1 Device Physics

Before 1960, most electronic circuits depended upon vacuum tubes to perform the critical tasks of amplification and rectification. An ordinary mass-produced AM radio required five tubes, while a color television needed no fewer than twenty. Vacuum tubes were large, fragile, and expensive. They dissipated a lot of heat and were not very reliable. So long as electronics depended upon them, it was nearly impossible to construct systems requiring thousands or millions of active devices.

The appearance of the bipolar junction transistor in 1947 marked the beginning of the solid-state revolution. These new devices were small, cheap, rugged, and reliable. Solid-state circuitry made possible the development of pocket transistor radios and hearing aids, quartz watches and touch-tone phones, compact disc players and personal computers.

A *solid-state device* consists of a crystal with regions of impurities incorporated into its surface. These impurities modify the electrical properties of the crystal, allowing it to amplify or modulate electrical signals. A working knowledge of device physics is necessary to understand how this occurs. This chapter covers not only elementary device physics but also the operation of three of the most important solid-state devices: the junction diode, the bipolar transistor, and the field-effect transistor. Chapter 2 explains the manufacturing processes used to construct these and other solid-state devices.

1.1 SEMICONDUCTORS

The inside front cover of the book depicts a long-form periodic table. The elements are arranged so those with similar properties group together to form rows and columns. The elements on the left-hand side of the periodic table are called *metals*, while those on the right-hand side are called *nonmetals*. Metals are usually good conductors of heat and electricity. They are also malleable and display a characteristic metallic luster. Nonmetals are poor conductors of heat and electricity, and those that are solid are brittle and lack the shiny luster of metals. A few elements in the middle

of the periodic table, such as silicon and germanium, have electrical properties that lie midway between those of metals and nonmetals. These elements are called *semiconductors*. The differences between metals, semiconductors, and nonmetals result from the electronic structure of their respective atoms.

Every atom consists of a positively charged nucleus surrounded by a cloud of electrons. The number of electrons in this cloud equals the number of protons in the nucleus, which also equals the atomic number of the element. Therefore a carbon atom has six electrons because carbon has an atomic number of six. These electrons occupy a series of *shells* that are somewhat analogous to the layers of an onion. As electrons are added, the shells fill in order from innermost outward. The outermost or *valence* shell may remain unfilled. The electrons occupying this outermost shell are called *valence electrons*. The number of valence electrons possessed by an element determines most of its chemical and electronic properties.

Each row of the periodic table corresponds to the filling of one shell. The leftmost element in the row has one valence electron, while the rightmost element has a full valence shell. Atoms with filled valence shells possess a particularly favored configuration. Those with unfilled valence shells will trade or share electrons so that each can claim a full shell. Electrostatic attraction forms a chemical bond between atoms that trade or share electrons. Depending upon the strategy adopted to fill the valence shell, one of three types of bonding will occur.

Metallic bonding occurs between atoms of metallic elements, such as sodium. Consider a group of sodium atoms in close proximity. Each atom has one valence electron orbiting around a filled inner shell. Imagine that the sodium atoms all discard their valence electrons. The discarded electrons are still attracted to the positively charged sodium atoms, but, since each atom now has a full valence shell, none accepts them. Figure 1.1A shows a simplified representation of a sodium crystal. Electrostatic forces hold the sodium atoms in a regular lattice. The discarded valence electrons wander freely through the resulting crystal. Sodium metal is an excellent electrical conductor due to the presence of numerous free electrons.[1] These same electrons are also responsible for the metallic luster of the element and its high thermal conductivity. Other metals form similar crystal structures, all of which are held together by metallic bonding between a sea of free valence electrons and a rigid lattice of charged atomic cores.

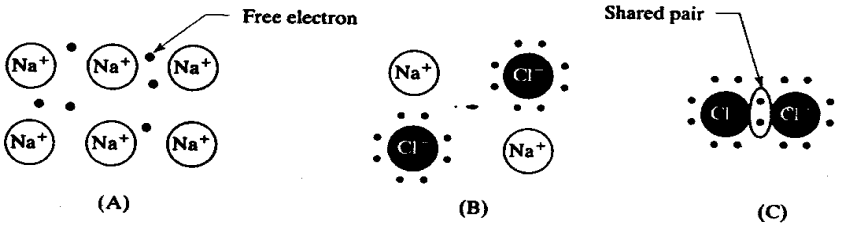

FIGURE 1.1 Simplified illustrations of various types of chemical bonding: metallically bonded sodium crystal (A), ionically bonded sodium chloride crystal (B), and covalently bonded chlorine molecule (C).

Ionic bonding occurs between atoms of metals and nonmetals. Consider a sodium atom in close proximity to a chlorine atom. The sodium atom has one valence electron, while the chlorine atom is one electron short of a full valence shell. The sodium atom can donate an electron to the chlorine atom and by this means both can achieve filled outer shells. After the exchange, the sodium atom has a net posi-

[1] Some metals conduct by means of holes rather than electrons, but the general observations made in the text still apply.

tive charge and the chlorine atom a net negative charge. The two charged atoms (or *ions*) attract one another. Solid sodium chloride thus consists of sodium and chlorine ions arranged in a regular lattice, forming a crystal (Figure 1.1B). Crystalline sodium chloride is a poor conductor of electricity, since all of its electrons are held in the shells of the various atoms.

Covalent bonding occurs between atoms of nonmetals. Consider two chlorine atoms in close proximity. Each atom has only seven valence electrons, while each needs eight to fill its valence shell. Suppose that each of the two atoms contributes one valence electron to a common pair shared by both. Now each chlorine atom can claim eight valence electrons: six of its own, plus the two shared electrons. The two chlorine atoms link to form a molecule that is held together by the electron pair shared between them (Figure 1.1C). The shared pair of electrons forms a *covalent bond.* The lack of free valence electrons explains why nonmetallic elements do not conduct electricity and why they lack metallic luster. Many nonmetals are gases at room temperature because the electrically neutral molecules exhibit no strong attraction to one another and thus do not condense to form a liquid or a solid.

The atoms of a semiconductor also form covalent bonds. Consider atoms of silicon, a representative semiconductor. Each atom has four valence electrons and needs four more to complete its valence shell. Two silicon atoms could theoretically attempt to pool their valence electrons to achieve filled shells. In practice this does not occur because eight electrons packed tightly together strongly repel one another. Instead, each silicon atom shares one electron pair with each of four surrounding atoms. In this way, the valence electrons are spread around to four separate locations and their mutual repulsion is minimized.

Figure 1.2 shows a simplified representation of a silicon crystal. Each of the small circles represents a silicon atom. Each of the lines between the circles represents a covalent bond consisting of a shared pair of valence electrons. Each silicon atom can claim eight electrons (four shared electron pairs), so all of the atoms have full valence shells. These atoms are linked together in a molecular network by the covalent bonds formed between them. This infinite lattice represents the structure of the silicon crystal. The entire crystal is literally a single molecule, so crystalline silicon is strong and hard, and it melts at a very high temperature. Silicon is normally a poor conductor of electricity because all of its valence electrons are used to form the crystal lattice.

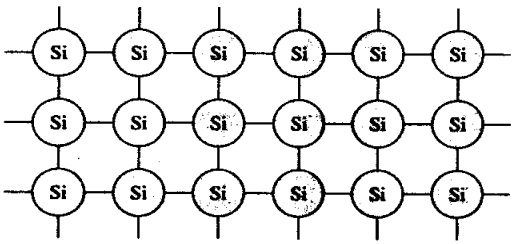

FIGURE 1.2 Simplified two-dimensional representation of a silicon crystal lattice.

A similar macromolecular crystal can theoretically be formed by any group-IV element,[2] including carbon, silicon, germanium, tin, and lead. Carbon, in the form of diamond, has the strongest bonds of any group-IV element. Diamond crystals

[2] The group-III, IV, V, and VI elements reside in columns III-B, IV-B, V-B, and VI-B of the long-form periodic table. The group-II elements may fall into either columns II-A or II-B. The A/B numbering system is a historical curiosity and the International Union of Pure and Applied Chemists (IUPAC) has recommended its abandonment; see J. Hudson, *The History of Chemistry* (New York: Chapman and Hall, 1992), pp. 122–137.

are justly famed for their strength and hardness. Silicon and germanium have somewhat weaker bonds due to the presence of filled inner shells that partially shield the valence electrons from the nucleus. Tin and lead have weak bonds because of numerous inner shells; they typically form metallically bonded crystals instead of covalently bonded macromolecules. Of the group-IV elements, only silicon and germanium have bonds of an intermediate degree of strength. These two act as true semiconductors, while carbon is a nonmetal, and tin and lead are both metals.

1.1.1. Generation and Recombination

The electrical conductivity of group-IV elements increases with atomic number. Carbon, in the form of diamond, is a true insulator. Silicon and germanium have much higher conductivities, but these are still far less than those of metals such as tin and lead. Because of their intermediate conductivities, silicon and germanium are termed *semiconductors*.

Conduction implies the presence of free electrons. At least a few of the valence electrons of a semiconductor must somehow escape the lattice to support conduction. Experiments do indeed detect small but measurable concentrations of free electrons in pure silicon and germanium. The presence of these free electrons implies that some mechanism provides the energy needed to break the covalent bonds. The statistical theory of thermodynamics suggests that the source of this energy lies in the random thermal vibrations that agitate the crystal lattice. Even though the average thermal energy of an electron is relatively small (less than 0.1 electron volt), these energies are randomly distributed, and a few electrons possess much larger energies. The energy required to free a valence electron from the crystal lattice is called the *bandgap energy*. A material with a large bandgap energy possesses strong covalent bonds and therefore contains few free electrons. Materials with lower bandgap energies contain more free electrons and possess correspondingly greater conductivities (Table 1.1).

TABLE 1.1 Selected properties of group-IV elements.[3]

Element	Atomic Number	Melting Point, °C	Electrical Conductivity $(\Omega cm)^{-1}$	Bandgap Energy, eV
Carbon (diamond)	6	3550	$\sim 10^{-16}$	5.2
Silicon	14	1410	$4 \cdot 10^{-6}$	1.1
Germanium	32	937	0.02	0.7
White Tin	50	232	$9 \cdot 10^{4}$	0.1

A vacancy occurs whenever an electron leaves the lattice. One of the atoms that formerly possessed a full outer shell now lacks a valence electron and therefore has a net positive charge. This situation is depicted in a simplified fashion in Figure 1.3. The ionized atom can regain a full valence shell if it appropriates an electron from a neighboring atom. This is easily accomplished since it still shares electrons with three adjacent atoms. The electron vacancy is not eliminated; it merely shifts to the

[3] Bandgap energies for Si, Ge: B. G. Streetman, *Solid State Electronic Devices*, 2nd ed. (Englewood Cliffs, NJ: Prentice-Hall, 1980), p. 443. Bandgap for C: N. B. Hanny, ed., *Semiconductors* (New York: Reinhold Publishing, 1959), p. 52. Conductivity for Sn: R. C. Weast, ed., *CRC Handbook of Chemistry and Physics*, 62nd ed. (Boca Raton, FL: CRC Press, 1981), pp. F135–F136. Other values computed. Melting points: Weast, pp. B4–B48.

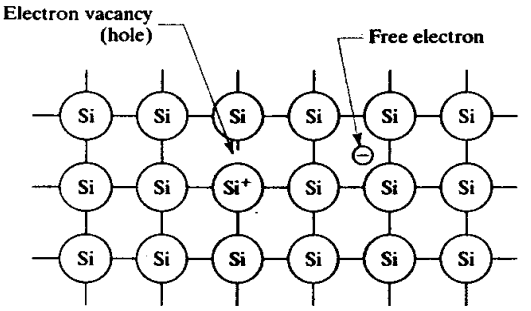

FIGURE 1.3 Simplified diagram of thermal generation in intrinsic silicon.

adjacent atom. As the vacancy is handed from atom to atom, it moves through the lattice. This moving electron vacancy is called a *hole.*

Suppose an electric field is placed across the crystal. The negatively charged free electrons move toward the positive end of the crystal. The holes behave as if they were positively charged particles and move toward the negative end of the crystal. The motion of the holes can be compared to bubbles in a liquid. Just as a bubble is a location devoid of fluid, a hole is a location devoid of valence electrons. Bubbles move upward because the fluid around them sinks downward. Holes shift toward the negative end of the crystal because the surrounding electrons shift toward the positive end.

Holes are usually treated as if they were actual subatomic particles. The movement of a hole toward the negative end of the crystal is explained by assuming that holes are positively charged. Similarly, their rate of movement through the crystal is measured by a quantity called *mobility.* Holes have lower mobilities than electrons; typical values in bulk silicon are $480 cm^2/V \cdot sec$ for holes and $1350 cm^2/V \cdot sec$ for electrons.[4] The lower mobility of holes makes them less efficient charge carriers. The behavior of a device therefore relies upon whether its operation involves holes or electrons.

A free electron and a hole are formed whenever a valence electron is removed from the lattice. Both particles are electrically charged and move under the influence of electric fields. Electrons move toward positive potentials, producing an electron current. Holes move toward negative potentials, producing a hole current. The total current equals the sum of the electron and the hole currents. Holes and electrons are both called *carriers* because of their role in transporting electric charge.

Carriers are always generated in pairs since the removal of a valence electron from the lattice simultaneously forms a hole. The generation of electron-hole pairs can occur whenever energy is absorbed by the lattice. Thermal vibration produces carriers, as do light, nuclear radiation, electron bombardment, rapid heating, mechanical friction, and any number of other processes. To consider only one example, light of a sufficiently short wavelength can generate electron-hole pairs. When a lattice atom absorbs a photon, the resulting energy transfer can break a covalent bond to produce a free electron and a free hole. Optical generation will occur only if the photons have enough energy to break bonds, and this in turn requires light of a sufficiently short wavelength. Visible light has enough energy to produce electron-hole pairs in most semiconductors. Solar cells make use of this phenomenon to convert sunlight into electrical current. Photocells and solid-state camera detectors also employ optical generation.

[4] Streetman, p. 443.

Just as carriers are generated in pairs, they also recombine in pairs. The exact mechanism of carrier recombination depends on the nature of the semiconductor. Recombination is particularly simple in the case of a *direct-bandgap semiconductor.* When an electron and a hole collide, the electron falls into the hole and repairs the broken covalent bond. The energy gained by the electron is radiated away as a photon (Figure 1.4A). Direct-bandgap semiconductors can, when properly stimulated, emit light. A *light-emitting diode* (LED) produces light by electron-hole recombination. The color of light emitted by the LED depends on the bandgap energy of the semiconductor used to manufacture it. Similarly, the so-called *phosphors* used in manufacturing glow-in-the-dark paints and plastics also contain direct-bandgap semiconductors. Electron-hole pairs form whenever the phosphor is exposed to light. A large number of electrons and holes gradually accumulate in the phosphor. The slow recombination of these carriers causes the emission of light.

Silicon and germanium are *indirect-bandgap semiconductors.* In these semiconductors, the collision of a hole and an electron will not cause the two carriers to recombine. The electron may momentarily fall into the hole, but quantum mechanical considerations prevent the generation of a photon. Since the electron cannot shed excess energy, it is quickly ejected from the lattice and the electron-hole pair reforms. In the case of an indirect-bandgap semiconductor, recombination can only occur at specific sites in the lattice, called *traps,* where flaws or foreign atoms distort the lattice (Figure 1.4B). A trap can momentarily capture a passing carrier. The trapped carrier becomes vulnerable to recombination because the trap can absorb the liberated energy.

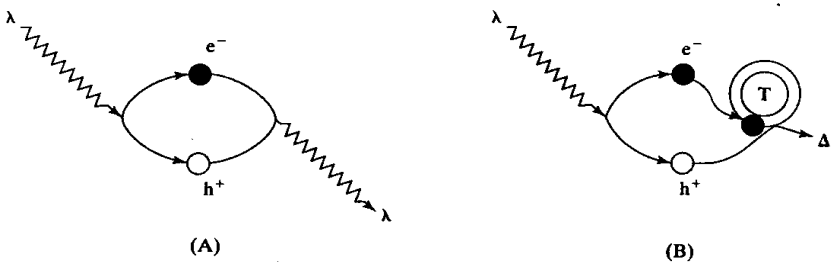

FIGURE 1.4 Schematic representations of recombination processes: (A) direct recombination, in which a photon, λ, generates a hole, h^+, and an electron, e^-, that collide and re-emit a photon; and (B) indirect recombination, in which one of the carriers is caught by a trap, T, and recombination takes place at the trap site with the liberation of heat, Δ.

Traps that aid the recombination of carriers are called *recombination centers.* The more recombination centers a semiconductor contains, the shorter the average time between the generation of a carrier and its recombination. This quantity, called the *carrier lifetime,* limits how rapidly a semiconductor device can switch on and off. Recombination centers are sometimes deliberately added to semiconductors to increase switching speeds. Gold atoms form highly efficient recombination centers in silicon, so high-speed diodes and transistors are sometimes made from silicon containing a small amount of gold. Gold is not the only substance that can form recombination centers. Many transition metals such as iron and nickel have a similar (if less potent) effect. Some types of crystal defects can also serve as recombination centers. Solid-state devices must be fabricated from extremely pure single-crystal materials in order to ensure consistent electrical performance.

1.1.2. Extrinsic Semiconductors

The conductivity of semiconductors depends upon their purity. Absolutely pure, or *intrinsic,* semiconductors have low conductivities because they contain only a few thermally generated carriers. The addition of certain impurities greatly increases the

number of available carriers. These *doped,* or *extrinsic,* semiconductors can approach the conductivity of a metal. A lightly doped semiconductor may contain only a few parts per billion of dopant. Even a heavily doped semiconductor contains only a few hundred parts per million due to the limited solid solubility of dopants in silicon. The extreme sensitivity of semiconductors to the presence of dopants makes it nearly impossible to manufacture truly intrinsic material. Practical semiconductor devices are, therefore, fabricated almost exclusively from extrinsic material.

Phosphorus-doped silicon is an example of an extrinsic semiconductor. Suppose a small quantity of phosphorus is added to a silicon crystal. The phosphorus atoms are incorporated into the crystal lattice in positions that would otherwise have been occupied by silicon atoms (Figure 1.5). Phosphorus, a group-V element, has five valence electrons. The phosphorus atom shares four of these with its four neighboring atoms. Four bonding electron pairs give the phosphorus atom a total of eight shared electrons. These, combined with the one remaining unshared electron, result in a total of nine valence electrons. Since eight electrons entirely fill the valence shell, no room remains for the ninth electron. This electron is expelled from the phosphorus atom and wanders freely through the crystal lattice. Each phosphorus atom added to the silicon lattice thus generates one free electron.

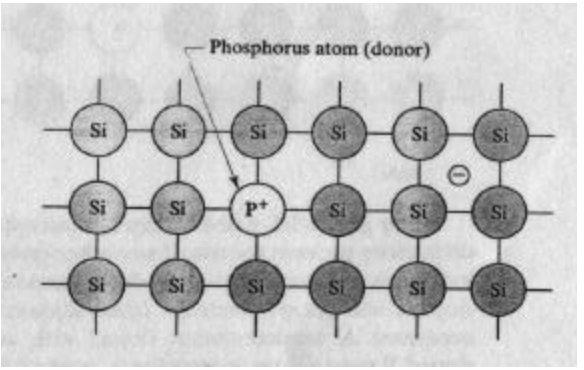

FIGURE 1.5 Simplified crystal structure of phosphorus-doped silicon.

The loss of the ninth electron leaves the phosphorus atom with a net positive charge. Although this atom is ionized, it does not constitute a hole. Holes are electron vacancies created by the removal of electrons from a filled valence shell. The phosphorus atom has a full valence shell despite its positive charge. The charge associated with the ionized phosphorus atom is therefore immobile.

Other group-V elements will have the same effect as phosphorus. Each atom of a group-V element that is added to the lattice will produce one additional free electron. Elements that donate electrons to a semiconductor in this manner are called *donors.* Arsenic, antimony, and phosphorus are all used in semiconductor processing as donors for silicon.

A semiconductor doped with a large number of donors has a preponderance of electrons as carriers. A few thermally generated holes still exist, but their numbers actually diminish in the presence of extra electrons. This occurs because the extra electrons increase the probability that the hole will find an electron and recombine. The large number of free electrons in N-type silicon greatly increases its conductivity (and greatly reduces its resistance).

A semiconductor doped with donors is said to be *N-type.* Heavily doped N-type silicon is sometimes marked N+, and lightly doped N-type silicon N–. The plus and minus symbols denote the relative numbers of donors, not electrical charges. Electrons are considered the *majority carriers* in N-type silicon due to their large

numbers. Similarly, holes are considered the *minority carriers* in N-type silicon. Strictly speaking, intrinsic silicon has neither majority nor minority carriers because both types are present in equal numbers.

Boron-doped silicon forms another type of extrinsic semiconductor. Suppose a small number of boron atoms are added to the silicon lattice (Figure 1.6). Boron, a group-III element, has three valence electrons. The boron atom attempts to share its valence electrons with its four neighboring atoms, but, because it has only three, it cannot complete the fourth bond. As a result, there are only seven valence electrons around the boron atom. The electron vacancy thus formed constitutes a hole. This hole is mobile and soon moves away from the boron atom. Once the hole departs, the boron atom is left with a negative charge caused by the presence of an extra electron in its valence shell. As in the case of phosphorus, this charge is immobile and does not contribute to conduction. Each atom of boron added to the silicon contributes one mobile hole.

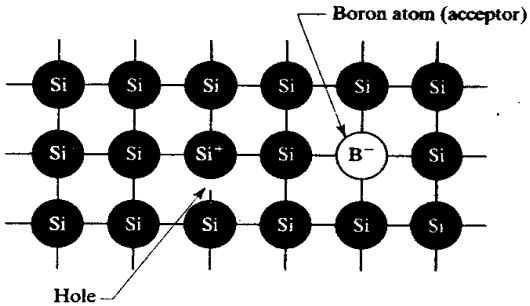

FIGURE 1.6 Simplified crystal structure of boron-doped silicon.

Other group-III elements can also accept electrons and generate holes. Technical difficulties prevent the use of any other group-III elements in silicon fabrication, but indium is sometimes used to dope germanium. Any group-III element used as a dopant will *accept* electrons from adjoining atoms, so these elements are called *acceptors*. A semiconductor doped with acceptors is said to be *P-type*. Heavily doped P-type silicon is sometimes marked P+, and lightly doped P-type silicon P–. Holes are the majority carriers and electrons are the minority carriers in P-type silicon. Table 1.2 summarizes some of the terminology used to describe extrinsic semiconductors.

TABLE 1.2 Extrinsic semiconductor terminology.

Semiconductor Type	Dopant Type	Typical Dopants for Silicon	Majority Carriers	Minority Carriers
N-type	Donors	Phosphorus, arsenic, and antimony	Electrons	Holes
P-type	Acceptors	Boron	Holes	Electrons

A semiconductor can be doped with both acceptors and donors. The dopant present in excess determines the type of the silicon and the concentration of the carriers. It is thus possible to invert P-type silicon to N-type by adding an excess of donors. Similarly, it is possible to invert N-type silicon to P-type by adding an excess of acceptors. The deliberate addition of an opposite-polarity dopant to invert the type of a semiconductor is called *counterdoping*. Most modern semiconductors are

made by selectively counterdoping silicon to form a series of P- and N-type regions. Much more will be said about this practice in the next chapter.

If counterdoping were taken to extremes, the entire crystal lattice would consist of an equal ratio of acceptor and donor atoms. The two types of atoms would be present in exactly equal numbers. The resulting crystal would have very few free carriers and would appear to be an intrinsic semiconductor. Such *compound semiconductors* actually exist. The most familiar example is *gallium arsenide*, a compound of gallium (a group-III element) and arsenic (a group-V element). Materials of this sort are called III-V compound semiconductors. They include not only gallium arsenide but also gallium phosphide, indium antimonide, and many others. Many III-V compounds are direct-bandgap semiconductors, and some are used in constructing light-emitting diodes and semiconductor lasers. Gallium arsenide is also employed to a limited extent for manufacturing very high-speed solid-state devices, including integrated circuits. II-VI compound semiconductors are composed of equal mixtures of group-II and group-VI elements. Cadmium sulfide is a typical II-VI compound used to construct photosensors. Other II-VI compounds are used as phosphors in cathode ray tubes. A final class of semiconductors includes IV-IV compounds such as silicon carbide, recently used on a small scale to fabricate blue LEDs.

Of all the semiconductors, only silicon possesses the physical properties required for high-volume, low-cost manufacture of integrated circuits. The vast majority of solid-state devices are fabricated in silicon, and all other semiconductors are relegated to niche markets. The remainder of this text, therefore, focuses upon silicon integrated circuits.

1.1.3. Diffusion and Drift

The motion of carriers through a silicon crystal results from two separate processes: *diffusion* and *drift*. Diffusion is a random motion of carriers that occurs at all times and places, while drift is a unidirectional movement of carriers under the influence of an electric field. Both of these processes contribute to conduction in semiconductors.

Diffusion closely resembles Brownian motion. That is, individual carriers move through the semiconductor until they collide with lattice atoms. The collision process scatters the carriers through unpredictable angles. After a very few collisions, the motion of the carriers becomes completely randomized. The carriers wander aimlessly about, tracing a sort of drunkard's walk (Figure 1.7A).

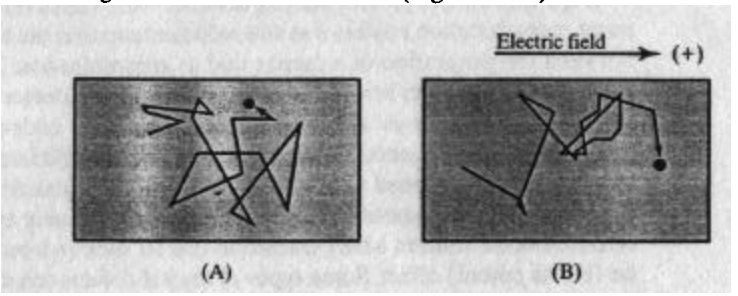

FIGURE 1.7 Comparison of conduction mechanisms for an electron: diffusion (A) and drift superimposed on diffusion (B). Notice the gradual motion of the electron toward the positive potential.

The diffusion of carriers through a semiconductor is analogous to the diffusion of dye molecules through still water. When a drop of concentrated dye falls into water, the dye molecules all initially occupy a small volume of liquid. The molecules gradually diffuse from regions of higher concentration to regions of lower concentration. Eventually the dye becomes distributed uniformly throughout the solution. Similarly, the diffusion of carriers across concentration gradients produces a *diffusion current*.

Unless some mechanism constantly adds more carriers, diffusion eventually redistributes them uniformly throughout the silicon and the diffusion current subsides.

The motion of a carrier under the influence of an electric field is called *drift*. Although the carrier still collides with the lattice and thus moves in a random drunkard's walk, it gradually drifts in a specific direction (Figure 1.7B). This subtle bias is caused by the electric field. No matter what direction the carrier moves, the field always and relentlessly acts upon it. If the carrier is moving opposite the field, its motion is retarded; if it is moving with the field, its motion is accelerated. Frequent collisions prevent the carrier from building up any appreciable velocity, but a subtle overall motion appears. Electrons move toward positive potentials, even if slowly and erratically. Holes likewise move toward negative potentials. A simple analogy to drift consists of the motion of the steel ball in a pinball machine. Although bumpers and pegs may divert the ball in any direction, the tilt of the board causes it to eventually move downward. Similarly, an electric field biases carriers toward motion in a specific direction, producing a *drift current*.

1.2 PN JUNCTIONS

Uniformly doped semiconductors have few applications. Almost all solid-state devices contain a combination of multiple P- and N-type regions. The interface between a P-type region and an N-type region is called a *PN junction*, or simply a *junction*.

Figure 1.8A shows two pieces of silicon. On the left is a bar of P-type silicon, and on the right is a bar of N-type silicon. No junction is present as long as the two are not in contact with one another. Each piece of silicon contains a uniform distribution of carriers. The P-type silicon has a large majority of holes and a few electrons; the N-type silicon has a large majority of electrons and a few holes.

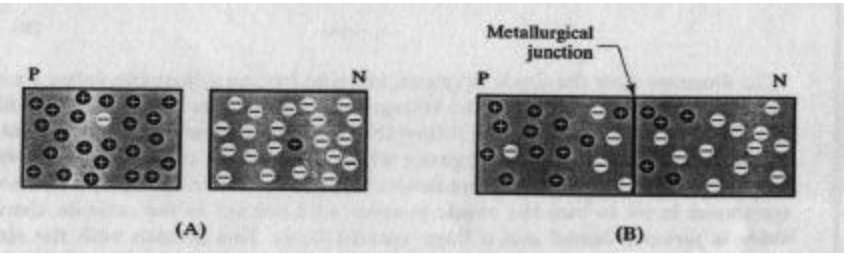

FIGURE 1.8 Carrier populations in silicon before the junction is assembled (A), and afterward (B).

Now, imagine the two pieces of silicon are brought into contact with one another to form a junction. No physical barrier to the motion of the carriers remains. There is a great excess of holes in the P-type silicon, and a great excess of electrons in the N-type silicon. Some of the holes diffuse from the P-type silicon to the N-type. Likewise, some of the electrons diffuse from the N-type silicon to the P-type. Figure 1.8B shows the result. A number of carriers have diffused across the junction in both directions. The concentration of minority carriers on either side of the junction has risen above that which would be produced by doping alone. The excess of minority carriers produced by diffusion across a junction is called the *excess minority carrier concentration*.

1.2.1. Depletion Regions

The presence of excess minority carriers on either side of a junction has two effects. First, the carriers produce an electric field. The extra holes in the N-type silicon represent a positive charge, while the excess electrons in the P-type silicon represent a

negative charge. Thus an electric potential develops across the PN junction that biases the N-side of the junction positive with respect to the P-side.

When carriers diffuse across the junction, they leave equal numbers of ionized dopant atoms behind. These atoms are rigidly fixed in the crystal lattice and cannot move. On the P-side of the junction lie ionized acceptors that produce a negative charge. On the N-side of the junction lie ionized donors that produce a positive charge. An electric potential again develops that biases the N-side of the junction positive with respect to the P-side. This potential adds to the one produced by the separation of the charged carriers.

Carriers tend to drift in the presence of an electric field. Holes are attracted to the negative potential on the P-side of the junction. Similarly, electrons are attracted to the positive potential on the N-side of the junction. The drift of carriers thus tends to oppose their diffusion. Holes diffuse from the P-side of the junction to the N-side and drift back. Electrons diffuse from the N-side of the junction to the P-side and drift back. Equilibrium occurs when the drift and diffusion currents are equal and opposite. The excess minority carrier concentrations on either side of the junction also reach equilibrium values, as does the voltage potential across the junction.

The voltage difference across a PN junction in equilibrium is called its *built-in potential,* or its *contact potential.* In a typical silicon PN junction, the built-in potential can range from a few tenths of a volt to as much as a volt. Heavily doped junctions have larger built-in potentials than lightly doped ones. Because of the higher doping levels, more carriers diffuse across the heavily doped junction and thus a larger diffusion current flows. In order to restore equilibrium, a larger drift current is also needed and thus a stronger electric field develops. Heavily doped junctions therefore have larger built-in potentials than lightly doped ones.

Although the built-in potential is quite real, it cannot be measured with a voltmeter. This apparent paradox can be explained by a closer examination of a circuit containing a PN junction and a voltmeter (Figure 1.9). The two probes of the meter are made of metal, not silicon. The points of contact between the metal probes and the silicon also form junctions, each of which has a contact potential of its own. Because the silicon beneath the two probes has different doping levels, the two contact potentials of the probe points are unequal. The difference between these two contact potentials exactly cancels the built-in potential of the PN junction, and no current flows in the external circuit. This situation must occur because any current flow would constitute a free energy source, or a sort of perpetual motion machine. The cancellation of the built-in potentials ensures that energy cannot be extracted from a PN junction in equilibrium and thus prevents a violation of the laws of thermodynamics.

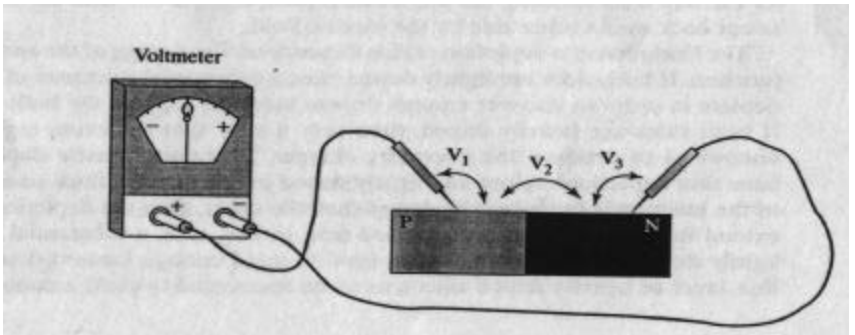

FIGURE 1.9 Demonstration of the impossibility of directly measuring a built-in potential. Contact potentials V_1 and V_3 exactly cancel built-in potential V_2.

The built-in potential has two causes: the separation of ionized dopant atoms and the separation of charged carriers. The carriers are free to move, but the dopant atoms are rigidly fixed in the crystal lattice. If the dopant atoms could move, they would be drawn together by their opposite charges. They remain separated because they are anchored to the lattice. The region occupied by these charged atoms is subject to a strong electric field. Any carrier that enters this region must move quickly or it will be swept out again by the field. As a result, this region contains very few carriers at any given instant in time. This region is sometimes called a *space charge layer* because of the presence of the charged dopant atoms. More commonly, it is called a *depletion region* because of the relatively low concentration of carriers found there.

If the depletion region contains few carriers, then the excess minority carriers must pile up on either side of it. Figure 1.10 graphically shows the resulting distributions of excess minority carriers. The concentration gradients cause these carriers to diffuse into the electrically neutral regions beyond the junction. The electric field produced by the separation of these charged carriers pulls them back toward the junction. An equilibrium is soon established, resulting in steady-state distributions of minority carriers resembling those shown in Figure 1.10.

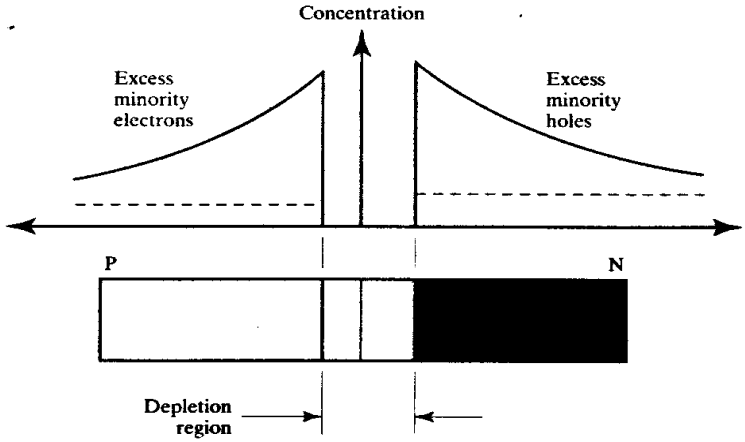

FIGURE 1.10 Diagram of excess minority carrier concentrations on either side of a PN junction in equilibrium.

The behavior of a PN junction can be summarized as follows: the diffusion of carriers across the junction produces excess minority carrier concentrations on either side of a depletion region. The separation of ionized dopant atoms causes an electric field to form across the depletion region. This field prevents most of the majority carriers from crossing the depletion region, and the few that do are eventually swept back to the other side by the electric field.

The thickness of a depletion region depends on the doping of the two sides of the junction. If both sides are lightly doped, then a substantial thickness of silicon must deplete in order to uncover enough dopant atoms to support the built-in potential. If both sides are heavily doped, then only a very thin depletion region need be uncovered to produce the necessary charges. Therefore heavily doped junctions have thin depletion regions and lightly doped junctions have thick ones. If one side of the junction is more heavily doped than the other, then the depletion region will extend further into the lightly doped side. In this case, a substantial thickness of lightly doped silicon must be uncovered to yield enough ionized dopants. Only a thin layer of heavily doped silicon need be uncovered to yield a counterbalancing

charge. Figure 1.10 illustrates this case since the N-side of the junction is more lightly doped than the P-side.

1.2.2. PN Diodes

A PN junction forms a very useful solid-state device called a *diode.* Figure 1.11 shows a simplified diagram of the structure of a PN diode. The diode has, as its name suggests, two terminals. One terminal, called the *anode,* connects to the P-side of the junction. The other terminal, called the *cathode,* connects to the N-side. These two terminals are used to connect the diode to an electrical circuit. The schematic symbol for a diode consists of an arrowhead representing the anode and a perpendicular line representing the cathode. Diodes conduct current preferentially in one direction—that indicated by the arrowhead.

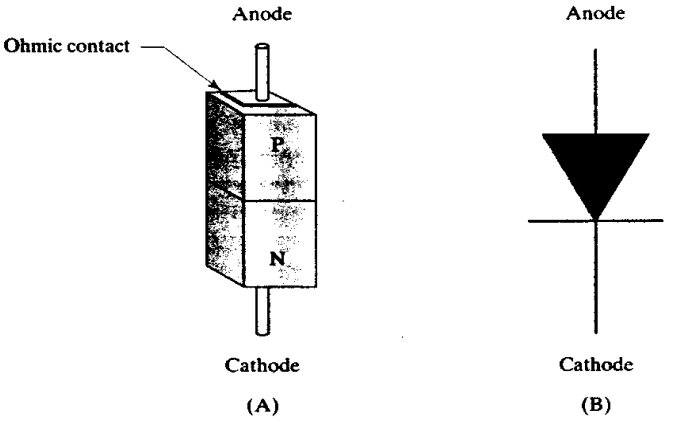

FIGURE 1.11 PN junction diode: simplified structure (A) and standard schematic symbol (B).

To illustrate how the diode operates, imagine that an adjustable voltage source has been connected across it. If the voltage source is set to zero volts, then the diode is under *zero bias.* No current will flow through a zero-biased diode. If the voltage source is set to bias the anode negative with respect to the cathode, then the diode is *reverse biased.* Very little current flows through a reverse-biased diode. If the voltage source is set to bias the anode positive with respect to the cathode, then the diode is *forward biased* and a large current flows. This accords with the simple mnemonic: *current flows with the arrow, not against it.* Devices that conduct in only one direction are called *rectifiers.* They find frequent application in power supplies, radio receivers, and signal processing circuits.

Diode rectification depends upon the presence of a junction. Each of the three bias conditions can be explained by an appropriate analysis of carrier flows across this junction. The case of the zero-bias diode is particularly simple since it is identical to the case of the equilibrium junction already discussed. The only potential present across the junction is the built-in potential. When the diode is connected in a circuit, the contact potentials of the leads touching the silicon balance the built-in potential of the junction. Thus no current flows in the circuit.

The behavior of a reverse-biased diode is also simple to explain. The reverse bias makes the N-side of the junction even more positive with respect to the P-side. The voltage seen across the junction increases, so excess minority carriers continue to be swept back across it and majority carriers continue to be held on their respective sides of the junction. The increased voltage across the junction causes the ionization

of additional dopant atoms on either side, so the depletion region widens as the reverse bias increases.

The behavior of a forward-biased junction is somewhat more complex. The voltage applied to the terminals opposes the built-in potential. The voltage across the junction therefore lessens and the depletion region thins. The drift currents caused by the electric field are simultaneously reduced. More and more majority carriers make the transit across the depletion region without being swept back by the electric field. Figure 1.12 shows graphically the overall flow of carriers: holes are injected across the junction from anode to cathode (left to right), while electrons are injected across the junction from cathode to anode (right to left). In the illustrated diode, the hole current across the junction outweighs the electron current because the anode is more heavily doped than the cathode and there are more majority holes available in the anode than there are majority electrons in the cathode. Once these carriers have been injected across the junction, they become minority carriers and recombine with majority carriers present on the other side. Currents are drawn in from the terminals in order to replenish the supply of majority carriers in the neutral silicon. This illustration is somewhat simplified since it only shows the general flow of carriers through the diode. Some of the carriers injected across the junction are swept back by the electric field before they can recombine. Such carriers do not contribute to the net current flow through the diode, so they are not illustrated. Likewise, the tiny numbers of thermally generated minority carriers that cross the junction are not shown since they form an insignificant portion of the overall current flow through a forward-biased diode.

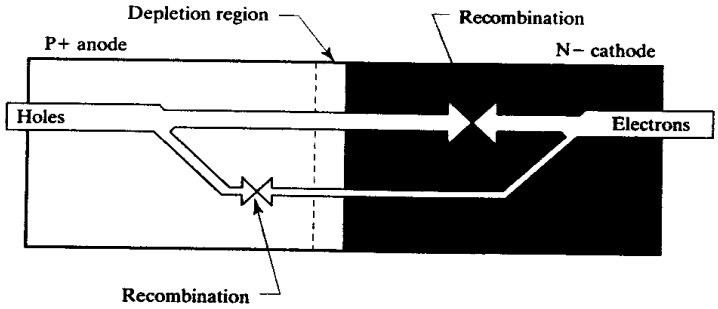

FIGURE 1.12 Carrier flow in a forward-biased PN junction.

The current through a forward-biased diode depends exponentially upon the applied voltage (Figure 1.13). About 0.6V suffices to produce substantial forward conduction in a silicon PN junction at room temperature.[5] Because diffusion is caused by the thermal motions of carriers, higher temperatures cause an exponential increase in diffusion currents. The forward current through a PN junction thus increases exponentially as temperature increases. Expressed another way, the forward bias required to sustain a constant current in a silicon PN junction decreases by approximately 2mV/°C.

Figure 1.13 also shows a low level of current flow when the diode is reverse-biased. This current flow is called *reverse conduction* or *leakage*. Leakage currents are produced by the few minority carriers thermally generated in the silicon. The electric field opposes the flow of majority carriers across a reverse-biased junction,

[5] The most widely quoted value is 0.7V, but in practice a typical integrated circuit base-emitter junction under microamp-level bias at 25°C exhibits a value nearer 0.6V than 0.7V.

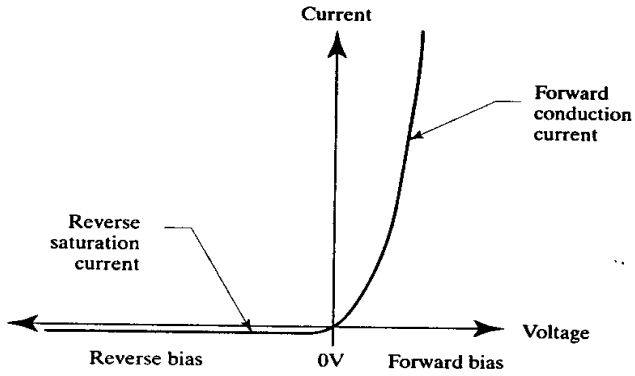

FIGURE 1.13 Diode conduction characteristics. The current scale is greatly magnified to show the reverse saturation current, which typically equals no more than a few picoamps at 25°C.

but it aids the flow of minority carriers. The application of a reverse bias sweeps these minority carriers across the junction. Because the rate of generation of minority carriers in the bulk silicon is essentially independent of electric fields, the leakage current does not vary much with reverse bias. Thermal generation does increase with temperature, and leakage currents are therefore temperature-dependent. In silicon, leakage currents double approximately every eight degrees Celsius. At high temperatures, the leakage currents begin to approach the operating currents of the circuit. The maximum operating temperature of a semiconductor device is therefore limited by leakage current. A maximum junction temperature of 150°C is widely accepted for silicon-integrated circuits.[6]

1.2.3. Schottky Diodes

Rectifying junctions can also form between a semiconductor and a metal. Such junctions are called *Schottky barriers*. The behavior of a Schottky barrier is somewhat analogous to that of a PN junction. Schottky barriers can, for example, be used to construct *Schottky diodes*, which behave much like PN diodes. Schottky barriers can also form in the contact regions of an integrated circuit's interconnection system.

The *work function* of a material equals the amount of energy required to remove an electron from it. Each material has its own characteristic work function that depends upon the properties of its crystal lattice as well as its composition. When two materials with different work functions are brought into contact, the electrons in each material have different initial energies. A voltage difference called the *contact potential* therefore exists between the two materials. Consider the case of a PN junction. The two semiconductors on either side of the junction have the same lattice structure. The contact potential of a PN junction, or its *built-in potential*, depends only upon doping. In the case of a Schottky barrier, the different lattice structures of the metal and the semiconductor also contribute to the contact potential.

A typical rectifying Schottky barrier results when aluminum metal touches lightly doped N-type silicon (Figure 1.14B). The carriers must redistribute in order to counterbalance the contact potential. Electrons diffuse from the semiconductor into the metal, where they pile up to form a thin film of negative charge. This exodus of electrons from the silicon leaves behind a zone of ionized dopant atoms that form a

[6] Integrated circuits can be built that work at 200°C, but many standard design practices do not apply. See R. J. Widlar and M. Yamatake, "Dynamic Safe-Area Protection for Power Transistors Employs Peak-Temperature Limiting." *IEEE J. Solid-State Circuits*, SC-22, #1, 1987, p. 77–84.

depletion region (Figure 1.14A). The electric field generated by the depletion region draws electrons from the metal back into the semiconductor. Equilibrium occurs when the drift and diffusion currents are equal. The potential difference across the Schottky barrier now equals the contact potential. Few minority carriers exist on the semiconductor side of the Schottky barrier, so the Schottky diode is called a *majority-carrier device.*

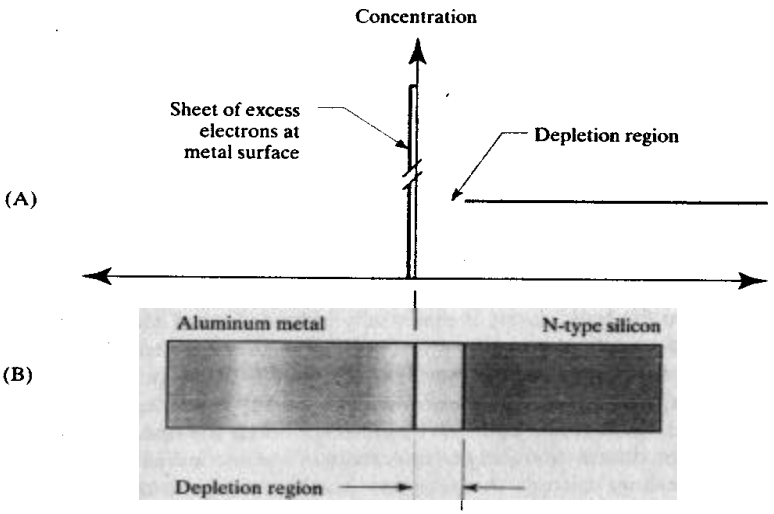

FIGURE 1.14 Diagram of excess carrier concentrations on either side of the Schottky barrier (A) and cross-section of the corresponding Schottky structure (B).

The behavior of a Schottky diode under bias can be similarly analyzed. The N-type silicon forms the *cathode* of the diode, while the metal plate forms the *anode.* The case of a zero-biased Schottky diode is identical to the case of the equilibrium Schottky barrier analyzed above. A reverse-biased Schottky has an external voltage connected in order to bias the semiconductor positively with respect to the metal. The resulting voltage difference adds to the contact potential. The depletion region widens to counterbalance the increased voltage difference, equilibrium is restored, and very little current flows through the diode.

A forward-biased Schottky diode has an external voltage connected in order to bias the metal positively with respect to the semiconductor. The resulting voltage difference across the junction opposes the contact potential, and the width of the depletion region shrinks. Eventually the contact potential is entirely offset, and a depletion region attempts to form on the metal side of the junction. The metal, being a conductor, cannot support an electric field, and no depletion region can form to oppose the externally applied potential. This potential begins to sweep electrons across the junction from the semiconductor into the metal, and a current flows through the diode.

Schottky diodes exhibit current-voltage characteristics similar to those of a PN diode (Figure 1.13). Schottky diodes also exhibit leakage currents caused by low levels of minority carrier injection from the metal into the semiconductor. These conduction mechanisms are accelerated by high temperatures, producing temperature dependencies similar to those of a PN diode.

Despite many apparent similarities, there are a few fundamental differences between Schottky diodes and PN junction diodes. Schottky diodes are majority-carrier devices since they rely primarily upon majority-carrier conduction. At high current densities a few holes do flow from the metal to the semiconductor, but

these contribute only a small fraction of the total current. Schottky diodes do not support large excess minority-carrier populations. Since the switching speed of a diode is a function of the time required for excess minority carriers to recombine, Schottky diodes can switch very rapidly. Some types of Schottky diodes also exhibit lower forward bias voltages than PN diodes. This combination of low forward-voltage drop and highly efficient switching make Schottky diodes very useful devices.

Schottky diodes can also be formed to P-type silicon, but the forward biases required for conduction are usually quite low. This renders P-type Schottky diodes rather leaky and they are therefore rarely used.[7] Most practical Schottky diodes result from the union between lightly doped N-type silicon and a class of materials called *silicides*. These substances are definite compounds of silicon and certain metals, for example platinum and palladium. Silicides exhibit very stable work functions and therefore form Schottky diodes that have consistent and repeatable characteristics.

1.2.4. Zener Diodes

Under normal conditions, only a small current flows through a reverse-biased PN junction. This leakage current remains approximately constant until the reverse bias exceeds a certain critical voltage, beyond which the PN junction suddenly begins to conduct large amounts of current (Figure 1.15). The sudden onset of significant reverse conduction is called *reverse breakdown,* and it can lead to device destruction if the current flow is not limited by some external means. Reverse breakdown often sets the maximum operating voltage of a solid-state device. However, if appropriate precautions are taken to limit the current flow, a junction in reverse breakdown can provide a fairly stable voltage reference.

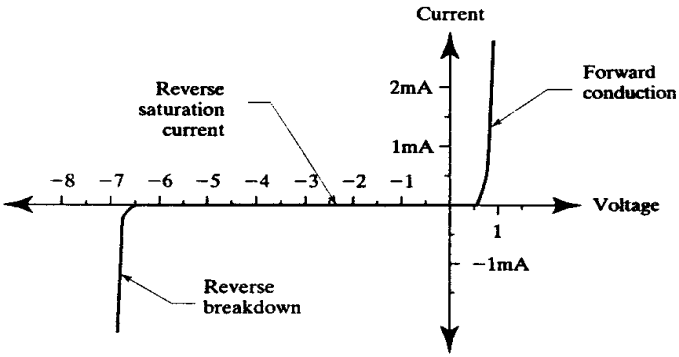

FIGURE 1.15 Reverse breakdown in a PN junction diode.

One of the mechanisms responsible for reverse breakdown is called *avalanche multiplication.* Consider a PN junction under reverse bias. The width of the depletion region increases with bias, but not fast enough to prevent the electric field from intensifying. The intense electric field accelerates the few carriers crossing the depletion region to extremely high velocities. When these carriers collide with lattice atoms, they knock loose valence electrons and generate additional carriers. This

[7] For example, compare the differences in work functions for platinum with respect to N-type silicon (0.85V) and P-type silicon (0.25V): R. S. Muller and T. I. Kamins, *Device Electronics for Integrated Circuits,* 2nd ed. (New York: John Wiley and Sons, 1986), p. 157.

process is aptly named because a single carrier can spawn literally thousands of additional carriers through collisions, just as a single snowball can start an avalanche.

The other mechanism behind reverse breakdown is called *tunneling*. Tunneling is a quantum-mechanical process that allows particles to move short distances regardless of any apparent obstacles. If the depletion region is thin enough, then carriers can leap across it by tunneling. The tunneling current depends strongly on both the depletion region width and the voltage difference across the junction. Reverse breakdown caused by tunneling is called *Zener breakdown*.

The reverse breakdown voltage of a junction depends on the width of its depletion region. Wider depletion regions produce higher breakdown voltages. As previously explained, the more lightly doped side of a junction sets its depletion region width and therefore its breakdown voltage. When the breakdown voltage is less than five volts, the depletion region is so thin that Zener breakdown predominates. When the breakdown voltage exceeds five volts, avalanche breakdown predominates. A PN diode designed to operate in reverse conduction is called either a *Zener diode* or an *avalanche diode*, depending on which of these two mechanisms predominates. Zener diodes have breakdown voltages of less than five volts, while avalanche diodes have breakdown voltages of more than five volts. Engineers traditionally call all breakdown diodes Zeners regardless of what mechanism underlies their operation. This can lead to confusion because a 7V Zener conducts primarily by avalanche breakdown.

In practice, the breakdown voltage of a junction depends on its geometry as well as its doping profile. The above discussion analyzed a *planar junction* consisting of two uniformly doped semiconductor regions intersecting in a planar surface. Although some real junctions approximate this ideal, most have curved sidewalls. The curvature intensifies the electric field and reduces the breakdown voltage. The smaller the radius of curvature, the lower the breakdown voltage. This effect can have a dramatic impact on the breakdown voltages of shallow junctions. Most Schottky diodes have sharp discontinuities at the edge of the metal-silicon interface. Electric field intensification can drastically reduce the measured breakdown voltage of a Schottky diode unless special precautions are taken to relieve the electric field at the edges of the Schottky barrier.

Figure 1.16 shows schematic symbols for all of the diodes discussed above. The PN junction diode uses a straight line to denote the cathode, while the Schottky diode and Zener diode are indicated by modifications to the cathode bar. In all cases, the arrow indicates the direction of conventional current flow through the forward-biased diode. In the case of the Zener diode, this arrow can be somewhat misleading because Zeners are normally operated in reverse bias. To the casual observer, the symbol may thus appear to be inserted "the wrong way around."

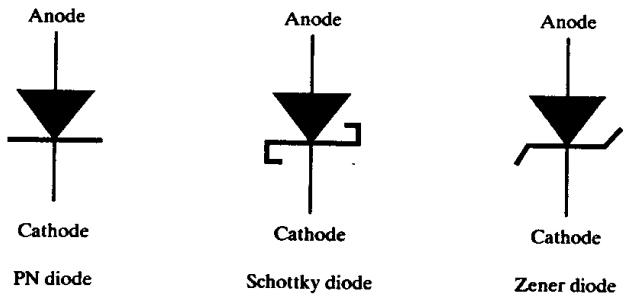

FIGURE 1.16 Schematic symbols for PN junction, Schottky, and Zener diodes. Some schematics show the arrowheads unfilled or show only half the arrowheads.

1.2.5. Ohmic Contacts

Contacts must be made between metals and semiconductors in order to connect solid-state devices into a circuit. These contacts would ideally be perfect conductors, but in practice they are *Ohmic contacts* that exhibit a small amount of resistance. Unlike rectifying contacts, these Ohmic contacts will conduct current equally well in either direction.

Schottky barriers can exhibit Ohmic conduction if the semiconductor material is doped heavily enough. The high concentration of dopant atoms thins the depletion region to the point where carriers can easily tunnel across it. Unlike normal Zener diodes, Ohmic contacts can support tunneling at very low voltages. Rectification does not occur since the carriers can effectively bypass the Schottky barrier by tunneling through it.

An Ohmic contact can also form if a Schottky barrier's contact potential causes surface accumulation rather than surface depletion. In accumulation, a thin layer of majority carriers forms at the semiconductor surface. In the case of an N-type semiconductor, this layer consists of excess electrons. The metal is a conductor and therefore cannot support a depletion region. A thin film of charge thus appears at the surface of the metal to counterbalance the accumulated carriers in the silicon. The lack of a depletion region on either side of the barrier prevents the contact from supporting a voltage differential, and any externally applied voltage will sweep carriers across the junction. Carriers can flow in either direction, so this type of Schottky barrier forms an Ohmic contact rather than a rectifying one.

In practice, rectifying contacts form to lightly doped silicon, and Ohmic contacts form to heavily doped silicon. The exact mechanism behind Ohmic conduction is unimportant since all Ohmic contacts behave in essentially the same manner. A lightly doped silicon region can be Ohmically contacted only if a thin layer of more heavily doped silicon is placed beneath the contact. Contact resistances of less than $50\Omega/\mu m^2$ can be obtained if a heavily doped silicon layer is used in combination with a suitable metal system. This resistance is small enough that it can be neglected for most applications.

Any junction between dissimilar materials exhibits a contact potential equal to the difference between the work functions of the materials. This rule applies to Ohmic contacts as well as to PN junctions and rectifying Schottky barriers. If all the contacts and junctions are held at the same temperature, then the sum of the contact potentials around any closed loop will equal zero. Contact potentials are, however, strong functions of temperature. If one of the junctions is held at a different temperature than the others, then its contact potential will shift and the sum of the contact potentials will no longer equal zero. This *thermoelectric effect* has significant implications for integrated circuit design.

Figure 1.17 shows a block of N-type silicon contacted on either side by aluminum. If one end of the block is heated, then a measurable voltage develops across the

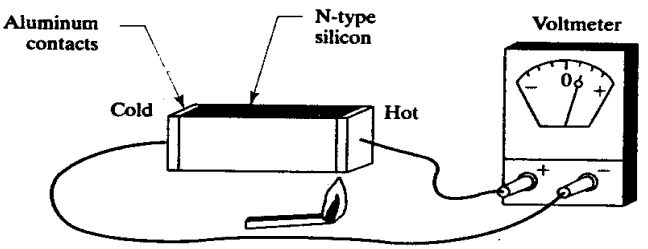

FIGURE 1.17 The thermoelectric effect produces a net measurable voltage if the two contacts are held at different temperatures.

block due to the mismatch between the two contact potentials. This voltage drop is typically 0.1–1.0mV/°C.[8] Many integrated circuits rely upon voltages matching within a millivolt or two, so even small temperature differences are enough to cause such circuits to malfunction.

1.3 BIPOLAR JUNCTION TRANSISTORS

While diodes are useful devices, they cannot amplify signals, and almost all electronic circuits require amplification in one form or another. One device that can amplify signals is called a *bipolar junction transistor* (BJT).

The structures of the two types of bipolar junction transistors are shown in Figure 1.18. Each transistor consists of three semiconductor regions called the *emitter, base,* and *collector.* The base is always sandwiched between the emitter and the collector. An NPN transistor consists of an N-type emitter, a P-type base, and an N-type collector. Similarly, a PNP transistor consists of a P-type emitter, an N-type base, and a P-type collector. In these simplified cross-sections, each region of the transistor consists of a uniformly doped section of a rectangular bar of silicon. Modern bipolar transistors have somewhat different cross-sections, but the principles of operation remain the same.

Figure 1.18 also shows the symbols for the two types of transistors. The arrowhead placed on the emitter lead indicates the direction of conventional current flow through the forward-biased emitter-base junction. No arrow appears on the collector lead even though a junction also exists between the collector and the base. In the simplified transistors of Figure 1.18, the emitter-base and collector-base junctions appear to be identical. One could apparently swap the collector and emitter leads without affecting the behavior of the device. In practice, the two junctions have different doping profiles and geometries and are not interchangeable. The emitter lead is distinguished from the collector lead by the presence of the arrowhead.

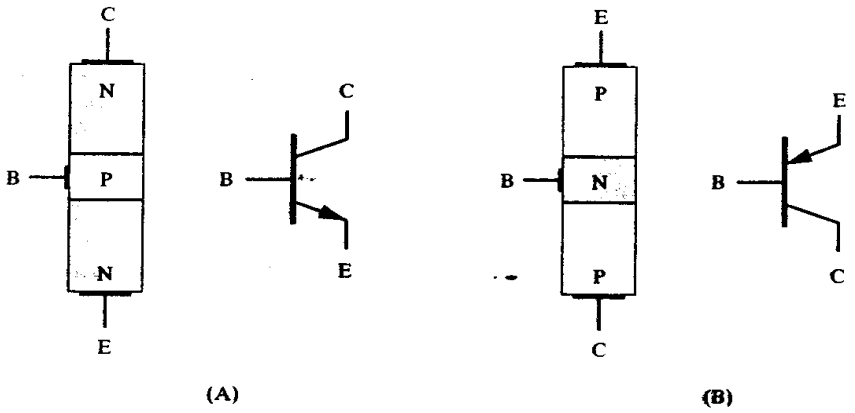

FIGURE 1.18 Structures and schematic symbols for the NPN transistor (A) and the PNP transistor (B).

A bipolar junction transistor can be viewed as two PN junctions connected back-to-back. The base region of the transistor is very thin (about 1–2μm wide). When the two junctions are placed in such close proximity, carriers can diffuse from one junction to the other before they recombine. Conduction across one junction therefore affects the behavior of the other junction.

[8] Lightly-doped silicon exhibits a higher Seebeck voltage; these values are taken from Widlar, *et al.*, p. 79.

Figure 1.19A shows an NPN transistor with zero volts applied across the base-emitter junction and five volts applied across the base-collector junction. Neither junction is forward biased, so very little current flows through any of the three terminals of the transistor. A transistor with both junctions reverse biased is said to be in *cutoff*. Figure 1.19B shows the same transistor with ten microamps of current injected into its base. This current forward biases the base-emitter junction to a potential of about 0.65V. A collector current a hundred times larger than the base current flows across the base-collector junction even though this junction remains reverse biased. This current is a consequence of the interaction between the forward-biased base-emitter junction and the reverse-biased base-collector junction. Whenever a transistor is biased in this manner, it is said to operate in the *forward active* region. If the emitter and collector terminals are interchanged so that the base-emitter junction becomes reverse-biased and the base-collector junction becomes forward-biased, the transistor is said to operate in the *reverse active* region. In practice, transistors are seldom operated in this manner.

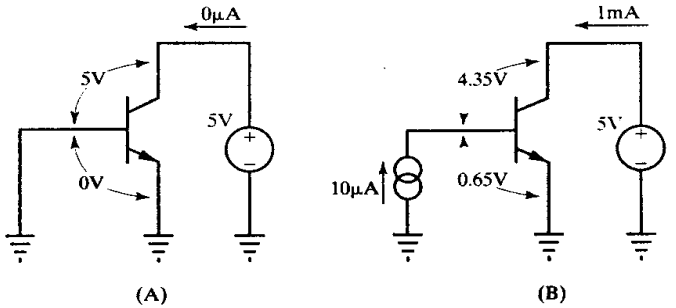

FIGURE 1.19 An NPN transistor operating in cutoff (A) and in the forward active region (B).

Figure 1.20 helps explain why collector current flows across a reverse-biased junction. Carriers flow across the base-emitter junction as soon as it becomes forward biased. Most of the current flowing across this junction consists of electrons injected from the heavily doped emitter into the lightly doped base. Most of these electrons diffuse across the narrow base before they recombine. The base-collector junction is reverse biased, so very few majority carriers can flow from the base into the collector. The same electric field that opposes the flow of majority carriers actually aids the flow of minority carriers. The electrons are minority carriers in the base, so they are swept across the reverse-biased base-collector junction into the collector. Here they again become majority carriers flowing toward the collector terminal. The collector current consists of the electrons that successfully complete the journey from emitter to collector without recombining in the base.

Some of the electrons injected into the base do not reach the collector. Those that do not reach the collector recombine in the base region. Base recombination consumes holes that are replenished by a current flowing in from the base terminal. Some holes are also injected from the base into the emitter, where they rapidly recombine. These holes represent a second source of base terminal current. These recombination processes typically consume no more than 1% of the emitter current, so only a small base current is required to maintain the forward bias across the base-emitter junction.

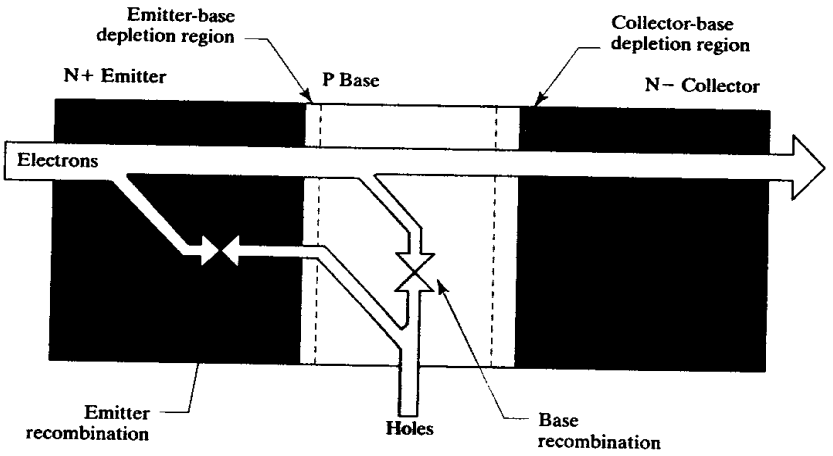

FIGURE 1.20 Current flow in an NPN transistor in the forward-active region.

1.3.1. Beta

The current amplification achieved by a transistor equals the ratio of its collector current to its base current. This ratio has been given various names, including *current gain* and *beta.* Likewise, different authors have used different symbols for it, including β and h_{FE}. A typical integrated NPN transistor exhibits a beta of about 150. Certain specialized devices may have betas exceeding 10,000. The beta of a transistor depends upon the two recombination processes illustrated in Figure 1.20.

Base recombination occurs primarily within the portion of the base between the two depletion regions, which is called the *neutral base* region. Three factors influence the base recombination rate: neutral base width, base doping, and the concentration of recombination centers. A thinner neutral base reduces the distance that the minority carriers must traverse and thus lessens the probability of recombination. Similarly, a more lightly doped base region minimizes the probability of recombination by reducing the majority carrier concentration. The *Gummel number* Q_B measures both of these effects. It is calculated by integrating the dopant concentration along a line traversing the neutral base region. In the case of uniform doping, the Gummel number equals the product of the base dopant concentration and the width of the neutral base. Beta is inversely proportional to the Gummel number.

The switching speed of transistors depends primarily on how quickly the excess minority carriers can be removed from the base, either through the base terminal or through recombination. Gold-doping is sometimes used to deliberately increase the number of recombination centers in bipolar junction transistors. The elevated recombination rate helps speed transistor switching, but it also reduces transistor beta. Few analog integrated circuits are built on gold-doped processes because of their low betas.

Bipolar transistors typically use a lightly doped base and a heavily doped emitter. This combination helps ensure that almost all of the current injected across the base-emitter junction consists of carriers flowing from emitter to base and not *vice-versa.* Heavy doping enhances the recombination rate in the emitter, but this has little impact since so few carriers are injected into the emitter in the first place. The

ratio of current injected into the emitter to that injected into the base is called the *emitter injection efficiency.*

Most NPN transistors use a wide, lightly doped collector in combination with a heavily doped emitter and a thin. moderately doped base. The light collector doping allows a wide depletion region to form in the neutral collector. This permits a high collector operating voltage without avalanching the collector-base junction. The asymmetric doping of emitter and collector helps explain why bipolar transistors do not operate well when these terminals are swapped.[9] A typical integrated NPN transistor with a forward beta of 150 has a reverse beta of less than 5. This difference is primarily due to the drastic reduction in emitter injection efficiency caused by the substitution of a lightly doped collector for a heavily doped emitter.

Beta also depends upon collector current. Beta is reduced at low currents by leakage and by low levels of recombination in the depletion regions. At modest current levels, these effects become insignificant and the beta of the transistor climbs to a peak value determined by the mechanisms discussed above. High collector currents cause beta to roll off due to an effect called *high-level injection.* When the minority carrier concentration approaches the majority carrier concentration in the base, extra majority carriers accumulate to maintain the balance of charges. The additional base majority carriers cause the emitter injection efficiency to decrease, which in turn causes beta to decrease. Most transistors are operated at moderate current levels to avoid beta roll-off, but power transistors must often operate in high-level injection because of size constraints.

The behavior of the PNP transistor is very similar to that of the NPN transistor. The beta of a PNP transistor is lower than that of an NPN of comparable dimensions and doping profiles, because the mobility of holes is lower than that of electrons. In many cases, the performance of the PNP is further degraded because of a conscious choice to optimize the NPN transistor at the expense of the PNP. For example, the material used to construct the base region of an NPN is often used to fabricate the emitter of a PNP. Since the resulting emitter is rather lightly doped, emitter injection efficiency is low, and the onset of high-level injection occurs at moderate current levels. Despite their limitations, PNP transistors are very useful devices, and most bipolar processes support their construction.

1.3.2. I-V Characteristics

The performance of a bipolar transistor can be graphically depicted by drawing a family of curves that relate base current, collector current, and collector-emitter voltage. Figure 1.21 shows a typical set of curves for an integrated NPN transistor. The vertical axis measures collector current I_C, while the horizontal axis measures collector-to-emitter voltage V_{CE}. A number of curves are superimposed upon the same graph, each representing a different base current I_B. This family of curves shows a number of interesting features of the bipolar junction transistor.

In the *saturation region,* the collector-emitter voltage remains so small that the collector-base junction is slightly forward-biased. The electric field that sweeps minority carriers across the collector-base junction still exists, so the transistor continues to conduct current. The collector-emitter voltage remains so low that Ohmic resistances in the transistor (particularly those in the lightly doped collector) become significant. The current supported in saturation is therefore less than that supported in the forward active region. The saturation region is of particular interest to integrated circuit

[9] This is only part of the explanation. The effective base width of the transistor also increases when it is operated in the reverse active mode.

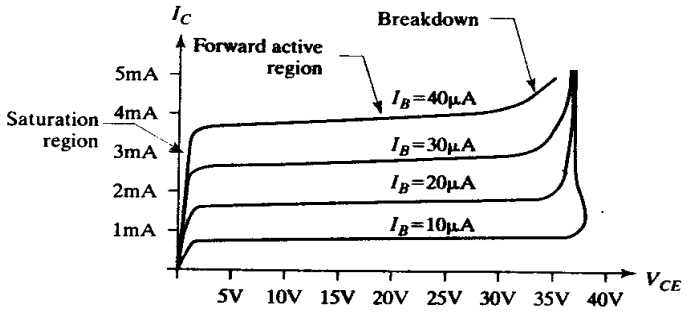

FIGURE 1.21 Typical I-V plot of an NPN transistor.

designers because the forward biasing of the collector-base junction injects minority carriers into the neutral collector. Section 8.1.4 discusses the effects of saturation upon integrated bipolar transistors in greater detail.

The collector-emitter voltage in the forward active region is large enough to reverse bias the collector-base junction. Ohmic drops in the collector no longer significantly reduce the electric field across the collector-base junction, so the current flow through the transistor now depends solely upon beta. The slight upward tilt to the current curves results from the *Early effect.* As the reverse bias on the collector-base junction increases, the depletion region at this junction widens and consequently the neutral base narrows. Since beta depends on base width, it increases slightly as the collector-emitter voltage rises. The Early effect can be minimized by using a very lightly doped collector, so the depletion region extends primarily into the collector rather than into the base.

Beyond a certain collector-emitter voltage, the collector current increases rapidly. This effect limits the maximum operating voltage of the transistor. In the case of a typical integrated NPN transistor, this voltage equals some 30V–40V. The increased current flow results from either one of two effects, the first of which is avalanche breakdown. The collector-base junction will avalanche if it is sufficiently reverse-biased. A wide lightly doped collector region can greatly increase the avalanche voltage rating, and discrete power transistors can achieve operating voltages of more than a thousand volts.

The second limiting mechanism is *base punchthrough.* Punchthrough occurs when the collector-base depletion region reaches all the way through the base and merges with the base-emitter depletion region. Once this occurs, carriers can flow directly from emitter to collector, and current is limited only by the resistance of the neutral collector and emitter. The resulting rapid increase in collector current mimics the effects of avalanche breakdown.

Base punchthrough is often observed in high-gain transistors. For example, *super-beta* transistors use an extremely thin base region to obtain betas of a thousand or more. Base punchthrough limits the operating voltage of these devices to a couple of volts. Super-beta transistors also display a pronounced Early effect because of the encroachment of the collector-base depletion region into the extremely thin neutral base. General-purpose transistors use wider base regions to reduce the Early effect, and their operating voltages are usually limited by avalanche instead of base punchthrough (Section 8.1.2).

1.4 MOS TRANSISTORS

The bipolar junction transistor amplifies a small change in input current to provide a large change in output current. The gain of a bipolar transistor is thus defined as the ratio of output to input current (beta). Another type of transistor, called a *field-*

effect transistor (FET), transforms a change in input voltage into a change in output current. The gain of an FET is measured by its *transconductance*, defined as the ratio of change in output current to change in input voltage.

The field-effect transistor is so named because its input terminal (called its *gate*) influences the flow of current through the transistor by projecting an electric field across an insulating layer. Virtually no current flows through this insulator, so the gate current of a FET transistor is vanishingly small. The most common type of FET uses a thin silicon dioxide layer as an insulator beneath the gate electrode. This type of transistor is called a ***metal-oxide-semiconductor*** (MOS) transistor, or alternatively, a ***metal-oxide-semiconductor*** *field-effect transistor* (MOSFET). MOS transistors have replaced bipolars in many applications because they are smaller and can often operate using less power.

The MOS transistor can be better understood by first considering a simpler device called a *MOS capacitor*. This device consists of two electrodes, one of metal and one of extrinsic silicon, separated by a thin layer of silicon dioxide (Figure 1.22A). The metal electrode forms the *gate*, while the semiconductor slab forms the *backgate* or *body*. The insulating oxide layer between the two is called the *gate dielectric*. The illustrated device has a backgate consisting of lightly doped P-type silicon. The electrical behavior of this MOS capacitor can be demonstrated by grounding the backgate and biasing the gate to various voltages. The MOS capacitor of Figure 1.22A has a gate potential of 0V. The difference in work functions between the metal gate and the semiconductor backgate causes a small electric field to appear across the dielectric. In the illustrated device, this field biases the metal plate slightly positive with respect to the P-type silicon. This electric field attracts electrons from deep within the silicon up toward the surface, while it repels holes away from the surface. The field is weak, so the change in carrier concentrations is small and the overall effect upon the device characteristics is minimal.

Figure 1.22B shows what occurs when the gate of the MOS capacitor is biased positively with respect to the backgate. The electric field across the gate dielectric strengthens and more electrons are drawn up from the bulk. Simultaneously, holes are repelled away from the surface. As the gate voltage rises, a point is reached where more electrons than holes are present at the surface. Due to the excess electrons, the surface layers of the silicon behave as if they were N-type. The apparent reversal of doping polarity is called *inversion* and the layer of silicon that inverts is called a *channel*. As the gate voltage increases still further, more electrons accumulate at the surface and the channel becomes more strongly inverted. The voltage at which the channel just begins to form is called the *threshold voltage* V_t. When the voltage difference between gate and backgate is less than the threshold voltage, no channel forms. When the voltage difference exceeds the threshold voltage, a channel appears.

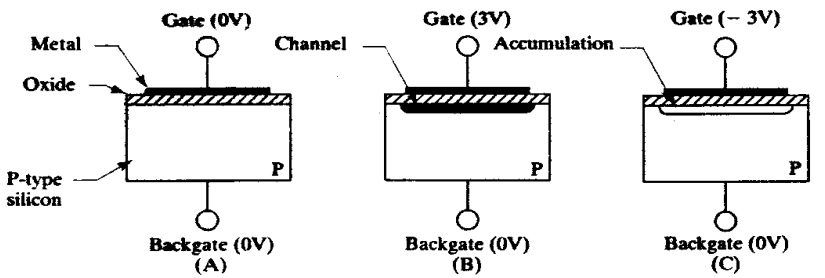

FIGURE 1.22 MOS capacitor: (A) unbiased (V_{BG} = 0V), (B) inversion (V_{BG} = 3V), (C) accumulation (V_{BG} = –3V).

Figure 1.22C shows what happens if the gate of the MOS capacitor is biased negatively with respect to the backgate. The electric field now reverses, drawing holes toward the surface and repelling electrons away from it. The surface layers of silicon appear to be more heavily doped, and the device is said to be in *accumulation.*

The behavior of the MOS capacitor can be utilized to form a true MOS transistor. Figure 1.23A shows the cross section of the resulting device. The gate, dielectric, and backgate remain as before. Two additional regions are formed by selectively doping the silicon on either side of the gate. One of these regions is called the *source* and the other is called the *drain.* Imagine that the source and backgate are both grounded and that a positive voltage is applied to the drain. As long as the gate-to-backgate voltage remains less than the threshold voltage, no channel forms. The PN junction formed between drain and backgate is reverse-biased, so very little current flows from drain to backgate. If the gate voltage exceeds the threshold voltage, a channel forms beneath the gate dielectric. This channel acts like a thin film of N-type silicon shorting the source to the drain. A current consisting of electrons flows from the source across the channel to the drain. In summary, drain current will only flow if the gate-to-source voltage V_{GS} exceeds the threshold voltage V_t.

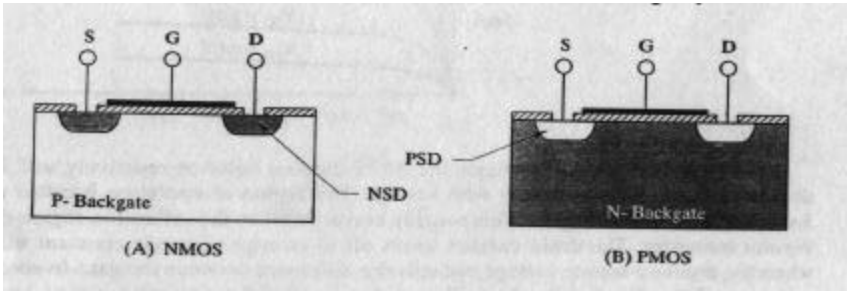

FIGURE 1.23 Cross sections of MOSFET transistors: NMOS (A) and PMOS (B). In these diagrams, S = Source, G = Gate, and D = Drain. The backgate connections, though present, are not illustrated.

The source and drain of a MOS transistor are interchangeable, as both are simply N-type regions formed in the P-type backgate. In many cases, these two regions are identical and the terminals can be reversed without changing the behavior of the device. Such a device is said to be *symmetric.* In a symmetric MOS transistor the labeling of source and drain becomes somewhat arbitrary. By definition, carriers flow out of the source and into the drain. The identity of the source and the drain therefore depends on the biasing of the device. Sometimes the bias applied across the transistor fluctuates and the two terminals swap roles. In such cases, the circuit designer must arbitrarily designate one terminal the drain and the other the source.

Asymmetric MOS transistors are designed with different source and drain dopings and geometries. There are several reasons why transistors may be made asymmetric, but the result is the same in every case. One terminal is optimized to function as the drain and the other as the source. If source and drain are swapped, then the performance of the device will suffer.

The transistor depicted in Figure 1.23A has an N-type channel and is therefore called an *N-channel MOS transistor,* or NMOS. *P-channel MOS* (PMOS) transistors also exist. Figure 1.23B shows a sample PMOS transistor consisting of a lightly doped N-type backgate with P-type source and drain regions. If the gate of this transistor is biased positive with respect to the backgate, then electrons are drawn to the surface and holes are repelled away from it. The surface of the silicon accumulates, and no channel forms. If the gate is biased negative with respect to the backgate, then holes are drawn to the surface, and a channel forms. The PMOS transistor thus

has a negative threshold voltage. Engineers often ignore the sign of the threshold voltage since it is normally positive for NMOS transistors and negative for PMOS transistors. An engineer might say, "The PMOS V_t has increased from 0.6V to 0.7V" when in actuality the PMOS V_t has shifted from –0.6V to –0.7V.

1.4.1. Threshold Voltage

The *threshold voltage* of a MOS transistor equals the gate-to-source bias required to just form a channel with the backgate of the transistor connected to the source. If the gate-to-source bias is less than the threshold voltage, then no channel forms. The threshold voltage exhibited by a given transistor depends on a number of factors, including backgate doping. dielectric thickness, gate material, and excess charge in the dielectric. Each of these effects will be briefly examined.

Backgate doping has a major effect on the threshold voltage. If the backgate is doped more heavily, then it becomes more difficult to invert. A stronger electric field is required to achieve inversion, and the threshold voltage increases. The backgate doping of an MOS transistor can be adjusted by performing a shallow implant beneath the surface of the gate dielectric to dope the channel region. This type of implant is called a *threshold adjust implant* (or *V_t adjust implant*).

Consider the effects of a V_t adjust implant upon an NMOS transistor. If the implant consists of acceptors, then the silicon surface becomes more difficult to invert and the threshold voltage increases. If the implant consists of donors, then the surface becomes easier to invert and the threshold decreases. If enough donors are implanted, the surface of the silicon can actually become counterdoped. In this case, a thin layer of N-type silicon forms a permanent channel at zero gate bias. The channel becomes more strongly inverted as the gate bias increases. As the gate bias is decreased, the channel becomes less strongly inverted and at some point it vanishes. The threshold voltage of this NMOS transistor is actually negative. Such a transistor is called a *depletion-mode NMOS*, or simply a *depletion NMOS*. In contrast, an NMOS with a positive threshold voltage is called an *enhancement-mode NMOS*, or *enhancement NMOS*. The vast majority of commercially fabricated MOS transistors are enhancement-mode devices, but there are a few applications that require depletion-mode devices. A depletion-mode PMOS can also be constructed. Such a device will have a positive threshold voltage.

Depletion-mode devices should always be explicitly identified as such. One cannot rely on the sign of the threshold voltage to convey this information, because many engineers customarily ignore threshold polarities. Therefore, one should say "a depletion-mode PMOS with a threshold of 0.7V," rather than a PMOS with a threshold of 0.7V." Many engineers would interpret the latter statement as indicating an enhancement PMOS with a threshold of –0.7V rather than a depletion PMOS with a threshold of +0.7V. Explicitly referring to depletion-mode devices as such eliminates any possibility of confusion.

Special symbols are often used to distinguish between different types of MOS transistors. Figure 1.24 shows a representative collection of these symbols.[10] Symbols A and B are the standard symbols for NMOS and PMOS transistors, respectively. These symbols are not commonly used in the industry; instead symbols

[10] Symbols A, B, E, F, G, and H are used by various authors; see A. B. Grebene, *Bipolar and MOS Analog Integrated Circuit Design* (New York: John Wiley and Sons, 1984), pp. 112–113; also P. R. Gray and R. G. Meyer, *Analysis and Design of Analog Integrated Circuits*, 3rd ed. (New York: John Wiley and Sons, 1993), p. 60. The *J. Solid State Circuits* also uses three-terminal MOS symbols but differentiates PMOS devices by placing a bubble on their gate leads.

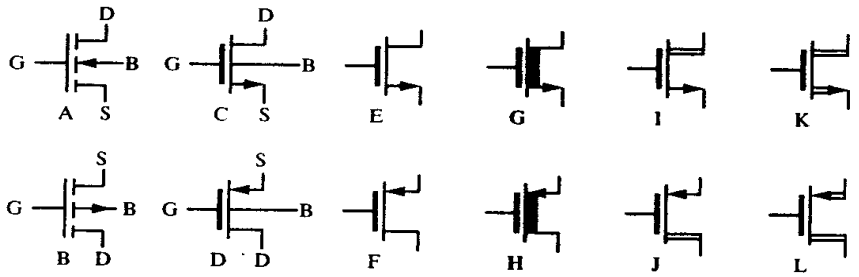

FIGURE 1.24 MOSFET symbols: A, B: standard symbols; C, D: industry symbols (four-terminal); E, F: industry symbols (three-terminal); G, H: depletion-mode devices; I, J: asymmetric high-voltage MOS symbols; K, L: symmetric high-voltage MOS symbols.

C and D are preferred for NMOS and PMOS transistors, respectively. These symbols intentionally resemble NPN and PNP transistors. This convention helps highlight the essential similarities between MOS and bipolar circuits. Symbols E and F are sometimes employed when the backgates of the transistors connect to known potentials. Every MOS transistor has a backgate, so this terminal must always connect to something. Symbols E and F are potentially confusing, because the reader must infer the backgate connections. These symbols are nonetheless very popular because they make schematics much more legible. Symbols G and H are often used for depletion-mode devices, where the solid bar from drain to source represents the channel present at zero bias. Symbols I and J are sometimes employed for asymmetric transistors with high-voltage drains, and symbols K and L are used for symmetric transistors with high-voltage terminations for both source and drain. There are many other schematic symbols for MOS transistors; the ones shown in Figure 1.24 form only a representative sample.

Returning to the discussion of threshold voltage, the dielectric also plays an important role in determining the threshold voltage. A thicker dielectric weakens the electric field by separating the charges by a greater distance. Thus, thicker dielectrics increase the threshold voltage while thinner ones reduce it. In theory, the material of the dielectric also affects the electric field strength. In practice, almost all MOS transistors use pure silicon dioxide as the gate dielectric. This material can be grown in extremely thin films of exceptional purity and uniformity; no other material has comparable properties. Alternate dielectric materials therefore have very limited application.[11]

The gate electrode material also affects the threshold voltage of the transistor. As mentioned above, an electric field appears across the gate oxide when the gate and backgate are shorted together. This field is produced by the difference in work functions between the gate and backgate materials. Most practical transistors use heavily doped polysilicon for the gate electrode. The work function of polysilicon can be varied to a limited degree by changing its doping.

A potentially troublesome source of threshold voltage variation comes from the presence of excess charges in the gate oxide or along the interfaces between the oxide and the silicon surface. These charges may consist of ionized impurity atoms, trapped carriers, or structural defects. The presence of trapped electric charge in the dielectric or along its interfaces alters the electric field and therefore the threshold voltage. If the amount of trapped charge varies with time, temperature, or applied bias, then the threshold voltage will also vary. This subject is discussed in greater detail in Section 4.2.2.

[11] A few devices have been fabricated using high-permittivity materials such as silicon nitride for the gate dielectric. Some authors use the term *insulated-gate field effect transistor* (IGFET) to refer to all MOS-like transistors, including those with non-oxide dielectrics.

1.4.2. I-V Characteristics

The performance of an MOS transistor can be graphically illustrated by drawing a family of I-V curves similar to those used for bipolar transistors. Figure 1.25 shows a typical set of curves for an enhancement NMOS. The source and backgate were connected together to obtain these particular curves. The vertical axis measures drain current I_D, while the horizontal axis measures drain-to-source voltage V_{DS}. Each curve represents a specific gate-to-source voltage V_{GS}. The general character of the curves resembles that of the bipolar transistor shown in Figure 1.21, but the family of curves for an MOS transistor are obtained by stepping gate voltage, while those for a bipolar transistor are obtained by stepping base current.

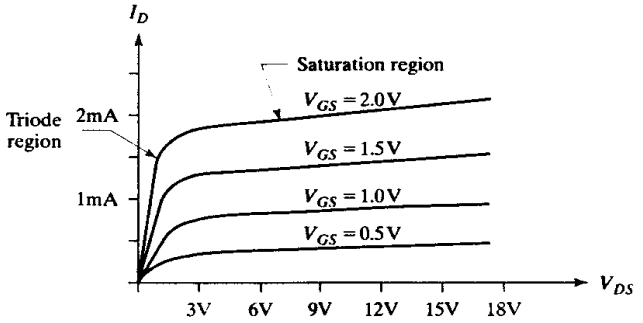

FIGURE 1.25 Typical I-V plot of an NMOS transistor.

At low drain-to-source voltages the MOS channel behaves resistively, and the drain current increases linearly with voltage. This region of operation is called the *linear region* or *triode region*. This roughly corresponds to the saturation region of a bipolar transistor. The drain current levels off to an approximately constant value when the drain-to-source voltage exceeds the difference between the gate-to-source voltage and the threshold voltage. This region is called the *saturation region*, and it roughly corresponds to the forward active region of a bipolar transistor. The term *saturation* thus has very different meanings for MOS and bipolar transistors.

The behavior of the MOS transistor in the linear region is easily explained. The channel acts as a film of doped silicon with a characteristic resistance that depends upon the carrier concentration. The current increases linearly with voltage, exactly as one would expect of a resistor. Higher gate voltages produce larger carrier concentrations and therefore lessen the resistance of the channel. PMOS transistors behave similarly to NMOS transistors, but since holes have lower mobilities than electrons, the apparent resistance of the channel is considerably greater. The effective resistance of an MOS transistor operating in the triode region is symbolized $R_{DS(on)}$.

MOS transistors saturate because of a phenomenon called *pinch-off*. While the drain-to-source voltage remains small, a depletion region of uniform thickness surrounds the channel (Figure 1.26A). As the drain becomes more positive with respect to the source, the depletion region begins to thicken at the drain end. This depletion region intrudes into the channel and narrows it. Eventually the channel depletes all the way through and it is said to have *pinched off* (Figure 1.26B). Carriers move down the channel propelled by the relatively weak electric field along it. When they reach the edge of the pinched-off region, they are sucked across the depletion region by the strong electric field. The voltage drop across the channel does not increase as the drain voltage is increased; instead the pinched-off region widens. Thus, the drain current reaches a limit and ceases to increase.

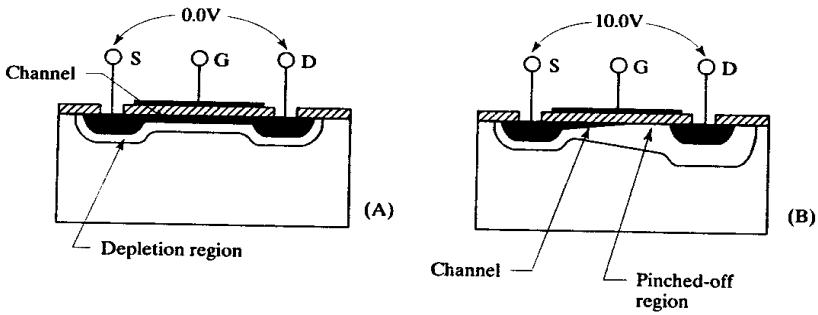

FIGURE 1.26 Behavior of a MOS transistor under bias: (A) V_{DS} = 0V (triode region); (B) V_{DS} = 10V (saturation region).

The drain current curves actually tilt slightly upward in the saturation region. This tilt is caused by *channel length modulation*, which is the MOS equivalent of the Early effect. Increases in drain voltage cause the pinched-off region to widen and the channel length to shorten. The shorter channel still has the same potential drop across it, so the electric field intensifies and the carriers move more rapidly. The drain current thus increases slightly with increasing drain-to-source voltage.

The I-V curves of Figure 1.25 were obtained with the backgate of the transistor connected to the source. If the backgate is biased independently of the source, then the apparent threshold voltage of the transistor will vary. If the source of an NMOS transistor is biased above its backgate, then its apparent threshold voltage increases. If the source of a PMOS transistor is biased below its backgate, then its threshold voltage decreases (it becomes more negative). This *backgate effect*, or *body effect*, arises because the backgate-to-source voltage modulates the depletion region beneath the channel. This depletion region widens as the backgate-to-source differential increases, and it intrudes into the channel as well as into the backgate. A high backgate-to-source differential will thin the channel, which in turn raises the apparent threshold voltage. The intrusion of the depletion region into the channel becomes more significant as the backgate doping rises, and this in turn increases the magnitude of the body effect.

MOS transistors are normally considered majority carrier devices, which conduct only after a channel forms. This simplistic view does not explain the low levels of conduction that occur at gate-to-source voltages just less than the threshold voltage. The formation of a channel is a gradual process. As the gate-to-source voltage increases, the gate first attracts small numbers of minority carriers to the surface. The concentration of minority carriers rises as the voltage increases. When the gate-to-source voltage exceeds the threshold, the number of minority carriers becomes so large that the surface of the silicon inverts and a channel forms. Before this occurs, minority carriers can still move from the source to the drain by diffusion. This *subthreshold conduction* produces currents that are much smaller than those that would flow if a channel were present. However, they are still many orders of magnitude greater than junction leakages. Subthreshold conduction is typically significant only when the gate-to-source voltage is within about 0.3V of the threshold voltage. This is sufficient to cause serious "leakage" problems in low-V_t devices. Some electrical circuits actually take advantage of the exponential voltage-to-current relationship of subthreshold conduction, but these circuits cannot operate at temperatures much in excess of 100°C because the junction leakages become so large that they overwhelm the tiny subthreshold currents.

As with bipolar transistors, MOS transistors can break down by either avalanche or punchthrough. If the voltage across the depletion region at the drain becomes so large that avalanche multiplication occurs, the drain current increases rapidly. Similarly, if the entire channel pinches off, then the source and drain will be shorted by the resulting depletion region and the transistor will punch through.

The operating voltage of an MOS transistor is often limited to a value considerably below the onset of avalanche or punchthrough by a long-term degradation mechanism called *hot carrier injection.* Carriers that traverse the pinched-off portion of the drain are accelerated by the strong electric field present here. The carriers can achieve velocities far beyond those normally associated with room-temperature thermal diffusion, so they are called *hot carriers.* When these carriers collide with atoms near the silicon surface, some of them are deflected up into the gate oxide, and a few of these become trapped. Slowly, over a long period of operation, the concentration of these trapped carriers increases and the threshold voltage shifts. Hot hole injection occurs less readily than hot electron injection because the lower mobility of holes limits their velocity and therefore their ability to surmount the oxide interface. For this reason, NMOS transistors are frequently limited to lower operating voltages than PMOS transistors of similar construction. Various techniques have been devised to limit hot carrier injection (Section 12.1).

1.5 JFET TRANSISTORS

The MOS transistor represents only one type of field-effect transistor. Another is the *junction field-effect transistor* or JFET. This device uses the depletion regions surrounding reverse-biased junctions as a gate dielectric. Figure 1.27A shows a cross-section of an N-channel JFET. This device consists of a bar of lightly doped N-type silicon called the *body* into which two P-type diffusions have been driven from opposite sides. The thin region of N-type silicon remaining between the junctions forms the *channel* of the JFET. The two diffusions act as the *gate* and the *backgate* and the opposite ends of the body form the *source* and the *drain.*

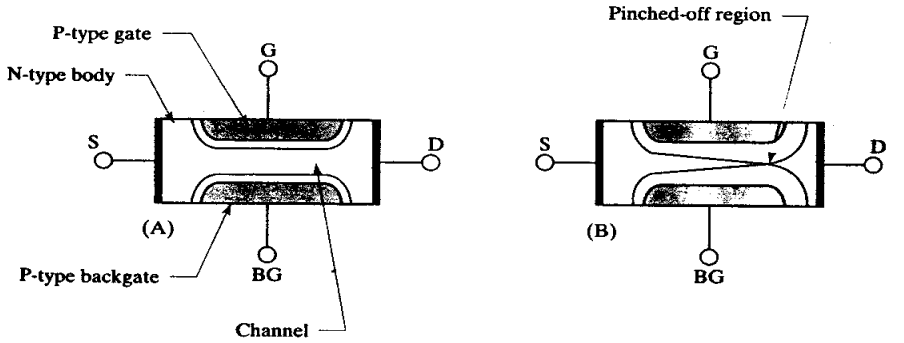

FIGURE 1.27 Cross sections of an N-channel JFET transistor operating in the linear region (A) and in saturation (B). In both diagrams, S = Source, D = Drain, G = Gate, and BG = Backgate.

Suppose that all four terminals of the N-channel JFET are grounded. Depletion regions form around the gate-body and backgate-body junctions. These depletion regions extend into the lightly doped channel, but they do not actually touch one another. A channel therefore exists from the drain to the source. If the drain voltage rises above the source voltage, then a current flows through the channel from

drain to source. The magnitude of this current depends on the resistance of the channel, which in turn depends on its dimensions and doping. As long as the drain-to-source voltage remains small, it does not significantly alter the depletion regions bounding the channel. The resistance of the channel therefore remains constant and the drain-to-source voltage varies linearly with drain current. Under these conditions, the JFET is said to operate in its *linear region*. This region of operation corresponds to the linear (or triode) region of an MOS transistor. Since a channel forms at $V_{GS} = 0$, the JFET resembles a depletion-mode MOSFET rather than an enhancement-mode one.

The depletion regions at the drain end of the JFET widen as the drain voltage increases. The channel becomes increasingly constricted by the encroachment of the opposing depletion regions. Eventually the depletion regions meet and pinch off the channel (Figure 1.27B). Drain current still flows through the transistor even though the channel has pinched off. This current originates at the source terminal and consists of majority carriers (electrons). These carriers move down the channel until they reach the pinched-off region. The large lateral electric field across this region draws the carriers across into the neutral drain.

Further increases in drain voltage have little effect once the channel has pinched off. The pinched-off region widens slightly, but the dimensions of the channel remain about the same. The resistance of the channel determines the magnitude of the drain current, so this also remains approximately constant. Under these conditions, the JFET is said to operate in *saturation*.

The gate and backgate electrodes also influence the current that flows through the channel. As magnitudes of the gate-body and backgate-body voltages increase, the reverse biases across the gate-body and backgate-body junctions slowly increase. The depletion regions that surround these junctions widen and the channel constricts. Less current can flow through the constricted channel, and the drain-to-source voltage required to pinch the channel off decreases. As the magnitudes of the gate and backgate voltages continue to increase, eventually the channel will pinch off even at $V_{DS} = 0$. Once this occurs, no current can flow through the transistor regardless of drain-to-source voltage, and the transistor is said to operate in *cutoff*.

Figure 1.28 shows the I-V characteristics of an N-channel JFET whose gate and backgate electrodes have been connected to one another. Each curve represents a different value of the gate-to-source voltage V_{GS}. The drain currents are at their greatest when $V_{GS} = 0$, and they decrease as the magnitude of the gate voltage

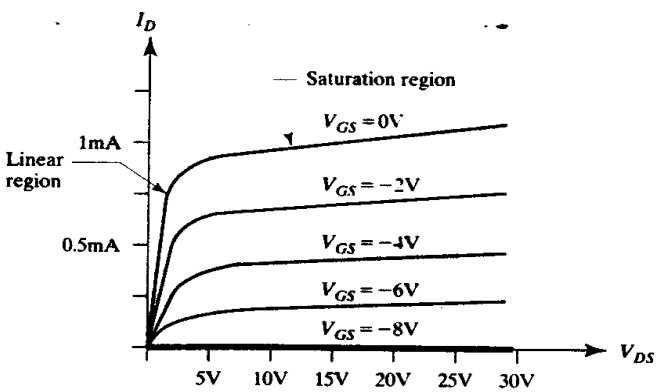

FIGURE 1.28 Typical I-V plot of an N-JFET transistor with $V_T = -8V$.

increases. Conduction ceases entirely when the gate voltage equals the *turnoff voltage* V_T. The turnoff voltage qualitatively corresponds to the threshold voltage of an MOS transistor. The comparison must not be taken too far, however, as the conduction equations of the two devices differ considerably.

The drain current curves of the N-JFET tilt slightly upward in saturation due to *channel length modulation*. This effect is analogous to that which occurs in MOS transistors. The pinched-off region of the JFET lengthens as the drain-to-source voltage increases. Any increase in the length of the pinched-off region produces a corresponding decrease in the length of the channel. The effect of channel length modulation is usually quite small because the channel length greatly exceeds the length of the pinched-off region.

The source and drain terminals of a JFET can often be interchanged without affecting the performance of the device. The JFET structure of Figure 1.27A is an example of such a *symmetric* device. More complex JFET structures sometimes exhibit differences in source and drain geometries that render them *asymmetric*.

Almost all JFET structures short the gate and backgate terminals. Consider the device of Figure 1.27A. The channel is bounded on the left by the source, on the right by the drain, on the top by the gate, and on the bottom by the backgate. The drawing does not show what bounds the channel on the front or the rear. In most cases, these sides of the channel are also bounded by reverse-biased junctions that are extensions of the gate-body and backgate-body junctions. This arrangement necessitates shorting the gate and backgate.

Figure 1.29 shows the conventional schematic symbols for N-channel and P-channel JFET transistors. The arrowhead on the gate lead shows the orientation of the PN junction between the gate and the body of the device. The symbol does not explicitly identify the source and drain terminals, but most circuit designers orient the devices so that the drain of an N-JFET and the source of a P-JFET lie on top.

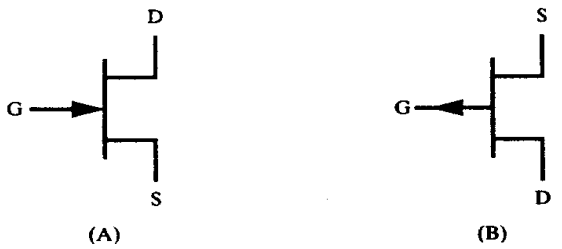

FIGURE 1.29 Symbols for an N-channel JFET (A) and a P-channel JFET (B).

1.6 SUMMARY

Device physics is a complex and ever-evolving science. Researchers constantly develop new devices and refine existing ones. Much of this ongoing research is highly theoretical and therefore lies beyond the scope of this introductory text. The functionality of most semiconductor devices can be satisfactorily explained using relatively simple and intuitive concepts.

This chapter emphasizes the role of majority and minority carrier conduction across PN junctions. If a junction is reverse-biased, then the majority carriers on either side of it are repelled and a depletion region forms. If the junction is forward-biased, then majority carriers diffuse across and recombine to create a net current flow across the PN junction. The PN junction diode employs this phenomenon to rectify signals.

When two junctions are placed in close proximity, carriers emitted by one junction can be collected by the other before they can recombine. The bipolar junction transistor (BJT) consists of just such a pair of closely spaced junctions. The voltage across the base-emitter junction of the BJT controls the current flowing from collector to emitter. If the transistor is properly designed, then a small base current can control a much larger collector current. The BJT therefore serves as an amplifier capable of transforming weak signals into much stronger ones. Thus, for example, a BJT can amplify a weak signal picked up by a radio receiver into a signal strong enough to drive a loudspeaker.

The metal-oxide-semiconductor (MOS) transistor relies upon electrical fields projected across a dielectric to modulate the conductivity of a semiconductor material. A suitable voltage placed upon the gate of a MOS transistor produces an electric field that attracts carriers up from the bulk silicon to form a conductive channel. The gate is insulated from the remainder of the transistor, so no gate current is required to maintain conduction. MOS circuitry can thus potentially operate at very low power levels.

The junction diode, the bipolar junction transistor, and the MOS transistor are the three most important semiconductor devices. Together with resistors and capacitors, they form the vast majority of the elements used in modern integrated circuits. The next chapter will examine how these devices are fabricated in a production environment.

1.7 EXERCISES

1.1. What are the relative proportions of aluminum, gallium, and arsenic atoms in intrinsic aluminum gallium arsenide?

1.2. A sample of pure silicon is doped with exactly 10^{16} atoms/cm^3 of boron and exactly 10^{16} atoms/cm^3 of phosphorus. Is the doped sample P-type or N-type?

1.3. The *instantaneous* velocity of carriers in silicon is almost unaffected by weak electric fields, yet the *average* velocity changes dramatically. Explain this observation in terms of drift and diffusion.

1.4. A layer of intrinsic silicon 1μ thick is sandwiched between layers of P-type and N-type silicon, both heavily doped. Draw a diagram illustrating the depletion regions that form in the resulting structure.

1.5. A certain process incorporates two different N+ diffusions that can be combined with a P– diffusion to produce Zener diodes. One of the resulting diodes has a breakdown voltage of 7V, while the other has a breakdown voltage of 10V. What causes the difference in breakdown voltages?

1.6. When the collector and emitter leads of an integrated NPN transistor are swapped, the transistor continues to function but exhibits a greatly reduced beta. There are several possible reasons for this behavior; explain at least one.

1.7. If a certain transistor has a beta of 60, and another transistor has a base twice as wide and half as heavily doped, then what is the approximate beta of the second transistor? What other electrical characteristics of the devices will vary, and how?

1.8. A certain MOS transistor has a threshold voltage of –1.5V. If a small amount of boron is added to the channel region, the threshold voltage shifts to –0.6V. Is the transistor PMOS or NMOS, and is it an enhancement or a depletion device?

1.9. If a depletion PMOS transistor has a threshold voltage of 0.5V when constructed using a 200Å oxide, will this threshold voltage increase or decrease if the oxide is thickened to 400Å?

1.10. A certain NMOS transistor has a threshold voltage of 0.5V; the gate-to-source voltage V_{GS} of the transistor is set to 2V, and the drain-to-source voltage V_{DS} is set to 4V.

What is the relative effect of doubling the gate-to-source voltage versus doubling the drain-to-source voltage, and why?

1.11. A certain silicon PN junction diode exhibits a forward voltage drop of 620mV when operated at a forward current of 25μA at a temperature of 25°C. What is the approximate forward drop of this diode at –40°C? At 125°C?

1.12. Two JFET transistors differ only in the separation between their gate and backgate; in one transistor these two regions are twice as far apart as in the other transistor. In what ways do the electrical properties of the two transistors differ?