of the microcomputer, however, depends not so much on the upper
limit of its speed and processing power as it does on its miniscule size and its incredibly
low cost. It is this combination that makes a microcomputer suitable not only as a general-
purpose machine which must be reprogrammed for each specific end use but also as a pre-
programmed and dedicated “engine” that loses its identity as a computer and becomes an
electronic ignition system, a smart cash register with built-in inventory control, a control
system for home appliances, or one of countless other machines in everyday use.

27. Computer architecture. The diagram in Fig. 6.78 shows the basic components
of a typical computer. The heart of the system is the central processing unit (CPU), which
consists of an arithmetic-logic unit (ALU) and a control unit. In a microcomputer, these two
entities are normally processed on a single chip called a microprocessor unit (MPU).
Together, these two sections provide all the action within the computer: the control unit
“fetches” the instructions and the data from the memory section in the proper sequence and
feeds these to the ALU; the ALU performs the actual arithmetic and logic operations in
accordance with the instructions it receives from the control unit and passes the results onto
the output section. The output section converts the results into the corresponding electrical
stimulus needed to activate the peripheral equipment associated with the computer. This
equipment may be a printer, a display terminal, a control mechanism such as a motor or
relay, a modem (for transmission to a remote location over telephone lines), or any other
electronic or electromechanical system.

The memory system represents the brain of the computer. The main memory stores all
the instruction (program) to be run and all the data which are to be manipulated. In a dedi-
cated microcomputer, the memory section is invisible to the user. It consists of an electronic
read-only memory (ROM), which stores the code sequence for the particular program that is to be run. In a general-purpose computer, much of the internal memory is random-access (RAM), often called a read-write memory because it can be programmed and reprogrammed at will. With such computers, the specific program to be run must first be loaded into the RAM from an external source through the input section.

The main-memory system shown in the diagram represents that portion of the system which is an integral part of the microcomputer itself. In a microcomputer dedicated to one specific application, this is all the memory needed. General-purpose computers operate in conjunction with associated mass memory, which is external to the computer. Such memories are in the form of disks, tapes, and cassettes which represent permanent storage for the various programs that the computer may be called upon to run at different times. These programs must be transferred from the external memory to the computer’s internal main memory before they can be used effectively.

The input section converts the output of a signal source into a format acceptable to the microprocessor and memory. The signals from such input peripherals as the keyboard, disk and tape readers, etc., may already be in the proper format. Other potential input systems such as pressure and temperature sensors, modems, and light or presence detectors have output signals that are incompatible with either the computer memory or the MPU. These must first be converted into the proper format by the input section.

28. Computer language. The language of the modern computer is binary. That is, all data and all instructions processed by the computer consist of a series of binary digits (bits) representing either a zero (0) or a one (1).

A zero is normally represented by a low voltage level, while a one is represented by a high voltage level. The actual voltage value representing a 1 or a 0 depends upon the type of logic employed.

The basic computer word consists of eight bits, called a byte (Fig. 6.79). A single byte can represent decimal numbers from 0 to 256 (2 × 10^8); 2 bytes (16 bits) can represent...
numbers up to $65,536 (2 \times 10^{16})$, etc. Similarly, a 2-byte (16-bit) word can encompass over 65,000 memory addresses or program instructions.

Within the computer, information is processed in blocks of parallel bits (bytes or multiple bytes) which are transferred from one section to another along parallel paths called buses. A microcomputer generally has three distinct buses: a data bus, an address bus, and a control bus (Fig. 6.80).

The data bus carries the data words between the ALU and the main memory. It is therefore bidirectional and usually has as many lines as the number of data bits which the ALU is able to process in one operation. Thus, an 8-bit MPU has an 8-bit data bus; a 16-bit machine has a 16-bit data bus, etc. An 8-bit machine could be used to process a 16-bit data work, but it would require two cycles of operation and would be considerably slower than a 16-bit machine.

The address bus is used by the control unit to fetch data and instructions from specific locations in the main memory and transport it to the ALU for processing. The most popular 8-bit machines utilize 16-bit address buses to access in excess of 65,000 memory address locations.

The control bus accesses all portions of the microcomputer. It controls the sequence of events that carries out a particular instruction, activating the read and write lines of the memory, as needed, initiating the transfer of instructions and data to various parts of the CPU in the proper sequence, and generally controlling the entire sequence of program execution.
29. Microcomputer hierarchy. The first major component of the microcomputer era was a PMOS 4-bit microprocessor introduced by Intel Corp. in 1971. This limited-capability chip was soon joined by 8-bit NMOS MPUs from a number of manufacturers. As processing technology improved to permit increased component density, processor power increased first to 16-bit and, most recently, to 32-bit data manipulation on a single chip of silicon. The latter utilize CMOS technology, which, owing to its extremely low power requirements, permits the operation of tens of thousands of transistors in such a small area without demanding extraordinary measures of heat dissipation.

While microprocessor power was expanding from 4 to 32 bits, microcomputer-chip architecture was expanding in other directions as well. Recall that a microprocessor consists of a single chip housing the ALU and control-unit functions. A complete microcomputer requires memory and I/O circuitry as well. Understandably, as increasing component densities facilitated the increase of on-chip componentry, designers began to utilize the excess chip space to place memory and I/O functions on the same chip with the MPU. This resulted in a diversified series of components called microcomputer units (MCUs), which provide virtually complete microcomputer capabilities on a single chip. The repertoire of MCUs is necessarily large because of the many different applications in which microcomputers are employed. These varying applications require different amounts of memory and, in many instances, different I/O functions. Thus, with a basic MPU structure as a core, semiconductor manufacturers are offering MCUs that are tailor-made for different end-use functions by surrounding the MPU with differing complements of on-chip peripheral circuits (Table 6). This capability is becoming so commonplace that some manufacturers are including microprocessor cores and associated peripheral choices as basic cells to be used in the design of semicustom (ASIC) microcomputers.

<table>
<thead>
<tr>
<th>Features</th>
<th>HCMOS MC68HC05</th>
<th>CMOS MC146805</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of pins</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>RAM (bytes)</td>
<td>176</td>
<td>112</td>
</tr>
<tr>
<td>User ROM (bytes)</td>
<td>4160</td>
<td>0</td>
</tr>
<tr>
<td>I/O lines, bidirectional</td>
<td>24</td>
<td>16</td>
</tr>
<tr>
<td>I/O lines, unidirectional</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Timer (bits)</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>Special features</td>
<td>SPI, SCI</td>
<td>Bus expander</td>
</tr>
<tr>
<td>EPROM version</td>
<td>. . .</td>
<td>. . .</td>
</tr>
</tbody>
</table>

NOTE: SPI = serial peripheral interface; SCI = serial communications interface; COP = computer-operating-properly reset timer which acts as a watchdog to reset the CPU automatically if it is not reset by a program sequence within a given time.

30. Product availability. The proliferation of microcomputer applications has spawned a large array of component building blocks with which to implement microcomputer systems. Fundamental, still, for low-level applications are the 8-bit MPUs with their wide assortments of associated memory and I/O chips. These have recently been joined by increasing numbers of 8- and 16-bit MCUs which can be selected for
dedicated applications without the need for additional peripheral chips except, perhaps, for interface circuits to match an MCU to the voltage and current requirements of specific input and output equipment.

For more sophisticated computer applications, 32-bit microprocessors have become commonplace, and 64-bit MPUs are in the early stages of implementation. These circuits are so complex that manufacturers have not yet combined them with memory and I/O peripherals on the same chip, but associated peripheral chips to match the characteristics of the MPUs are being introduced in considerable numbers. These peripherals are often as complex as the MPUs themselves and greatly reduce design time as well as the size of the resulting equipment.

Implementation of microcomputers in a wide range of personal computers as well as dedicated machines is further aided by the availability of increasing numbers of board-level products. These predesigned boards greatly speed up the equipment design cycle. They are being offered not only by the chip manufacturers themselves but also by other manufacturers. Regardless of manufacturer, the boards are designed in families that work together to form complete systems. They offer a choice of basic single-board–microcomputer modules as well as additional peripheral modules that provide a variety of memory capacity and interface choices.

Board families are defined in terms of the bus structure they support; that is, all boards within a given bus family, say, the VMEbus family, have identical pin assignments so that they will be properly interconnected when plugged into a backplane. It is possible, therefore, to purchase VMEbus boards (modules) from different vendors and be assured that they will be compatibly interconnected when plugged into a VMEbus chassis. The large variety of modules available for some bus structures has given rise to the open-system philosophy which permits rapid assembly of complete functional systems with tailor-made capabilities by choosing the appropriate modules and plugging them into the slots (sockets) of a prewired chassis (Fig. 6.81).

FIGURE 6.81  A prewired chassis and board assortment simplifies microcomputer design and assembly.
**ELECTRONIC MEMORIES**

While MCUs normally have on-chip memory sections with a memory capability deemed adequate (by the manufacturer) for the intended application, the MPUs implement their memory sections with peripheral chips. Since all microcomputers require some sort of memory, these devices are the most prevalent of the microprocessor peripherals.

There are many different types of memories. They are classified in a variety of ways, as shown in Table 7. These memories are electronic memories associated with the actual operation of the computer, not with the permanent mass storage of operating programs or large files of data. The mass-storage memories, such as disks, tapes, and cassettes, are separate from the microcomputers and are used only as sources of information to be entered into the computer when required. They are *nonvolatile*, meaning that the information they contain is "permanent" (unless deliberately removed), and they store large amounts of data in a serial fashion. This means that access time is relatively slow.

<table>
<thead>
<tr>
<th>TABLE 7</th>
<th>Memory Descriptors*</th>
</tr>
</thead>
<tbody>
<tr>
<td>By function</td>
<td>RAM, ROM, PROM, EPROM, EEPROM</td>
</tr>
<tr>
<td>By capacity (number of bits)</td>
<td>1K (1000), 16K, 64K, 256K, 1M (1,000,000)</td>
</tr>
<tr>
<td>By organization</td>
<td>4K × 1, 4K × 4, 2K × 8, etc.</td>
</tr>
<tr>
<td>By technology</td>
<td>NMOS, CMOS, bipolar (TTL, ECL)</td>
</tr>
<tr>
<td>By operation</td>
<td>Dynamic, static (RAM only)</td>
</tr>
</tbody>
</table>

*A representative description might be: a 256K × 1-bit NMOS dynamic RAM.

On the other hand, the so-called main memory, or operating memory, within the computer is electronic in nature, and information is stored in a matrix of cells that are accessed in a parallel manner, so that information transfer to and from such memories is relatively fast.

As with logic circuits, memory characteristics are determined by the technology employed. Bipolar ECL and TTL memories are the fastest, NMOS memories have the highest chip density and therefore the highest storage capacity, and CMOS memories have the lowest power dissipation.

31. **Memory organization.** An electronic memory consists of a number of storage units called *cells*. Each cell is a transistor circuit that can be charged to either of two binary states: a high (voltage) level or a low level. A high level is designated as 1, and the low level is designated as 0. In practice, the cells are operated in groups called *words* which store information in binary coded form. The content of a memory word can be the binary code for a number, a letter, or an instruction for manipulating data.

A memory is specified by two numbers that indicate the total number of cells (bits) in the memory and the manner in which these cells are grouped. A 1024 × 1-bit memory (commonly called 1K memory) contains 1024 cells, each of which is separately addressable and programmable. A 128 × 8-bit memory also contains 1024 cells. Each one of these is also individually programmable, but each is accessible only with a group of seven other cells with which it is permanently associated.

32. **Memory capacity.** Memory chips are available in a variety of sizes and cell organizations. Capacity has been increasing rapidly as computer power and processing technology have increased.
Early memories had a capacity of only 128 bits, or 16 bytes. The subsequent 1K-bit memory was considered quite a breakthrough but was quickly replaced by 4K, 16K, and 64K memories. Today, 265K-bit memories are available and 1M (megabit) devices are being introduced. These large-capacity memories are fabricated by NMOS technology, with smaller capacity units being available in CMOS and bipolar processing.

33. Memory functions. Functionally, electronic memories are divided into two basic classifications: random-access memories (RAMs) and read-only memories (ROMs). RAMs are volatile memories whose content disappears when power to the memory is interrupted. ROMs are nonvolatile, in that they retain their content even when power is shut off. Hence, RAMs are used for the actual manipulation of a program or the storing of a program on a temporary basis, while ROMs are employed for storing permanent data such as microprograms, look-up tables, and display graphics.

The addresses of both RAM and ROM are accessible in a random fashion, but RAMs are read-write memories whose contents can be changed at will, whereas ROMs are read-only memories. In general, ROMs are preprogrammed during manufacture with the custom program specified by the customer. There are, however, a number of variations—PROMs, EPROMs, EEPROMs—which are custom-programmable by the user.

Random-Access Memories (RAMs). There are two types of RAMs: static and dynamic. In a static RAM, the individual cells are flip-flops which, once set to either 1 or 0, will retain their setting indefinitely, until either deliberately reset or until power is removed. The basic cell (Fig. 6.82) consists of six transistors. \( T_1 \) and \( T_2 \) represent the actual flip-flop, \( T_3 \) and \( T_4 \) function as load resistors for the flip-flop, and \( T_5 \) and \( T_6 \) are isolation transistors that isolate the cell from the access lines (bit lines) until the cell is ready to be addressed.

The basic cell of a dynamic MOS RAM (Fig. 6.83) utilizes only a single transistor and therefore requires considerably less chip space. Data storage is accomplished by charging the associated capacitor for a 1 level or discharging it for a 0 level. In the diagram, the capacitor

![FIGURE 6.82 A typical six-transistor static MOS memory cell. \( T_1 \) and \( T_2 \) comprise a flip-flop, with \( T_3 \) and \( T_4 \) representing load resistors (note the fixed bias on these transistors). \( T_5 \) and \( T_6 \) are isolation transistors which, when turned on by a row-select signal, cause the state of the flip-flop to be reflected on the bit (data) lines.](image-url)
is charged when both X and Y lines are made high and discharged when the X line is high and the Y line is low. Unfortunately, the small capacitor will not retain its charge very long. Therefore, dynamic RAMs must be refreshed (the charged cells must be recharged) periodically (say, every millisecond) in order to maintain their information. The refreshing circuitry adds substantial complexity to the circuit. On the other hand, dynamic RAMs dissipate a relatively small amount of power. For this reason and because of the smaller space requirements, they are preferred for memories with a large number of cells (large capacity).

Read-Only Memories (ROMs, PROMs, EPROMs, EEPROMs)

ROMs. Read-only memories are an integral part of all microcomputers. They store the operating routines of the system (microprograms) and provide conversion data, display codes, and other information that is periodically accessed by the computer but is not subject to change.

A ROM consists basically of a series of OR gates addressed by a decoder circuit (Fig. 6.84). The truth table associated with the diagram indicates the state of the four output

**FIGURE 6.83** A single-transistor dynamic RAM cell. Additional circuitry is needed to refresh all cells in the memory on a regular, periodic schedule.

**FIGURE 6.84** (a) Basic configuration of a typical read-only memory. (b) Truth table corresponding to the simple circuit.
lines, \( b_0 \) to \( b_3 \), as a function of the input signals \( A \) and \( B \). If, for example, the input signals are both HIGH, then output lines \( b_1 \) and \( b_2 \) are HIGH because the output of AND gate 11 in the decoder is HIGH. Similarly, if \( B \) is HIGH and \( A \) is LOW, output lines \( b_0, b_4, \) and \( b_3 \) will be HIGH because the output of decoder gate 10 is HIGH. Output \( b_2 \) will be HIGH only when the output of NAND gate 01 is HIGH, which occurs only if \( A \) is HIGH and \( B \) is LOW. Thus, for any combination of input signals, the output can be determined by the number and placement of the diodes that comprise the OR gates.

The simple circuit described is a 4 × 4 ROM consisting of 4 words, with 4 bits per word. If the number of output lines were extended to 8, the resultant ROM would be 4 × 8. An increase in the capacity of the decoder will provide ROMs with thousands of different 8-bit output combinations.

ROMs are available as factory mask-programmable devices which are fabricated by the manufacturer to any set of custom patterns specified by the user. They can also be obtained as standard preprogrammed devices containing various popular functions such as character generators, code converters, look-up tables, etc.

**PROMs.** The ROM of Fig. 6.84 can be converted into a unit that is user-programmable (PROM) simply by placing diodes in all possible locations within the matrix and fabricating a fusible link in series with each diode (Fig. 6.85). The user can burn out any link, using a PROM programmer, in order to create any special set of codes desired. Of course, this can be done only once, so that this type of PROM is sometimes called a one-time programmable ROM.

**Erasable Programmable ROMs (EPROMs).** The ideal memory is a nonvolatile one which maintains its program even after power is removed from the system (as with a standard mask-programmed or one-time field-programmed PROM discussed above), yet one whose program can be altered at will (as with a standard RAM). Development efforts in that direction are bearing fruit, and several iterations of programmable ROMs have been introduced. These do not yet have the reprogramming speed and ease of the conventional RAM, but they are enjoying considerable popularity for a number of applications, and they are indicative of the rapidly expanding technology.

One of the earliest practical ROMs that was repeatedly erasable and reprogrammable in the field was labeled EPROM. The information stored in the memory cells could be erased by subjecting the device to ultraviolet light, which entered the package through a quartz window at the top (Fig. 6.86). While the EPROM can be erased and reprogrammed repeatedly, it must be removed from the system and bulk-erased by exposing the entire unit to high-intensity ultraviolet light source. There are a number of suitable ultraviolet programmers, but the erase time is on the order of half an hour, which falls far short of meeting the criteria for the ideal memory. Moreover, the EPROM can be unintentionally erased by sunlight and fluorescent light and requires special shielding after programming. Despite these limitations, the UV EPROM still enjoys considerable use, and its principles of operation are fundamental to even the more advanced devices currently emerging.
The EPROM is basically an MOS device whose structure is illustrated in Fig. 6.87. It is characterized by two stacked polysilicon gates which are separated from each other and from the substrate of the cell by an insulating layer of silicon dioxide. The gate closest to the substrate is a floating gate which is not connected in the circuit. The upper gate is the control gate, which is used to energize the cell. Under the initial condition, the cell acts as a conventional MOS transistor so that when a positive voltage is applied to the control gate and the drain, a channel is created between source and drain, permitting a drain-current flow.

If the drain voltage and the control-gate voltage are raised to a high level (say, 30 V), some of the electrons in the channel gain enough energy to cross the silicon dioxide barrier and impinge upon the floating gate, where they become trapped. This raises the turn-on threshold level of the MOS transistor so that it will not be turned on when normal operating voltages are applied. The cell then is said to be charged, representing an open circuit and yielding a high output level (1) when connected in a conventional circuit.

The charge on the floating gate can be removed by the application of light to the structure. Sunlight or fluorescent light will erase the charge but would require an excessive amount of time. A strong dose of ultraviolet light will bulk-erase the entire memory in about 30 min.

Electrically Erasable Programmable ROMs (EEPROMs). The EEPROM is very similar in theory and structure to the EPROM. The principal difference is in the thickness of the oxide between the floating gate and the substrate. Whereas the thickness of this oxide layer for the EPROM is on the order of 1000 Å, that of the EEPROM is on the order of 100 Å or less. As a result of this thinner oxide, the floating gate can be charged $and$ discharged by a voltage applied to the control gate. Write and erase times for typical EEPROMs with present technology are on the order of 10 ms. While the EEPROM is not competitive with RAM speeds, its in-circuit program-erase capability makes it suitable for a great many applications. Devices of this type are currently available with 64K-bit capacity. They are also included as on-chip peripheral circuits for a number of MCUs to expand MCU versatility.
34. Accessing the memory. The total number of cells in a single memory chip can range from several hundred on the low end to 1 million (1 megabit) with current state-of-the-art technology. The cells are arranged in rows and columns (Fig. 6.88), with the exact location of each cell specified by its row and column number. There are two basic organizational schemes for accessing the memory: bit organization and word organization.

Bit Organization. In a bit-organized memory chip (e.g., 1K × 1, 64K × 1, etc.), each cell is addressed in conjunction with similarly located cells on other chips of the same capacity and organization. In Fig. 6.89, for example, eight identical chips are used to store information in byte-sized (8-bit) words. Each chip stores one of the 8 bits, and all chips are addressed simultaneously for programming or reading.

Word Organization. A word-organized memory chip is one in which a number of columns corresponding to the number of bits in the word are addressed simultaneously. In this manner, a complete word can be stored in a single chip (Fig. 6.90).
35. **Memory circuitry.** A memory array receives a number of signals from the microprocessor. The primary ones are as follows:

1. The address signal which activates the rows and columns on each chip that represent the cells associated with a given address
2. The binary word representing the data to be entered into the addressed cells in the case of a programming step
3. A read-write signal generated by the processor which determines whether information is to be entered into the memory or extracted from it

The circuitry required to respond to these signals and generate the desired results resides directly on the memory chip along with the array of memory cells. Typical operation can be described with the aid of diagrams utilized in a representative MOS static memory chip.

**The Memory Cell.** Figure 6.91 is a conventional MOS flip-flop in which $T_1$ and $T_2$ are cross-coupled inverters and $T_3$ and $T_4$ represent their respective load resistors. A positive pulse applied to point $A$ appears at the gate of $T_1$, turning this transistor on. As a result, the voltage at point $B$ goes to 0 (ground), and since this voltage is applied to the gate of $T_2$, the latter is turned off. Voltage at $A$ therefore

**Figure 6.90** A 1K (128 × 8) byte-organized chip addresses all bits associated with a complete word on a single chip.

**Figure 6.91** A typical MOS static memory cell.
goes high and remains high even when the external signal is removed. In this condition, the cell is considered to store a 1.

To reprogram the cell to contain a 0, a positive pulse is applied to point B, turning \( T_1 \) on and \( T_2 \) off. Again, this condition remains stable until it is deliberately changed by another positive signal to point A. Clearly, the cell is programmable.

**Row Address.** In practice, a cell is programmed by means of a two-rail bit-line system (Fig. 6.92), in which a signal and its complement are simultaneously applied to points \( A \) and \( B \). To be applied to the cell, however, isolation transistors \( T_5 \) and \( T_6 \) must be turned on. Programming is accomplished by applying the desired programming signals on the bit lines and then momentarily turning on \( T_5 \) and \( T_6 \) with a pulse applied simultaneously to their gates. Once programmed, a cell is read by briefly turning \( T_5 \) and \( T_6 \) on without having an external signal on the bit lines.

![Figure 6.92](image)

**Column Address.** In Fig. 6.93, the column-select circuitry and a write buffer have been added to Fig. 6.90. Note that the write buffer generates the programming signal and its complement, to be applied to the bit lines. To access the bit lines, however, transistors \( T_7 \) and \( T_8 \) must be turned on. The address code defines which of the columns is to be activated. In Fig. 6.93, each of the columns can be separately addressed, as required for a single-bit organization. In Fig. 6.90, all eight of the column-select circuits are activated simultaneously, as required for a byte-organized orientation.

**Read-Write Circuit.** In Fig. 6.94, a read buffer circuit has been added to the previous diagram, permitting the data on the bit lines to be applied to the data bus for transmittal back to the MPU. Note, however, that the data line that carries the data from the memory to the MPU is the same as the line that carries the programming data from the MPU to the memory. To avoid undesirable feedback, the read buffer must be disabled during a write operation. This is accomplished with a three-state read buffer circuit as in Fig. 6.94b.

**Three-State Buffer.** A three-state buffer is a circuit whose output can be high, low, or open-circuited. A high output at bit line 1 turns \( T_4 \) on and \( T_1 \) off. If \( T_2 \) and \( T_3 \) are turned on by a separate high-level signal on their gates, then the output line is connected to the positive-voltage source through \( T_3 \) and \( T_4 \), and the output is high. A high output at bit line 1 turns \( T_1 \), turns off \( T_4 \), and presents a low signal to the output. But if \( T_2 \) and \( T_3 \) are turned off, then regardless of the state of \( T_1 \) and \( T_4 \) the output is isolated from the inputs because of the very high resistance of \( T_2 \) and \( T_3 \).
FIGURE 6.93 Column-address circuitry. Data can be entered into the memory only when transistors \( T_7 \) and \( T_8 \) are turned on.

FIGURE 6.94 (a) Connection of read buffer to memory column. (b) Three-state read-buffer schematic diagram.
$T_2$ and $T_3$ are turned off during writing operations and turned on at all other times. Thus, the cell is always connected to the data line through the read buffer except during a write cycle, during which the read buffer is in the open-circuit state.

### 36. Address decoding.

The number of unique bit patterns required to address all rows in a single-column, word-organized memory equals $2^n$, where $n$ is the number of rows. Thus, a small $128 \times 8$-bit memory, for example, requires $128$, or $2^{7}$, unique address patterns, which can be generated with a 7-bit binary code. These 128 unique address patterns can be applied to a single-address decoder that has 128 output lines. For each address, the decoder provides a high-level signal on a different output line. There are many different decoder circuit configurations. The circuit shown in Fig. 6.95 is for illustration and does not necessarily represent any specific design.

**Decoder Circuit.** The decoder circuit consists of 128 seven-input NOR gates having a common driver circuit. Each NOR gate is activated (puts out a high-level signal) only when all seven inputs are low. The driver circuit, in turn, is configured so that each of the 128 possible input patterns activates a different NOR gate.

![Decoder Circuit Diagram](image_url)
Each NOR gate consists of seven parallel-connected transistors (switches) in series with a single fixed-bias transistor (resistor). When all seven parallel-connected transistors are turned off by a low-level input to each, the output is high and the row of cells associated with that gate is activated.

To provide the required all-zero signals to the gate to be activated, the 7-bit energizing signal is first sent through double-inverter driver stages, one for each bit. The output of each driver then makes available both the signal (A, B, C, . . .) and its complement (Ā, Ā, Ā, . . .).

For example, if gate 0 (row 0) is to be activated by the first address pattern (0000000), then gate-0 input terminals are connected to outputs (A, B, C, D, E, F, G) of the driver. To activate gate 1 by address No. 2 (0000001), gate-1 terminals are connected to driver outputs Ā, Ā, Ā, Ā, Ā, Ā, Ā. The last gate in the system (No. 127) has all its inputs connected to the inverted outputs of the driver. It is possible, therefore, to connect the driver outputs to each gate in such a manner that each gate receives an all-zero input for a different combination of 1s and 0s at the decoder input.

Most memories, instead of being arranged in one continuous sequence of rows, are arranged physically and electrically in different blocks. A $128 \times 8$-bit memory, for example, could contain four blocks, each with 32 rows and 8 columns. Here, the 8-bit data signal could be applied via a data bus to all eight columns of each block simultaneously. However, this arrangement needs separate decoding circuits to activate not only the desired row but the desired eight-column block as well. Accordingly, a 2-bit block-select decoder (Fig. 6.96) is needed for four-pattern ($2^2$) selectivity. However, since the row-select decoder need now decode only 32 lines ($2^5$) rather than 128, a 5-bit decoder is sufficient for row selection. Thus, again, a 7-bit decoder handles the entire job.

![FIGURE 6.96 A row-and-block–select decoder.](image-url)
The only difference between this circuit and the previous one is that four gates of the decoder are used to select the blocks while the row-select function is handled by 32 five-input NOR gates. The same binary address pattern that accesses a memory address in a one-block system can access the same address in a four-block organization and other organizations as well.

NOTE Much of the information provided in this division is based on the work of technical and marketing specialists of the Motorola Semiconductor Sector. The author gratefully acknowledges the kind permission of Motorola for the use of material from its various data sheets, application notes, and other documents.

PRODUCT RELIABILITY

37. Semiconductor reliability. In today’s marketplace, quality and reliability are paramount to the success of a product. Quality is defined here as the ability of a device to operate within its specifications when received by the user; reliability refers to its capability to operate within specifications over long periods of time and under the most adverse environmental conditions to be encountered in practice.

Semiconductors are inherently extremely reliable. Properly designed and used, they can be expected to outlive the equipment in which they are employed. Yet they are subject to a wide range of failure mechanisms that must be anticipated and circumvented during manufacture in order to achieve this inherent feature.

To ensure high reliability, the government has developed a stringent series of tests and screens involving all phases of manufacturing for military applications. The manufacturing techniques required to meet these qualifications have been adopted in some measure by manufacturers of products intended for commercial applications so that, except for very special requirements, there are no longer significant variations in inherent reliability between commercial and military products.

Basic Military Qualifications Tests. The Basic MIL-qualification tests, called JAN (Joint Army-Navy), are divided into groups to satisfy three objectives:

1. Group A—verifies that form, fit, and function conform to design specifications.
2. Group B—assures manufacturing integrity and verifies reliability in ground support applications.
3. Group C—provides evidence of long-term reliability under harsh environmental conditions where severe mechanical and environmental stresses exist.

Examples of the types of tests included in each group are given below.

Group A Tests

Visual/mechanical. Consists of a sampling of devices that is examined to determine whether they meet the applicable materials, design, construction, marking, and workmanship standards.

DC tests. Verify the major voltage, current, and other dc parameters of the device first under normal (25°C) temperature conditions, then at an appropriate high and/or low temperature limit to confirm satisfactory performance over the entire temperature range.

AC tests. Check parameters associated with dynamic operating conditions, such as capacitances, noise figure, switching time, etc.

Safe operating area (SOA) tests. Limited to power transistors, these tests corroborate the power limits over which the devices are designed to operate.
Current surge. Applies principally to diodes, e.g., rectifiers.

Selected tests. Assigned to specific devices for unique specifications that do not fit a general specification.

Thus, the complete spectrum of electrical device characteristics is accounted for, and successful completion of this test sequence provides assurance that the devices are capable of operating, at least initially, in accordance with their design.

Group B Tests. This sequence of tests includes screens that are intended to verify that the devices are mechanically sound and that they can be expected to operate satisfactorily over time and under adverse operating conditions. Since a number of these screens involve stress factors that could result in ultimate performance degradation, the electrical parameters expected to be affected are tested before and after the applied screen to ascertain that the performance change remains within prescribed limits.

Solderability. Determines the solderability of wires, lugs, tabs, and all types of terminals that are normally joined by a soldering operation. The procedure includes an accelerated aging test that simulates a minimum of 2 to 10 years natural aging under a combination of various storage conditions that have different deleterious effects.

Resistance to solvents. Verifies that the markings will not become illegible when the component is subject to various solvents employed by equipment manufacturers.

Thermal shock. Simulates transferring equipment from a heated shelter to the outside environment in an arctic area. While temperature limits may vary with specified conditions, the test usually involves exposure of the components to temperatures of $-55^\circ C$ and $+125^\circ C$, or the maximum rated specification of the component.

Surge current test. Subjects devices under test to high forward current stress conditions to determine the ability of the chip and the contacts to withstand current surges.

Hermetic seal. Determines the effectiveness of the seal of semiconductors with internal cavities. There are two types of required tests—fine leak and gross leak.

Steady-state operation life. Designed to weed out components that might be subject to “infant mortality” on the proven premise that if a semiconductor device is likely to fail, it will do so during the early hours of normal operation.

Intermittent operation life. Similar to the steady-state operation life test, above, except that the devices are subject to sudden sequential on-off periods to provide added stress.

Blocking life. Performed on rectifier diodes only, this test is normally run at an elevated temperature of $150^\circ C$ for 340 h, with the primary blocking junction reverse-biased. The reverse-bias voltage is 80 to 85 percent of the rated voltage of the device.

Decap internal visual. Verifies conformance to the original specifications by “opening” samples of a completed lot and comparing the internal structure with the qualified design report.

Scanning electron microscope (SEM) inspection. Verifies the quality of the metalization on the semiconductor die by checking for defects such as voids, separations, notches, cracks, depressions, or tunnels and other processing-related faults.

Bond strength. Verifies the integrity of wire (or clip) bonds within the package by applying stresses that will cause the maximum potential for wire breaks and associated die failures.

Thermal resistance tests. Determine the efficiency of the chip-to-header interface in transferring heat from the chip to the header.

High-temperature (nonoperating) life. Verifies that the devices will not degrade after storage at maximum temperature for 340 h.

Group C Tests. Whereas previous test groups deal with the design and manufacturing integrity of semiconductors, this group of tests adds additional assurances that the products will withstand severe environmental stresses over prolonged time periods.
Thermal shock (glass strain). Exposes the devices to sudden extreme changes in temperature by alternately immersing them in hot and cold liquids ranging from 0 to 100°C and, subsequently, testing them for hermeticity and electrical parameters to determine if any failures occurred.

Terminal strength. Checks the capabilities of the device leads, welds, and seals for their ability to withstand pulls and bends and other physical stresses.

Moisture resistance. An accelerated test method of evaluating the resistance of the device to the deteriorative effects of high humidity and heat typical of tropical conditions.

Mechanical shock. Consists of mounting the devices in a fixture that is subsequently dropped repeatedly from a predetermined height to produce a high-impact force on the various mechanical parts and bonds.

Vibration. Simulates the use of devices in the field, when mounted on a jeep, truck, or airplane, in order to confirm the integrity of the package and the wire bonds within the package.

Constant acceleration. Designed to detect structural and mechanical weaknesses not detected in other shock and vibration tests.

Salt atmosphere. An accelerated laboratory corrosion test simulating the effect of shipboard or seacoast atmospheres on devices by subjecting them to a controlled salt atmosphere fog stream to detect subsequent flaking, pitting, or corrosion that will interfere with device application.

The above tests are required for all standard military (MIL-Qualified) semiconductors and are described in detail in documents MIL-S-19500 and MIL-STD-750 maintained by the Department of the Navy. In addition, many of them are employed routinely during the manufacturing process of commercial products as a safeguard against processing faults and to provide a continual check of end-product quality and reliability.

38. Semiconductor packages. Semiconductor devices are available in a seemingly endless variety of packages. Some of these variations are cost-related, others are function-related, and still others are preference-related.

Cost-related selections normally include metal and ceramic packages on the high end of the scale, and plastic packages on the low end. Plastic packages are dominant in commercial equipment, while metal and ceramic packages are usually mandated for military and high-reliability applications, or where power requirements are beyond the capabilities of plastic packages.

For this discussion, function-related packages relate to sizes and shapes dictated by the functions contained on the chip. Thus, discrete components may utilize from two to four pins, while integrated circuit packages may require hundreds of pins to accommodate the I/O requirements of the enclosed function.

Preference-related packages are those that offer the same or similar functions in different package configurations to accommodate design preferences. Typical choices include such variations as leaded or surface-mount packages, in-line or dual in-line pin configurations or special structures such as press-fit and stud packages.

Integrated circuits, typically, are housed in dual in-line packages (Fig. 6.97a) to accommodate a large number of leads in an acceptable amount of space. Large-scale integrated circuits, whose I/O requirements cannot easily be satisfied even with dual in-line structures are accommodated by special VLSI packages such as pin grid arrays (PGAs) and leaded chip carriers (LCCs) (Fig. 6.97b). And for equipment where available space is at a premium, equivalent space-saving small outline integrated circuit (SOIC) packages for surface-mount applications (Fig. 6.97c) are gaining in popularity and availability. While integrated circuits for military and hi-rel applications are usually made of ceramic, a few metal packages with relatively high pin counts are also available (Fig. 6.97d).
The selection of discrete component packages is equally prolific. Higher-power devices come in standard JEDEC-registered (TO-numbered) metal and plastic housings (Fig. 6.98a) and a wide variety of package variations carrying a manufacturer’s proprietary case numbers. Often, these offer similar specifications in a number of different package choices. Small-signal components are equally endowed. Typical examples of JEDEC-registered three-leaded transistors and two-terminal diodes are shown in Fig. 6.98b. Packages housing multiple transistors and diodes take the form of conventional dual in-line integrated circuits, with the number of leads depending on the number of devices contained.

**Mounting Considerations for Power Packages**. Heat is a principal enemy of semiconductor devices. Except for lead-mounted parts used at low currents, a heat exchange is required to prevent junction temperatures from exceeding their rated limits, thereby risking device failure. Proper mounting of power devices, during manufacture and eventual substitution or replacement, is a prerequisite for reliable operation. This necessitates attention to the following areas:

1. Mounting surface preparation
2. Application of thermal compounds
3. Electrical insulation
4. Assembly fastening
5. Lead bending and soldering

Intimate thermal contact between the package and the heat sink is essential for proper heat dissipation. Rough or scratched heat sink surfaces, excessively large mounting holes, and oxidized or unclean surfaces all can have deleterious effects on heat transfer, thereby contributing to potential device degradation in high-power circuits. Significantly improved contact can be achieved through the use of thermal joint compounds (“grease”), which fill air voids between all mating surfaces. Satisfactory joint compounds can reduce the thermal resistivity of such voids from 1200°C·in/W to approximately 60°C·in/W. This compares much more favorably to the 0.10°C·in/W for copper film.

Since power semiconductors normally have their collectors or anodes connected to the case, many applications require that the case be electrically insulated from circuit ground. This can be accomplished most effectively by insulating the entire heat sink–semiconductor assembly from ground rather than using an insulator between the semiconductor package and the heat sink. Where this is not possible, in cases where the chassis serves as the heat

---

*Abstracted from Motorola Application Note AN1040.*
FIGURE 6.98  Representative sampling of packages used for discrete devices.
sink, or where the heat sink is common to several devices, insulating (mica or silicone rubber) washers can be used between the package and the heat sink.

When electrical insulation is required prior to heat sinking, the need for a thermal grease becomes more pronounced since there are now two isolating surfaces between the package and the heat sink. Data obtained by Thermalloy, Inc. (Fig. 6.99) show relative thermal resistance values between different types of insulating materials and also for various mounting screw torque values. It is quite evident that the use of grease is much more effective in reducing thermal resistance than an increase in the amount of force used to mate the two surfaces. Indeed, excessive tightening of the mounting screw can cause potential problems, particularly if the mounting hole is too large, by pulling the package into the hole (Fig. 6.100). The resulting buckling not only reduces surface contact, but can also cause cracking of the die in the package. Die rupture can also occur when the leads of a device must be bent in order to fit into a fixture or socket.

Typical examples of mounting several types of semiconductor packages are illustrated in Fig. 6.101.

![Interface thermal resistance for various types of insulating materials as a function of mounting screw torque—with and without the use of grease.](image-url)
FIGURE 6.100 Exaggerated example of improperly mounted semiconductor package.

FIGURE 6.101 Typical methods for mounting various types of power device packages.
To reduce insulation problems, some of the more recent power devices are being produced in insulated packages. These newer packages, housing multiple-chip and monolithic integrated power circuits as well as discrete components, come in two basic forms. The first (Fig. 6.102a) uses a mounting plate that is insulated from the chip and can be fastened directly to a grounded chassis or heat sink. The second (Fig. 6.102b) utilizes a thermal conductive plastic overmold to cover the metal mounting base of the basic package.

39. Transient suppression*. A not-infrequent cause for semiconductor problems is the incidence of transients, which can stress the devices beyond their electrical design limits. Some transients are internally generated as a result of inductive switching, commutation voltage spikes, etc. These are easily suppressed because their energy content is known and predictable. Others, however, may be created outside the circuit and coupled into it. These are more difficult to anticipate, and their suppression is often beyond the control of the circuit designer. Such transients include lightning-generated spikes and substation problems, but can also be caused by switching of parallel loads within the same branch of a power distribution system. Effective transient suppression requires that the impulse energy within the transient be dissipated in an added transient suppressor at a low enough voltage so that the capabilities of the circuit or associated components will not be exceeded.

Several types of transient suppressors have been in widespread use. Among these are the carbon block spark gaps, gas tubes, selenium rectifiers, metal oxide varactors, and zener diodes. Of these, zener diodes, especially those designed specifically for surge suppression, e.g., the 1N6267–1N6303 series, have the most favorable set of characteristics.

There are two characteristics of primary importance for surge suppressors. The first, the zener breakdown voltage rating, must be equal to or lower than the maximum voltage of the equipment or circuits it is to protect. The second is the maximum power rating of the suppressor, beyond which it may itself be damaged. The latter must take into account both the maximum peak voltage and current ratings of the device, as well as the voltage excursions and pulse width of the transients.

*Abstracted from Motorola Application Note AN843.
Typical waveshapes for indoor and outdoor locations (Fig. 6.103) are defined in IEEE standard 28, ANSI standard C62.1, and can be considered as realistic representations. Peak ac power line voltage transients in indoor locations are usually limited to approximately 6.0 kV due to the spark-over spacing between conductors used in standard wiring practices. In outdoor locations, such as electrical service entrances, power line connections to additional buildings, etc., no such limitations exist.

The amplitude of randomly induced voltage transients and their energy content are difficult to define, but data from surge counters and other sources (Fig. 6.104) have provided some insight. In this plot, data for low exposure were taken from locations with little load switching or lightning activity; medium exposure is considered as areas with severe switching transients and frequent lightning activity; high-exposure systems are those supplied by long unprotected overhead lines with high spark-over clearances. Surge characteristics deemed representative of various locations within a building are given in Table 8.

FIGURE 6.103 Typical indoor (a) and outdoor (b) transient waveshapes.
FIGURE 6.104 Peak surge voltages versus surges per year. [EIA Paper, 587.1/F. May 1979, p. 10]

TABLE 8 Surge Voltages and Currents Anticipated at Various Locations

<table>
<thead>
<tr>
<th>Location</th>
<th>Surge voltage</th>
<th>Surge current</th>
<th>Energy (joules*) dissipated in suppressor with clamp voltage of</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.0 kV</td>
<td>200 A</td>
<td>0.4 0.8 1.6</td>
</tr>
<tr>
<td>2</td>
<td>6.0 kV</td>
<td>500 A</td>
<td>1.0 2.0 4.0</td>
</tr>
<tr>
<td>3</td>
<td>10 kV or more</td>
<td>10 kV or more</td>
<td>20 40 80</td>
</tr>
</tbody>
</table>

*Joules = power × time.
1. Outlets and circuits a long distance from electrical service entrance
2. Major bus lines and circuits a short distance from service entrance
3. Electrical service entrance and outdoor locations

A safe approximation of the peak power contained in a transient can be obtained by considering its shape to be that of a rectangular pulse with the same peak power. For example, an exponential discharge with a time constant of 1.0 ms can be approximated with a rectangular 1.0-ms pulse with the same peak power as the transient.

Because pulse width determines the heating effect of the transient, the peak power dissipation ratings of typical zener diodes are also a function of pulse width, as shown in Fig. 6.105a. And when the transient is expected to be repetitive, peak power must be further derated as in Fig. 6.105b.

![Figure 6.105a](image)

**Figure 6.105a**  (a) Peak power ratings of typical zener diodes as a function of pulse width.  (b) Power derating as a function of duty cycle.