CHAPTER

2

ECLECTRONICS

This chapter introduces a relation between the study of biological neural systems and that of electronic neural systems. The chapter title, **eclectronics**, is derived in the following manner:

ec·lec·tic **1.** selecting what is thought best in various doctrines, methods, or styles. **2.** consisting of components from diverse sources.

e·lec·tron·ic of, based on, or operated by the controlled flow of charge carriers, especially electrons.

e·lec·tron·ics the science and technology of electronic phenomena and devices.

ec·lec·tron·ics the common framework of electrical properties used for information processing in both brain and silicon.

As we have mentioned, both neural and electronic systems represent information as electrical signals. Neurobiologists deal with neural systems, and have evolved a viewpoint, notation, jargon, and set of preconceptions that they use in any discussion of neural networks. Likewise, electrical engineers have developed an elaborate language and symbolism that they use to describe and analyze transistor circuits. In both cases, the language, viewpoint, and cultural bias derive partly from the properties inherent in the technology, and partly from the perspectives and ideas of early influential workers in the field. By now, it is extremely difficult

to separate the conceptual framework and vernacular of either field from the properties of the devices and systems being studied.

Because it is the express purpose of this book to explore the area of potential synergy of these two fields, we must develop a common conceptual framework within which both can be discussed. Such an undertaking will, of necessity, require a reevaluation of the underlying assumptions of the existing lore. In particular, it will be possible neither to preserve all the detailed distinctions prevalent in the current literature in either field, nor to pay lip service to the many schools of thought that intersect in a plethora of combinations. Rather, we will present a simple, unifying perspective within which the function of either technology can be visualized, described, and analyzed. Such a viewpoint is possible for two reasons:

1. The operation of elementary devices in both technologies can be described by the aggregate behavior of *electrically charged entities* interacting with *energy barriers*. In both cases, the rate at which dynamic processes take place is determined by the energy due to random thermal motion of the charged entities, which is accurately described by the Boltzmann distribution. The steady-state value of any quantity of interest is, in both cases, the result of equalization of the rate of processes tending to increase, and those tending to decrease, that quantity's value.
2. The properties of devices in both technologies are not well controlled. The operation of any robust system formed from the devices of either technology, therefore, must not be dependent on the detailed characteristics of any particular device. For system purposes, a device can be adequately described by an abstraction that captures its essential behavior and omits the finer detail.

In the process of any simplification, it is, by definition, necessary to omit detail. By "essential behavior," we mean those relationships that are necessary to reason about the correct operation of the system. We will attempt to develop a simplification that does not lose these relationships. In other words, as far as they go, the explanations in this book are intended to provide a conceptually correct foundation in the following sense: Gaining a deeper understanding of any given point should not require unlearning any conceptualization or formulation. History alone will determine the extent to which we approach this ideal.

ELECTRICAL QUANTITIES

WARNING: If you are already familiar with electric circuits, Boltzmann statistics, energy diagrams, and neural and transistor physics, you will be bored to tears with the following material—we urge you to skip the remainder of this chapter. If you have a background in solid-state physics, you should read the introduction to the elements of neuroscience in Chapter 4. If you are an expert in neuroscience, you should read the introduction to transistor operation in Chapter 3. Read the following discussion if you lack a firm preparation in either discipline.

Energy

All electrical mechanisms are concerned with the interaction of electrical **charges**. The concept of an elementary charge is so ingrained in our curricula that we seldom question its origin. The electrical force f_e that attracts two electrical charges q_1 and q_2 of opposite type (one positive and the other negative) is

$$f_e = \frac{1}{4\pi\epsilon}\frac{q_1 q_2}{r^2} \tag{2.1}$$

where ϵ is called the **permittivity** of the medium in which the charges are embedded. When ϵ is given in terms of ϵ_0, the permittivity of free space (vacuum), it is called the **dielectric constant** of the medium. In the units we will use, $\epsilon_0 = 8.85 \times 10^{-12}$ farads per meter.

To aid us in visualizing the interaction of electrical charges, we will first examine the analogous behavior of masses interacting through the force of gravity. The gravitational force f_g between two masses m_1 and m_2 due to their mutual gravitational attraction is

$$f_g = G\frac{m_1 m_2}{r^2} \tag{2.2}$$

where G is the gravitational constant. Note that Equation 2.2 is of exactly the same form as Equation 2.1.

Gravitational force plays a key role in our everyday experience. We could not get up in the morning, throw a baseball, drive a car, or water a lawn without an intuitive understanding of its operation. Yet no one but the astronauts has experienced gravitation in the form shown in Equation 2.2. Being earthbound mortals, we walk on the surface of a planet of mass M and of radius r. We must be content to manipulate a mass m that is infinitesimal compared with M over distances that are much smaller than r. Under those conditions, the gravitational law we encounter in daily life is a simplified form of Equation 2.2, in which M, r, and G can all be lumped into a new constant g, defined by

$$\begin{aligned} f_g &= m\frac{MG}{r^2} \\ &= mg \end{aligned} \tag{2.3}$$

The quantity f_g in Equation 2.3 is called the **weight** of the object. The quantity

$$g = \frac{MG}{r^2}$$

is the force per unit mass, and is called the **gravitational field** due to the mass M.

We also know from common experience that an expenditure of energy is required to raise a mass to a higher elevation; that energy can be recovered by allowing the mass to fall in the gravitational field. The amount of **potential energy** (PE) stored in a mass at height h above the surface of the earth is just

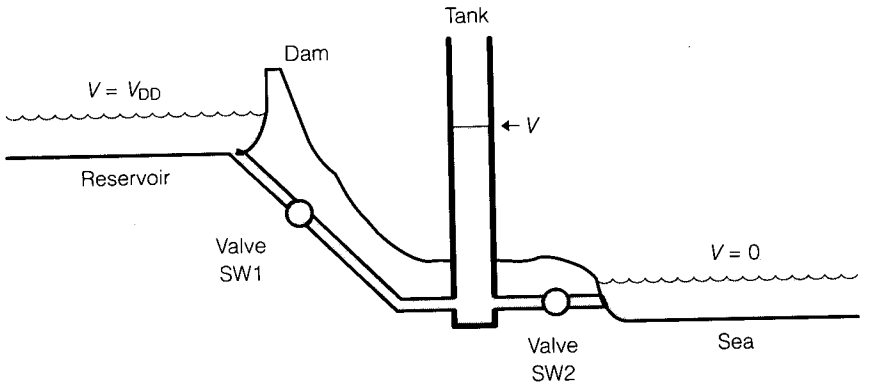

FIGURE 2.1 Hydraulic analogy of electronic or neural circuit. The power supply is a reservoir of water at potential V_{DD}. The reference potential is sea level, corresponding to ground in an electrical circuit. The water level in the tank corresponds to the output voltage. The tank can be filled by opening valve SW1, or emptied by opening valve SW2.

the integral of the force over the distance h. For values of h much smaller than r, the gravitational force is nearly constant and

$$\mathrm{PE} = mgh$$

The quantity gh is called the **gravitational potential**. It is the energy per unit mass of matter at height h above the earth's surface. We will use the symbol V for potential in both gravitational and electrical paradigms, to emphasize the similarity of the underlying physics. In both cases, the field is the gradient of the potential.

Fluid Model

In all system-level abstractions of device behavior, be those devices neural or electronic, electrical charges are present in sufficient number that they cannot be accounted for individually. In both disciplines, they are treated as an *electrical fluid*, the flow of which is called the **electrical current**. We can extend the gravitational-energy concepts of the preceding section to give us an intuitively simple yet conceptually correct picture of the operation of electrical systems. In this analog, shown in Figure 2.1, water represents the electrical fluid. The gravitational potential V, which is directly proportional to the height of the water, represents the **electrical potential**.

The electrical potential V also is called the **voltage;** the two terms are synonymous and are used interchangeably. The quantity of water represents the quantity of electrical charge Q. There is an underlying granularity to these quantities: Water is made up of water molecules; charge in a neuron is made up of ions; charge in a transistor is made up of electrons. In all cases, we can discuss the quantity in terms of either the number of elementary particles or a more

convenient macroscopic unit. For water, we use liters; for electrical charge we use **coulombs**. (Throughout this book, we will use the meter-kilogram-second-ampere [MKSA] system of units.) The magnitude q of the charge on an electron or monovalent ion is 1.6×10^{-19} coulombs. Similarly, we can measure the flow—or current—in terms of elementary particles per second or in macroscopic units. For water, we use liters per second; for the electrical current I, we use coulombs per second. Electrical current is such an important and universally used quantity that it has a unit of its own: 1 coulomb per second is called an **ampere**, or **amp.**

In all cases, it requires 1 unit of energy to raise a quantity of 1 unit to a potential of 1 unit. In the MKSA system, the unit of energy is the **joule**. To raise 1 kilogram 1 meter above the surface of the earth requires 1 joule of energy. To raise the potential of 1 coulomb of charge by 1 volt requires 1 joule of energy. It is often convenient to discuss microscopic processes in terms of the *energy per particle*. The energy required to raise the potential of one electronic charge by 1 volt is 1 **electron volt**. An electron volt is, of course, 1.6×10^{-19} joule.

Let us reexamine Figure 2.1. We see a reservoir filled with water, the surface level of which is called V_{DD} (dermis of the dam water). The reservoir is used to fill a tank, under the control of valve SW1. To empty the tank, we can open valve SW2, allowing water to discharge into the sea. Note that, with this system, the height of water in the tank cannot exceed V_{DD}, and cannot be reduced below sea level. We can measure the height of water from any reference point we choose. There is an advantage, however, to one particular choice. If we choose sea level as the **reference potential** ($V = 0$), the height always will be positive or zero. In an electric circuit, it is useful to have such a reference potential, from which all voltages are measured. The reference corresponding to sea level is called **ground**.

The electrical equivalent of the reservoir in Figure 2.1 is called a **power supply**. The reservoir can be kept full only if the water is replenished after it is depleted. In many cases, a pump is used for this purpose. So we need a mechanism to run the pump when the reservoir level is low, and to shut off the pump when the surface is above the desired level.

In neurons, voltage-sensitive pumps run by metabolic processes in the cell actively pump potassium ions into and sodium ions out of the cell's cytoplasm. Potassium ions exist as a minority ionic species in the extracellular fluid. This ionic gradient acts as the power supply for electrical activity in the neuron.

In an electronic circuit, a reservoir of charge is provided either by an electrochemical process (as in a battery), or by an active circuit that monitors the potential of the reservoir and adds charge as required. Such a circuit is called a **regulated power supply**.

Capacitance

For the moment, we will consider the role of the apparatus of Figure 2.1 to be the manipulation of the water level in the tank to the desired level. By closing valve SW1 and opening SW2, we can reduce the level to zero. The amount

f water required to increase the level in the tank by a certain amount—say, meter—is obviously dependent on the cross-sectional area of the tank. That rea can be expressed in acres, square feet, or some other arbitrary unit. For consistency, however, it is convenient to express it as the quantity of water required o raise the water level by 1 unit of potential. In an electrical circuit, such a storge tank is called a **capacitor**, and the electrical charge required to raise the otential level by 1 volt is its **capacitance**. The unit of capacitance—coulombs er volt —is called the **farad**. The total charge Q on a capacitor, like the total vater in the tank, is related to C, the capacitance, and to V, the voltage, by the xpression

$$Q = CV \tag{2.4}$$

Current I is, by definition, the rate at which charge is flowing:

$$I = \frac{dQ}{dt}$$

In the particular case where the capacitance C is constant, independent of the voltage V, the current flowing into the capacitor results in a rate of change of the potential

$$I = C\frac{dV}{dt}$$

In our water analogy, constant capacitance corresponds to a tank with constant cross-sectional area, independent of elevation.

Resistance and Conductance

If both valves in Figure 2.1 are opened, water will flow from the reservoir into the sea at a finite rate, restricted by the diameter of the pipe through which it must pass. If the water level in the reservoir is increased, water will flow more quickly. If the diameter of the pipe is decreased, water will flow more slowly. The property of the pipe that restricts the flow of water is called **resistance**. The electrical element possessing this quality is called a **resistor**. A current I through a resistor of resistance R is related to the voltage difference V between the two ends of the resistor by

$$V = IR \tag{2.5}$$

A voltage of 1 volt across a unit resistance will cause a current flow of 1 amp. The unit of resistance, the volt per amp, is called an **ohm**.

It often is convenient to view an electrical circuit element in terms of its *willingness* to carry current rather than of its reluctance to do so. When a nerve membrane is excited, it passes more current than it does when it is at rest. When a transistor has a voltage on its gate, it carries more current than it does when its gate is grounded. For this reason, both biological and electronic elements are

described by a **conductance** G. The conductance is defined as the current per unit voltage:

$$G = \frac{1}{R} = \frac{I}{V}$$

The unit of conductance, the amp per volt, is called a **mho**. In the neurobiology literature, the mho is called the **siemens**.

Equipotential Regions

We are all familiar with bodies of water that have flat surfaces (lakes, oceans), and with others that have sloping surfaces (rivers, streams). In Figure 2.1, we can identify regions in which the water level is flat (independent of position on the surface). The reservoir, the sea, and the inside of the tank are examples. To a first approximation, the water level in such **equipotential regions** will stay flat whether or not water is flowing in the pipes. In an electrical circuit, equipotential regions are called **electrical nodes**. By definition, a node is a region characterized by a single potential. As we proceed, we often will describe the dependence of one potential on that of other nodes, and will talk about that potential's evolution with respect to time. For these discussions, we will refer to names or labels attached to the nodes. Because there is a one-to-one relationship between nodes and voltages, we often will use the voltage label as the name of the node—as "I_1 and I_2 join at the V_1 node." This convention also will be applied to resistors and capacitors: The same label will be used for both their *name* and their *value*. This abuse of notation will be indulged only where there is a one-to-one correspondence, and thus no confusion can result.

SCHEMATIC DIAGRAMS

Once we have identified the nodes, we can construct an abstraction of the physical situation by lumping all the charge storage into capacitive elements, and all the resistance to current flow into resistors. The abstraction can be expressed either as a coupled set of differential equations or as a diagram called a **schematic**. Most people find it convenient to develop a schematic from the physical system, and then to write equations for the schematic. A schematic for Figure 2.1 is shown in Figure 2.2.

In any abstraction, certain details of the physical situation are omitted. The idealizations in Figure 2.2 are as follows:

1. We have treated the two valves as *switches*; they can be either on or off, but they cannot assume intermediate values
2. We have neglected the volume of water stored in the pipes
3. We have treated the reservoir and the mechanism by which it is kept full as a lumped constant *voltage source*, shown as a battery

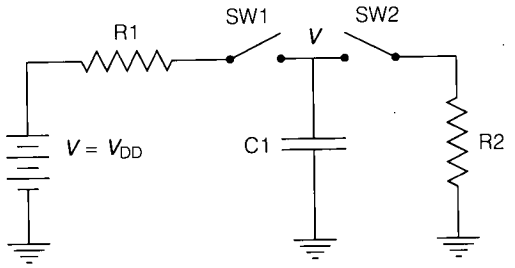

FIGURE 2.2 Schematic diagram representing the essential behavior of the physical system shown in Figure 2.1. The resistor R1 corresponds to the friction in the upper pipe, and R2 to that in the lower pipe. The capacitor represents the water-storage capacity of the tank. The two switches are analogous to the two valves. The battery is an abstraction of the reservoir, and must include a mechanism that keeps it full.

4. We have neglected any voltage drop (difference in water level) in the upper reservoir caused by water flowing out of the pipe through SW1
5. We have assumed all dependencies to be *linear*, as defined by Equations 2.5 and 2.4

The assumption of linearity in item 5 is a property, not of the schematic representation, but rather of the individual components. Many elements in both biological and electronic systems are highly nonlinear. The nonlinearities produced by most physical devices are smooth, however, and can be treated as linear for small excursions from any given operating point. We will encounter many examples in which the locally linear approximation gives us valuable information about an inherently nonlinear system.

These five idealizations have allowed us to construct a precise mathematical **model**, embodied in Figure 2.2. We will need such a model to analyze any physical system, because the myriad details in any real system are intractable to analysis. We intend the model to capture all effects that are relevant at the level of detail we are considering, and to omit the potentially infinite detail that would not affect the outcome in any substantial way. It is clear that no analysis can be any better than the model on which it is based. Constructing a good model requires consummate skill, judgment, experience, and taste. Once we have constructed an elegant model, analysis may present mathematical difficulties but does not require conceptual advances.

Models required for biological systems often are inherently *nonlinear*, in the sense that no linear system will behave even qualitatively as the observed system does. Analysis techniques for nonlinear systems have not evolved to nearly the level of generality as have those for linear systems. For this reason, when we wish to model a biological system, we have to pay more attention to the qualitative aspects of its behavior. Our models generally will evolve as we increase our understanding of the system. In fact, we can argue that the best models in a complex discipline *are* the embodiment of the understanding of that discipline.

The remainder of this book is devoted to a modeling technique quite different from any previously attempted. Not only are the resulting models directly related to the system under study, but the models themselves are real-time working physical systems. They can be used directly in engineering applications.

LINEAR SUPERPOSITION

Idealization 5 in the previous section has put us in a position to state the single most important principle in the analysis of electrical circuits: the **principle of linear superposition**. For any arbitrary network containing resistors and voltage sources, we can find the solution for the network (the voltage at every node and the current through every resistor) by the following procedure. We find the solution for the network in response to each voltage source individually, with all other voltage sources reduced to zero. The full solution, including the effects of all voltage sources, is just the sum of the solutions for the individual voltage sources. In addition to linearity of the component characteristics, there must be a well-defined reference value for voltages (ground), to which all node potentials revert when *all* sources are reduced to zero.

This principle applies to circuits containing current sources as well as to those containing voltage sources. It applies even if the sources are functions of time, as we will discuss in Chapter 8. It also applies to circuits containing capacitors, provided that any initial charge on a capacitor is treated as though it were a voltage source in series with the capacitor. Finally, the principle is applicable to networks containing transistors, or other elements with smooth nonlinearities, if the signal amplitudes are small enough that the circuit acts linearly within the range of signal excursions.

The analytical advantage we derive from this principle lies in the ability it gives us to treat the contribution of each individual input in isolation, knowing that the effect of each input is independent of the values of the other inputs. Thus, we need not worry that several inputs will combine in strange and combinatorially complex ways.

ACTIVE DEVICES

Although there are many details of any technology that can be changed without compromising our ability to create useful systems, there is one essential ingredient without which it simply is not possible to process information. That key ingredient is *gain*.

The nervous system is constantly bombarded by an enormous variety of sensory inputs. Perhaps the most important contribution of early sensory information-processing centers is to inhibit the vast majority of unimportant and therefore unwanted inputs, in order to concentrate on the immediately important stimuli. Of course, all inputs must be available at all times, lest an unseen preda-

tor leap suddenly from an inhibited region. Therefore, any real-time system of this sort will, at any time, develop its outputs from a small subset of its inputs. Because every input is the output of some other element, it follows that every element must be able to drive many more outputs than it receives as active inputs at any given time. An elementary device therefore must have **gain**—it must be able to supply more energy to its outputs than it receives from its input signals. Devices with this capability are called **active devices**.

In Figure 2.2, the active devices are shown as switches, but we have not specified what is required to open or close them. In both biology and electronics, valves are controlled not by some outside agent, as tacitly assumed in Figure 2.2, but rather by the *potential at some other point in the system.* Furthermore, valves cannot be treated as switches that are purely on or off; they assume intermediate values. Active devices play a central role in information processing; we will take a much closer look at how they work. In fact, this task will occupy us for the next two chapters.

THERMAL MOTION

The systems of particles that we will discuss—molecules in a gas, ions in a solution, or electrons in a semiconductor crystal—seem superficially so different that mentioning them in the same sentence may appear to be ridiculous. For many phenomena, we must exercise great care when drawing parallels between the behavior of these systems. For the phenomena of importance in computation, however, there is a deep similarity in the underlying physics of the three systems: in all cases, the trajectory of an individual particle is completely dominated by thermal motion. Collisions with the environment cause the particle to traverse a *random walk*, a familiar example of which is Brownian motion. We have no hope of following the detailed path of any given particle, let alone all the paths of a collection of particles. The quantities we care about—electrical charge, for example—are in any case sums over large numbers of particles. It suffices, then, to treat the average motion of a particle in a statistical sense.

By making three simplifications, we can derive the important properties of a collection of particles subject to random thermal motion by an intuitive line of reasoning:

1. We treat the inherently three-dimensional process as a one-dimensional model in the direction of current flow
2. We replace the distributions of velocities and free times by their mean values
3. We assume that electric fields are sufficiently small that they do not appreciably alter the thermal distribution

Although the resulting treatment is mathematically rough-and-ready, it is both mercifully brief and conceptually correct.

Drift

In any given environment, a particle will experience collisions at random intervals, either with others of its own kind (as in a gas), or with other thermally agitated entities (as in a semiconductor crystal lattice). In any case, there will be some **mean free time** (t_f) between collisions. We will make the simplest assumption about a collision: that all memory of the situation prior to the collision is lost, and the particle is sent off in a random direction with a random velocity. The magnitude of this velocity is, on average, v_T.

If the particle is subject to some external force f, from gravity or from an electric field, it will accelerate during the time it is free in accordance with Newton's law:

$$f = ma$$

Over the course of the time the particle is free, the particle will move with increasing velocity in the direction of the acceleration. Although the initial velocity is random after a collision, the small incremental change in velocity is always in the direction of the force. Over many collisions, the random initial velocity will average to zero, and we can therefore treat the particle as though it accelerated from rest after every collision. The distance s traveled in time t by a particle starting from rest with acceleration a is

$$s = \frac{1}{2}at^2$$

In the case of our model, over the time t_f between collisions, the acceleration $a = f/m$ will cause a net change δh in the position h of the particle:

$$\delta h = \frac{1}{2}at_f^2 = \frac{f}{2m}t_f^2$$

The average **drift velocity** (v_{drift}) of a large collection of particles subject to the force f per particle is just the net change in position δh per average time t_f between collisions:

$$v_{\text{drift}} = \frac{\delta h}{t_f} = \frac{ft_f}{2m} \tag{2.6}$$

Equation 2.6 describes the behavior of a uniform distribution of electrons or ions with charge q in the presence of an electric field E. The force on each particle is

$$f = qE$$

and therefore the drift velocity is linear in the electric field:

$$v_{\text{drift}} = \frac{qt_f}{2m}E = \mu E \tag{2.7}$$

The constant $\mu = qt_f/2m$ is called the **mobility** of the particle.

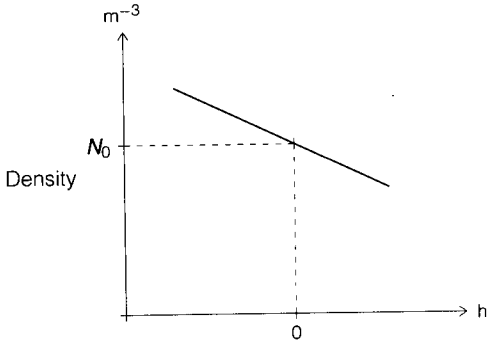

FIGURE 2.3 Density of particles as a function of some spatial dimension h. At $h = 0$, the particle density is N_0 per cubic meter. Particles will diffuse from regions of higher density (low h) to regions of lower density (high h).

Diffusion

In all structures that are interesting from an information-processing point of view, such as transistors and nerve membranes, there are large spatial gradients in the concentration of particles. Under such circumstances, there is a **diffusion** of particles from regions of higher density to those of lower density. Consider the situation shown in Figure 2.3. Particles diffusing from left to right cross the origin at some rate proportional to the gradient of the density N (number of particles per unit volume). That flow rate (J), given in particles per unit area per second, can be viewed as a movement of all particles to the right with some effective **diffusion velocity** (v_{diff}):

$$J = Nv_{\text{diff}} \tag{2.8}$$

For electrically charged particles, J usually is given in terms of charge per second per unit area (current per unit area), and is called the **current density**. This electrical current density is equal to the particle density given by Equation 2.8, multiplied by q.

A simple model of the diffusion process is shown in Figure 2.4. We have divided the spatial-dimension axis (h) into compartments small enough that a particle can, on the average, move from one to the other in one mean free time (t_f). Due to the local density gradient, there are $N_0 + \Delta N$ particles in the left compartment, but only N_0 particles in the right compartment. One t_f later, half of the particles in each compartment will have moved to the right, and half will have moved to the left. There is, therefore, a net movement of $\Delta N/2$ particles to the right. This flow is purely the result of the random movement of particles; the particles' individual velocities have no preferred direction. So $\Delta N/2$ particles move a distance $v_T t_f$ every t_f. Because there are N_0 total particles, the average velocity v_{diff} per particle in the $+h$ direction is $v_T \Delta N/(2N_0)$. The width of each

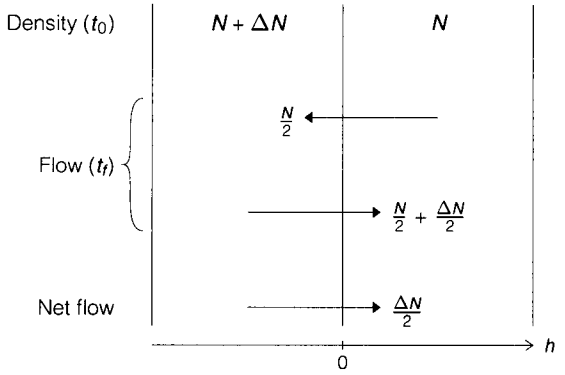

FIGURE 2.4 Flow of particles due to random thermal motion. The higher density to the left of the origin results in more particles crossing the axis to the right than to the left. The net flow to the right is the difference between these two rates.

compartment is $v_T t_f$, so

$$\Delta N \approx \frac{dN}{dh} v_T t_f$$

The diffusion velocity can thus be written in terms of the gradient of the particle distribution,

$$v_{\text{diff}} = -\frac{1}{2N}\frac{dN}{dh} t_f v_T^2 \tag{2.9}$$

The negative sign arises because the net flow of particles is from higher density to lower density.

In a one-dimensional model, a particle in thermal equilibrium has a mean kinetic energy that defines its **temperature** (T):

$$\frac{1}{2}mv_T^2 = \frac{1}{2}kT$$

Here k is Boltzmann's constant, m is the mass of the particle, and v_T is called the **thermal velocity**. At room temperature, kT is equal to 0.025 electron volt. Substituting v_T^2 into Equation 2.9, we obtain

$$v_{\text{diff}} = -\frac{1}{2N}\frac{dN}{dh} kT \frac{t_f}{m}$$

$$= -D\frac{1}{N}\frac{dN}{dh} \tag{2.10}$$

The quantity $D = kTt_f/2m$ is called the **diffusion constant** of the particle. Comparing Equation 2.7 with Equation 2.10, we can see that the mobility and

diffusion constants are related:

$$D = \frac{kT}{q}\mu \tag{2.11}$$

Although the preceding derivation was approximate, Equation 2.11 is an exact result called the **Einstein relation**; it was discovered by Einstein during his study of Brownian motion. It reminds us that drift and diffusion are not separate processes, but rather are two aspects of the behavior of an ensemble of particles dominated by random thermal motion.

The processes of drift and diffusion are the stuff of which all information-processing devices—both neural and semiconductor—are made. To understand the physics of active devices, we need one more conceptual tool—the energy diagram. Like a schematic, the energy diagram is a pictorial representation of a model. With the Boltzmann distribution, it forms a complete basis for the device physics that follows. We will discuss the Boltzmann distribution next, and will return to the energy diagram in Chapter 3.

Boltzmann Distribution

We all know that the earth's atmosphere is held to the earth's surface by gravitational attraction. Gas molecules have weight, and thus are subject to a force toward the center of the earth. If the temperature were reduced to absolute zero, the entire atmosphere would condense into a solid sheet about 5 meters thick. The distribution of matter in the atmosphere is the result of a delicate balance: Thermal agitation, through random collisions between molecules, tends to spread matter uniformly throughout space. Gravitational attraction tends to concentrate matter on the surface of the planet. In quantitative terms, the gravitational force produces a *drift velocity* toward the surface given by Equation 2.6, where the force f is just the weight w of each molecule, which we will assume to be independent of the height h over a small range of elevation. As molecules drift toward the surface, the density increases at lower elevations and decreases at higher elevations, thus forming a density gradient. This density gradient causes an upward *diffusion* of molecules, in accordance with Equation 2.10. Equilibrium is reached when the rate of diffusion upward due to the density gradient is equal to the rate of drift downward. Setting the two velocities equal, we obtain

$$v_{\text{drift}} = \frac{wt_f}{2m} = v_{\text{diff}} = -\frac{1}{2N}\frac{dN}{dh}kT\frac{t_f}{m}$$

We cancel out t_f and m, leaving a relationship between density and height:

$$\frac{1}{N}\frac{dN}{dh}kT = -w \tag{2.12}$$

Integration of Equation 2.12 with respect to h yields

$$kT \ln \frac{N}{N_0} = -wh \tag{2.13}$$

where N_0 is the density at the reference height $h = 0$. Exponentiation of both sides of Equation 2.13 gives

$$N = N_0 e^{-\frac{wh}{kT}} \tag{2.14}$$

The density of molecules per unit volume in the atmosphere decreases exponentially with altitude above the earth's surface.[1] The quantity wh is, of course, just the potential energy of the molecule.

Equation 2.13 can be generalized to any situation involving thermally agitated particles working against a gradient of potential energy. For charged particles, the potential energy is qV, and Equation 2.13 takes the form

$$V = -\frac{kT}{q}\ln\frac{N}{N_0} \tag{2.15}$$

The voltage V developed in response to a gradient in the concentration of a charged species, and exhibiting the logarithmic dependence on concentration shown in Equation 2.15, is called the **Nernst potential**. In the electrochemistry and biology literature, kT/q is written RT/F. Exponentiation of Equation 2.15 leads to

$$N = N_0 e^{-\frac{q}{kT}V} \tag{2.16}$$

Equation 2.16 is called the **Boltzmann distribution**. It describes the exponential decrease in density of particles in thermal equilibrium with a potential gradient. It is the basis for all exponential functions in the neural and electronic systems we will study.

[1] This treatment ignores all the complications present in a real atmosphere: convection, multiple gases, and so on.

CHAPTER

3

TRANSISTOR PHYSICS

The active devices in electronic systems are called **transistors**. Their function is to control the flow of current from one node based on the potential at another node. In terms of our analog of Figure 2.1 (p. 14), we construct a valve that is operated by the level in another tank. Although it is easy to imagine mechanical linkages that could cause a valve to operate, we will not succumb to the temptation to use that model. Rather, we will describe a somewhat less familiar but still intuitive model that accurately embodies all the necessary physics of the transistor.

BOLTZIAN HYDRAULICS

We saw in Chapter 2 that the density of gas molecules in the atmosphere decreases exponentially with height. For the planet Earth, the gravitational attraction, temperature, and molecular weight are such that the atmospheric density decreases by a factor of e for each approximately 20-kilometer increase in elevation. It is easy to imagine a planet with much higher gravitational attraction, lower temperature, and heavier molecules, such that the change in elevation required to decrease the density by a factor of e would be about 1 meter. Because such a planet would not be a hospitable place to live, we will travel to it only to illustrate the principles on which transistors operate. We will call this imaginary planet Boltzo.

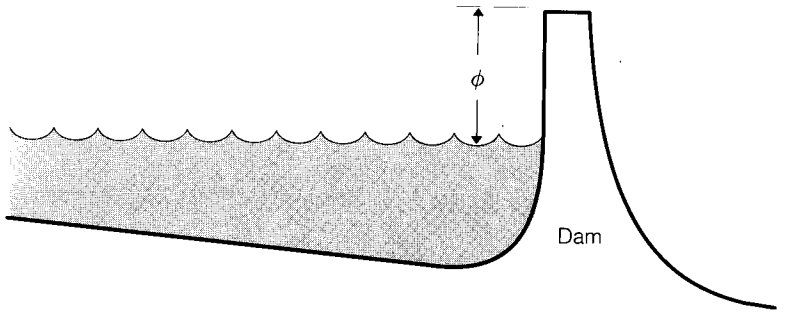

FIGURE 3.1 Cross-section of reservoir showing water level a distance ϕ below the top of the dam. Any water vapor must rise to this height to spill over the dam.

We will use a hydraulic power system similar to that shown in Figure 2.1 (p. 14). If we locate our Boltzian reservoir about 200 meters above sea level, the potential energy of a water molecule in the reservoir (measured in units of kT) will be about equal to that of an electron in the 5-volt power supply of an electronic circuit. (Had we located our reservoir at an altitude of 25 meters, we would have approximated the operation of a neural system, but that story is a bit more complicated. We will content ourselves for the moment with creating a hydraulic transistor.)

Early Boltzian engineers constructed a dam in the bottom of a deep canyon, as shown in Figure 3.1. The purpose of the dam was to create a reservoir to store water. When the rains were heavy and the reservoir nearly full, however, the Boltzians noticed that the water level did not stay constant, even if no water was withdrawn from the reservoir. They made an exhaustive search, but found no sign of leakage. Finally, they sought the advice of a wise and venerable Boltzian philosopher. After surveying the situation, he gave his reply: "Contemplate atop the dam in the quiet of the night."

This pronouncement instantly became the subject of much discussion in learned circles—what could he possibly mean by such a reply? One young engineering student named Lily Field grew tired of the endless arguments. Against prevailing sentiment, she undertook the long journey to the dam site. Sitting atop the massive structure as the noises of the day faded into darkness, she contemplated the meaning of the philosopher's words. All wind had ceased; not a blade of grass or a leaf stirred. Still she had the strong sense of a cold and heavy force against her back. Holding out her silk scarf like a sail, she noticed that it billowed out, as if pressed by some invisible force. When she raised it to eye level, the billowing was considerably weaker. She found a long stick, attached the scarf to its tip, and held it aloft; the billowing effect was scarcely detectable.

The mystery was solved. The water was not escaping in liquid form at all—it was *evaporating*, and the vapor was pouring over the dam in enormous quantities. There was a source of water behind the dam, and no source on the opposite side of the dam. There was thus a difference in density of water vapor on the two

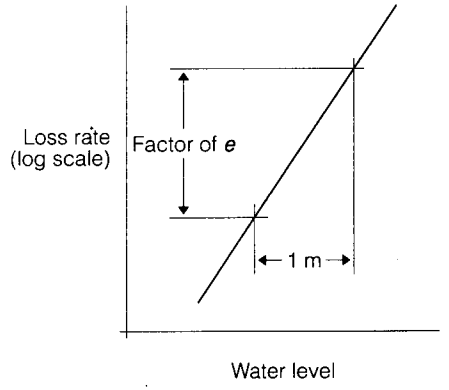

FIGURE 3.2 Loss rate of water vapor over the dam of Figure 3.1 as a function of water level in the reservoir. The distance ϕ decreases as the water level rises. Because the vapor density decreases exponentially with height above the water level, the loss rate increases exponentially with the water level.

sides of the dam. Water vapor was *diffusing* from the region of high vapor density to the region of low density. The density gradient, and hence the diffusion rate, was proportional to the density at the top of the dam. Because the density of the vapor decreased exponentially with height above the water level, the effect was noticeable only when the reservoir was very nearly full.

When Field returned to the city, she looked up data on the water loss rate and level over the several years that the dam had been operating. She plotted the log of the loss rate as a function of the height of the water level in the reservoir. The result was a straight line, as shown in Figure 3.2. For each meter that the water level rose, the loss rate increased by a factor of e.

Field presented her findings in a poster session at the next International Hydraulics Conference (IHC); her study soon became the talk of the entire meeting. A special evening discussion session was organized. Field showed the data, and pointed out that the engineers could greatly reduce the vapor loss by extending the height of the dam only a few meters. The additional structure would be required to support not the weight of liquid water, but only the density gradient of water *vapor*, so it could be constructed of light material such as wood or plastic, rather than of massive reinforced concrete.

In an invited paper at the IHC the following year, Dr. Field presented a mechanism for controlling the flow of vapor over the dam, shown in Figure 3.3. A barrier of light material is allowed to slide vertically in a slot in the middle of the dam. The barrier is supported on a series of floats. The floats are buoyed up by water from another reservoir or tank. If the level in the second tank is high, the barrier will be high and the flow of vapor over the dam will be low. If the level in the second tank is low, the barrier will be low, and the flow of vapor over the dam will be large. The flow rate I across the barrier will be directly proportional to the density gradient in the horizontal direction. That gradient will, in turn, be proportional to the density of water vapor at height ϕ above the liquid surface, given in Equation 2.14 (p. 25):

$$I = I_0 e^{-\frac{w\phi}{kT}} \tag{3.1}$$

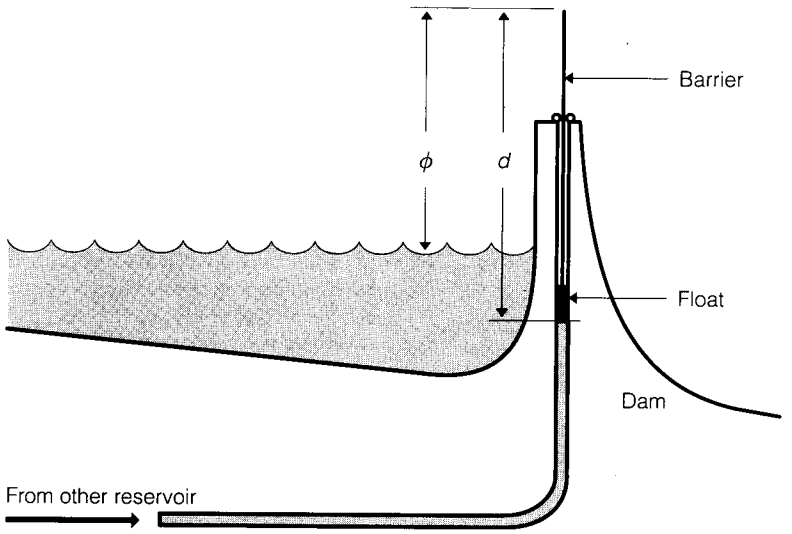

FIGURE 3.3 Hydraulic transistor. The barrier floats on water in a chamber filled from another reservoir. The water-vapor flow rate is an exponentially decreasing function of the water level in the second reservoir.

From the basic dimensions of the structure, Equation 3.1 can be expressed as

$$I = I_0 e^{-\frac{w}{kT}(h_2 - h_1 + d)} \tag{3.2}$$

where h_1 is the elevation of the main reservoir and h_2 is the elevation in the control reservoir. The physical height d of the barrier can be factored into a new preexponential constant,

$$I_1 = I_0 e^{-\frac{wd}{kT}}$$

The product wh of the molecular weight w and the reservoir level h is just the potential energy per molecule P. Equation 3.2 can thus be expressed as

$$I = I_1 e^{\frac{P_1 - P_2}{kT}} \tag{3.3}$$

The current decreases exponentially with the energy difference between the control reservoir and source reservoir. Equation 3.3 is the basic governing equation for the operation of a transistor. The only difference between Boltzo and silicon is that, in the latter, gravitational forces have been replaced by electrostatic ones. The Boltzian landscape is the energy profile of our electronic structures. Cross-sections of the energy terrain are called **energy diagrams**, and appear in all discussions of electron device physics. We will see the energy diagram for an MOS transistor later in this chapter.

SEMICONDUCTORS

All physical structures—neural, electronic, or mechanical—are built out of atoms. Before proceeding, we will briefly review the properties of these basic building blocks out of which transistors and neurons are made.

Atoms and All That

Atoms can be viewed as a swarm of electrons circulating around a nucleus containing protons and neutrons. The negatively charged electrons are, of course, attracted to the positively charged nucleus, and will orbit as close to it as the laws of quantum mechanics allow. An **element** is a type of atom; each atom of a given element will have the same number of electrons. The total number of electrons in orbit around an atom of a particular element is called the **atomic number** of that element. The electrons are arranged in quantum-mechanical orbits or shells. There is a maximum number of electrons each shell can hold. We can thus classify elements according to the number of electrons in the outermost shell. Such a classification is called a **periodic table** of the elements. A simplified periodic table showing the elements with which we will be concerned is given in Table 3.1.

Hydrogen, the element with atomic number 1, has one lone proton for a nucleus, and one electron in the innermost shell. It is a Group I element; the group number refers to the number of electrons in the outermost shell. The first shell can contain at most two electrons. Helium, a Group Zero element, has two electrons in its inner shell, which is therefore full. After one shell is full, additional electrons are forced to populate the next larger shell. The element lithium, atomic number 3, has the first shell full with two electrons, and one electron in the second

TABLE 3.1 Simplified periodic table of the elements, showing the valence and position of elements that form semiconductor crystals. The Group IV elements shown form a diamond lattice. Silicon is by far the most commonly used semiconductor. Boron, aluminum, and gallium are acceptor impurities in silicon; phosphorus and arsenic are donors. In addition, Group III elements can combine with Group V elements to form diamondlike crystals in which alternate lattice sites are occupied by atoms of each element. The best known of these Group III–V semiconductors is gallium arsenide, which is used for microwave transistors and light-emitting diodes. Group II–VI crystals also are semiconductors. Zinc sulfide is a common phosphor in television display tubes, and cadmium sulfide was the earliest widely used photosensitive material.

I	II	III	IV	V	VI	VII	Zero
H							He
Li	Be	B	C	N	O	F	Ne
Na	Mg	Al	Si	P	S	Cl	Ar
K	Zn	Ga	Ge	As	Se	Br	Kr
Rb	Cd	In	Sn	Sb	Te	I	Xe

shell. It is therefore a Group I element. The second, third, and fourth shells can each hold eight electrons. Neon, with atomic number 10, has both its first and second shells full.

In the world of atoms, having a full outer shell is a happy circumstance. So happy, in fact, that atoms fortunate enough to attain this condition have absolutely no desire to interact with other atoms. Such elements as helium, neon, and argon are **inert gases**, the ultimate snobs of the atomic pecking order. Other, less fortunate atoms strive mightily to emulate their austere brethren by forming alliances with similar atoms. These alliances, called **covalent bonds**, are based on sharing electrons such that all parties to the charade can make believe they have a full shell, even though some of the electrons that fill the outer shell are shared with neighbors. Small communal aggregates of this sort are called **molecules**. An example is methane, CH_4. Carbon, being a Group IV atom, has four electrons with which to play, but is desperately seeking four more to fill its outer shell. Each hydrogen has only one electron with which to play, but needs only one more to fill its outer shell. The ultimately blissful arrangement results when each of the four hydrogens shares its electron with the carbon, and the carbon shares one electron with each hydrogen. Not quite neon, but not too bad!

Crystals

Molecules can satisfy a social need for a few atoms, but for regimentation on a massive scale there is no substitute for a **crystal**. In the simplest crystal, every atom is an equivalent member of a vast army, arrayed in three dimensions as far as the eye can see. Three elements in Group IV of the periodic table crystallize naturally into a remarkable structure called the **diamond lattice**: carbon, silicon, and germanium. Each atom is covalently bonded to four neighbors arranged at the corners of a regular tetrahedron. Group IV is unique in chemistry. Eight electrons are required to complete an atomic shell. Atoms with four electrons can team up with four neighbors, sharing one electron with each. Such an arrangement fills the shell for everyone; no electrons are left over, and no bond is missing an electron. A pure crystal formed this way is called an **intrinsic** semiconductor; it is an electrical insulator, because there are no charged particles free to move around and to carry current.

Conduction

If we alter the crystal by replacing a small fraction of its atoms with impurity atoms of Group V, the crystal becomes conductive. The addition of impurities is called **doping**. Group V elements have one *more* electron than the four needed for the covalent bonds. This extra electron is only weakly bound to the impurity site in the lattice; at room temperature, it is free to move about and to carry current. Such atoms are called **donors**, because they donate a free electron to the crystal. The free electrons are negative, and a semiconductor crystal doped

with donors is said to be **n-type**. The entire crystal is charge neutral, because it is made of atoms that have as many positively charged protons in their nuclei as they have electrons in their shells. When an electron leaves its donor, the donor is said to be **ionized**. An ionized donor has a positive charge because it has lost one electron.

It also is possible to dope a Group IV semiconductor with atoms of Group III. These impurities have one *less* electron than is required for the four covalent bonds. Group III dopants are therefore called **acceptors**. The absence of one electron in a bond is called a **hole**. We can think of the hole as mobile at room temperature, moving about the crystal. (It is actually electrons that move, and the "motion" of the hole is in the opposite direction. We think of the hole as a "bubble" in the electronic fluid.) The hole, being the absence of an electron, carries a *positive* charge. Once the hole has been filled, the acceptor acquires a negative charge, and is said to be **ionized**. Doping a semiconductor with acceptors renders it conductive, the current being carried by positive holes. Such a crystal is called **p-type**.

It is thus possible to provide either positive or negative charge carriers by doping the crystal appropriately. The concentration of dopants can be controlled precisely over many orders of magnitude, from lightly doped (approximately 10^{15} atoms per cubic centimeter) to heavily doped (approximately 10^{19} atoms per cubic centimeter). Heavily doped n-type material is called $n+$, and heavily doped p-type material is called $p+$. The density of impurity atoms is always small compared to the approximately 5×10^{22} atoms per cubic centimeter in the crystal itself.

Because electrons and holes are both charged, and are both used to carry current, we refer to them generically as **charge carriers**.

MOS TRANSISTORS

A cross-section of the simplest transistor structure is illustrated in Figure 3.4. It shows an intrinsic substrate into which two highly doped regions have been fabricated. Consistent with our hydraulic analogy, one of the highly doped regions is called the **source**, and the other is called the **drain**. The entire surface is covered with a very thin layer of SiO_2 (quartz), which is an excellent electrical insulator. On top of the insulator is a metallic control electrode, called the **gate**, that spans the intrinsic region between source and drain. Current flows from source to drain in the region just under the gate oxide called the **channel**. This structure was first described by a freelance inventor named Lilienfeld in a patent issued in 1933 [Lilienfeld, 1928]. It is called an MOS transistor because the active region consists of a metallic gate, an oxide insulator, and a semiconductor channel. In today's technology, the metallic gate often is made of heavily doped polycrystalline silicon, called **polysilicon** or **poly** for short. The details of how transistors are fabricated are the subject of Appendix A. At this point, we will consider the electrical operation of such a transistor.

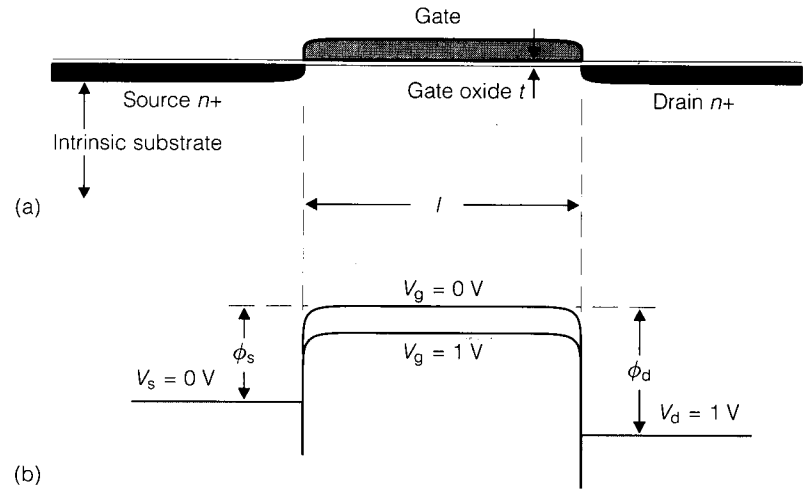

FIGURE 3.4 Cross-section (a) and energy diagram (b) of an *n*-channel transistor. In a typical 1988 process, the gate-oxide thickness is approximately 400 angstroms (0.04 micron), and the minimum channel length *l* is approximately 1.5 microns. When the circuit is in operation, the drain is biased positively; hence, the barrier for electrons is greater at the drain than at the source. Applying a positive voltage at the gate lowers the electron barrier at both source and drain, allowing electrons to diffuse from source to drain.

In Figure 3.4, an energy diagram of the cross-section of part (a) is shown in part (b). The energy of the charge carrier is plotted upward. The surface of the electronic "fluid" is called the **Fermi level**. The drain Fermi level is lower than that of the source by qV_{ds}, where V_{ds} is the drain voltage relative to the source, and q is the electronic charge.

The gate oxide is very thin compared to the source–drain spacing. In a typical 1988 process, the source–drain spacing l is 1.5 microns and the oxide thickness t is 400 angstroms (100 angstroms is equal to 0.01 micron). For this reason, the potential in the channel at the surface of the silicon, just under the gate, is dominated by the gate voltage. There are fringe effects of the source and drain voltages, which will be discussed in Appendix B. For the moment, we will ignore those effects. As a first approximation, then, we assume that any change in the potential on the gate will be reflected in an equal change in potential in the channel. As the gate potential is lowered, the lowest potential will be at the silicon surface. As this barrier between source and drain is lowered, more charges can flow along the surface from source to drain. The barrier for two gate voltages is shown in Figure 3.4(b). The device is exactly analogous to Dr. Field's hydraulic transistor, described at the beginning of the chapter.

The barrier ϕ_s from source to channel is lower than ϕ_d from drain to channel, so more charges will be able to surmount ϕ_s than can climb over ϕ_d. The charge-

carrier density at the source end of the channel thus will be larger than that at the drain end of the channel. Current flows through the channel by diffusion, from the region of high density at the source to that of low density at the drain. We will now compute the channel current.

The carrier density N_s at the source end of the channel is given by Equation 2.16 (p. 25):

$$N_s = N_0 e^{-\frac{\phi_s}{kT}} \tag{3.4}$$

where N_0 is the carrier density at the Fermi level. A similar relation holds for N_d, the carrier density at the drain end of the channel:

$$N_d = N_0 e^{-\frac{\phi_d}{kT}} \tag{3.5}$$

When the transistor was fabricated, there was a built-in barrier ϕ_0 between source and channel. The control of this barrier is the most crucial element of a high-quality processing line. As the gate potential is lowered, the barrier will be lowered accordingly:

$$\phi_s = \phi_0 + q(V_g - V_s) \tag{3.6}$$

$$\phi_d = \phi_0 + q(V_g - V_d) \tag{3.7}$$

We can now write the barrier energies in Equations 3.4 and 3.5 in terms of the source, gate, and drain voltages:

$$N_s = N_0 e^{-\frac{\phi_0 + q(V_g - V_s)}{kT}} \tag{3.8}$$

$$N_d = N_0 e^{-\frac{\phi_0 + q(V_g - V_d)}{kT}} \tag{3.9}$$

From Equations 3.8 and 3.9, we can compute the gradient of carrier density with respect to the distance z along the channel (z is equal to zero at the source). Because no carriers are lost as they travel from source to drain, the current is the same at any z, and the gradient will not depend on z. The density thus will be a linear function of z:

$$\frac{dN}{dz} = \frac{N_d - N_s}{l} = \frac{N_1}{l} e^{-\frac{qV_g}{kT}} \left(e^{\frac{qV_d}{kT}} - e^{\frac{qV_s}{kT}} \right) \tag{3.10}$$

where $N_1 = N_0 e^{-\phi_0/(kT)}$.

The electrical current is just the total number of charges times the average diffusion velocity given in Equation 2.10 (p. 23). The current per unit channel width w is thus

$$\frac{I}{w} = qNv_{diff} = -qD\frac{dN}{dz} \tag{3.11}$$

where D is the diffusion constant of carriers in the channel. Substituting Equation 3.10 into Equation 3.11, we obtain the general form of the MOS transistor

current:

$$I = I_0 e^{-\frac{qV_g}{kT}} \left(e^{\frac{qV_s}{kT}} - e^{\frac{qV_d}{kT}} \right) \tag{3.12}$$

The accumulated preexponential constants have been absorbed into one giant constant I_0.

For a transistor with its source connected to the power supply rail, V_s is equal to zero, and Equation 3.12 becomes:

$$I = I_0 e^{-\frac{qV_{gs}}{kT}} \left(1 - e^{\frac{qV_{ds}}{kT}} \right) \tag{3.13}$$

where V_{gs} and V_{ds} are the gate-to-source and drain-to-source voltages, respectively.

Because there are charge carriers with positive as well as negative charge, there are two kinds of MOS transistor: Those using electrons as their charge carriers are called **n-channel**, whereas those using holes are called **p-channel**; the technology is thus called **complementary MOS**, or CMOS. For positive q (p-channel device), the current increases as the gate voltage is made negative with respect to the source; for negative q (n-channel device), the opposite occurs. kT is the thermal energy per charge carrier, so the quantity kT/q has the units of potential; it is called the **thermal voltage**, and its magnitude is equal to 0.025 volt at room temperature. A carrier must slide down a potential barrier of kT/q to raise its energy by kT. As we noted in Chapter 2, electrochemists and biologists write RT/F in place of kT/q. The way it is written does not change its value.

We have made a number of simplifying assumptions, which will be addressed in Appendix B. Equation 3.12, however, captures all the essential quantitative principles of transistor operation. Notice that, in Equation 3.10, the roles of the source and of the drain are completely symmetrical; therefore, if we interchange them, the magnitude of the current given by Equation 3.12 is identical, with the current flowing in the opposite direction.

The energy diagrams for both types of transistors are identical to that shown in Figure 3.4, but the energy axis has a different meaning. For a p-channel device, upward means higher energy for positive charges, or positive voltage. For an n-channel device, upward means higher energy for negative charges, or negative voltage. Opposite charges attract. An n-channel device requires positive gate voltages to attract negative electrons out of its source into its channel; a p-channel device requires negative gate voltages to attract positive holes out of its source into its channel.

CIRCUIT PROPERTIES OF TRANSISTORS

The symbols used in schematic diagrams for both n- and p-channel transistors are given in Figure 3.5, which shows the source, gate, and drain terminals. We put a bubble on the gate of the p-channel symbol to remind us that the

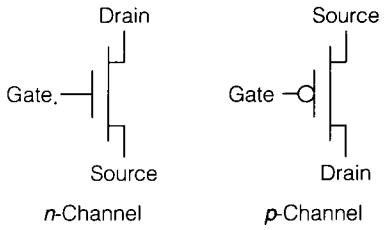

FIGURE 3.5 Circuit symbols for n- and p-channel transistors. These abstractions are used in schematic diagrams of transistor circuits. Note that the assignment of source and drain depends on the voltages to which these terminals are connected, because the physical device structure is symmetrical.

transistor turns on as we make the gate more negative relative to the source. We normally will draw the positive supply at the top of the diagram, and the most negative supply at the bottom. For this reason, the *sources* of p-channel devices usually are located at the *top*, whereas those of n-channel devices normally are at the *bottom*. Implicit in the schematic is a shadow of the Boltzian landscape, with upward meaning positive voltage, as in the energy diagram for p-channel transistors. Positive current (the flow of positive charges) is from high to low. Ground, the reference level, is the most negative supply, or sea level, for positive charges.

The measured current–voltage characteristics of a typical transistor are shown in Figure 3.6. The drain current is zero for $V_{ds} = 0$, as expected. For a given gate voltage, the drain current increases with V_{ds} and then saturates after a few kT/q, as predicted by Equation 3.13. The current in the flat part of the curves is nearly independent of V_{ds} and is called the **saturation current**, I_{sat}. A plot of the saturation current as a function of gate voltage V_{gs} is shown in Figure 3.7. I_{sat} increases exponentially with V_{gs} as predicted by Equation 3.13,

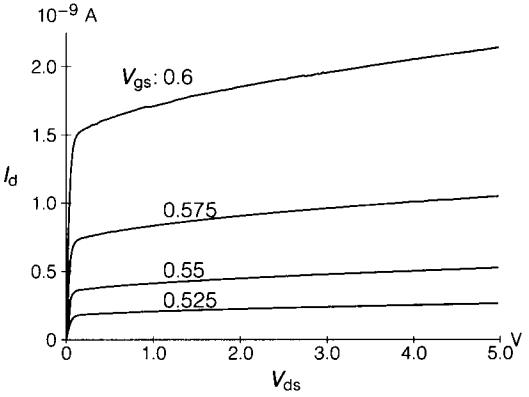

FIGURE 3.6 Drain current as a function of drain–source voltage for several values of gate–source voltage. The channel length of this transistor is 6 microns. The drain current increases rapidly with drain voltage, and saturates within a few kT/q to a gently sloping region of nearly constant current. The slope in the saturation region is due to the change in channel length with drain voltage. The slope in this region is proportional to drain current.

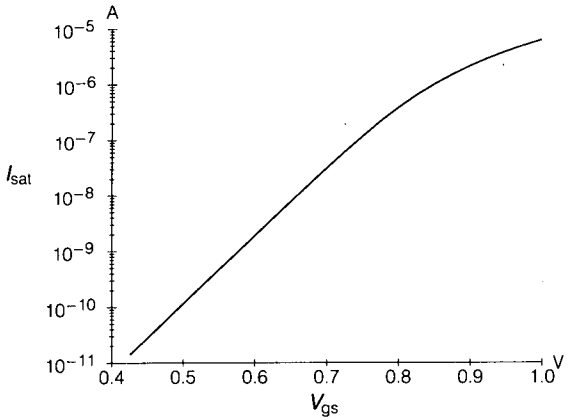

FIGURE 3.7 Saturation current of the transistor of Figure 3.6 as a function of gate voltage. The drain voltage was fixed at 2 volts. The current is an exponential function of gate voltage over many orders of magnitude. In this region, the current increases by a factor of e every 37 millivolts, corresponding to $\kappa = 0.676$. Above approximately 0.8 volt, the limiting effects of mobile charge on the current are evident. The nominal "threshold" of the device is approximately 0.9 volt.

but the voltage required for a factor of e increase in I_{sat} is 37 millivolts, rather than the 25 millivolts we expected.

In our simplified derivation, we assumed that the gate voltage was 100-percent effective in reducing the barrier potential. This assumption is valid for a structure built on an intrinsic substrate, as shown in Figure 3.4. Real transistors, such as those from which the data of Figure 3.7 were taken, are not built on intrinsic substrates. The n-channel transistors are fabricated on p-type substrates, and vice versa. Charges from the ionized donors or acceptors in the substrate under the channel reduce the effectiveness of the gate at controlling the barrier energy (see Appendix B). For our purposes, the effect can be taken into account by replacing kT/q by $kT/(q\kappa)$ in the gate term in Equation 3.12, and rescaling I_0. Equation 3.12 can thus be written

$$I = I_0 e^{-\frac{q\kappa V_g}{kT}} \left(e^{\frac{qV_s}{kT}} - e^{\frac{qV_d}{kT}} \right) \tag{3.14}$$

The transistor shown has a value of κ approximately equal to 0.7. The values of κ can vary considerably among processes, but are reasonably constant among transistors in a single fabrication batch. Throughout this book, we will treat kT/q as the unit of voltage. Because this quantity appears in nearly every expression, we have developed a shorthand notation for it. If the magnitude of kT/q is used to scale all voltages in an expression—as, for example, when a voltage appears as the argument of an exponential—we often will write the voltage as though it were dimensionless. Using this notation, we can write Equation 3.14 for an n-channel transistor as follows:

$$I = I_0 e^{\kappa V_g} \left(e^{-V_s} - e^{-V_d} \right) = I_{sat} \left(1 - e^{-V_{ds}} \right) \tag{3.15}$$

Where V_{ds} is the drain–source voltage. For a p-channel device, the signs of all voltages are reversed.

At the upper end of the current range of Figure 3.7, the current increases less rapidly than does the exponential predicted by Equation 3.15. This deviation from the exponential behavior occurs when the charge on the mobile carriers becomes comparable to the total charge on the gate. The gate voltage at which the mobile charge begins to limit the flow of current is called the **threshold voltage**. For gate voltages higher than threshold, the saturation current increases as the square of the gate voltage. Most circuits described in this book operate in **subthreshold**—their gate voltages are well below the threshold voltage. Typical digital circuits operate well above threshold. The detailed model described in Appendix B describes transistor characteristics over the entire range of operation.

Subthreshold operation has many advantages, three of which we are now in a position to appreciate:

1. Power dissipation is extremely low—from 10^{-12} to 10^{-6} watt for a typical circuit
2. The drain current saturates in a few kT/q, allowing the transistor to operate as a current source over most of the voltage range from near ground to V_{DD}
3. The exponential nonlinearity is an ideal computation primitive for many applications

We will encounter many more beneficial properties, and some limitations, of subthreshold operation as we proceed.

CURRENT MIRRORS

An often-used circuit configuration is shown in Figure 3.8. Here, each transistor is **diode-connected**; that is, its gate is connected to its drain. For a typical process, values of I_0 are such that even the smallest drain current used (10^{-12} amp) requires V_{gs} approximately equal to 0.4 volt. Higher drain currents

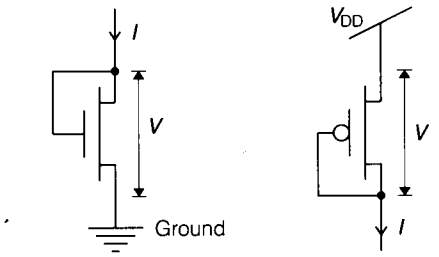

FIGURE 3.8 Diode connected n- and p-channel transistors. Because the drain–source voltage is always a few hundred millivolts, a device in this configuration is guaranteed to be saturated. The current–voltage characteristic is thus exponential, like that of Figure 3.7.

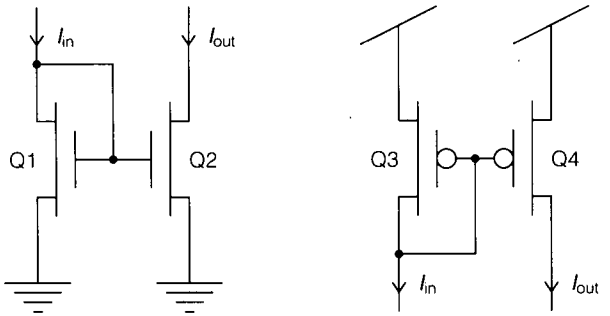

FIGURE 3.9 Current mirror connected *n*- and *p*-channel transistors. This configuration turns a source of current (the input) into an equal *sink* of current (the output). Reflecting currents in this manner is the most common operation in analog circuit design.

require higher values. Thus, the drain curves of Figure 3.6 are well into saturation for any useful V_{gs}. For this reason, the current through these diode-connected transistors has the same exponential dependence on voltage as that shown for the saturation current in Figure 3.7.

It is common to have a current of a certain sign—for example, a source of positive charges—and an input that requires an equal but opposite current—for example, a source of negative charges. A simple circuit that performs this inversion of current polarity is shown in Figure 3.9. The input current I_{in} biases a diode-connected transistor Q1. The resulting V_{gs} is just sufficient to bias the second transistor Q2 to a saturation current I_{out} equal to I_{in}. The value of I_{out} will be nearly independent of the drain voltage of Q2 as long as Q2 stays in saturation. A similar arrangement is shown for currents of the opposite sign using *p*-channel transistors Q3 and Q4. The *p*-channel circuit *reflects* a current to ground into a current from V_{DD}, so it is called a **current mirror**. The *n*-channel current mirror reflects a current from V_{DD} into a current to ground. We will use these circuit configurations in nearly every example in this book.

SUMMARY

We have seen how we can construct a physical structure that allows a voltage on one terminal to control the flow of current into another terminal. Although the first proposal for a device of this type was made in the 1930s [Lilienfeld, 1926], it took 3 decades to reduce the ideas to a production process. By the 1960s, MOS technology came into its own—today, it is the major technology on which the computer revolution has been built. For our purposes, MOS transistors are controlled sources of both positive and negative current. Their control terminals do not draw current from the nodes to which they are connected. MOS transistors are, in that sense, the most ideal active devices extant. The exponential dependence of drain current on gate voltage allows us to control current levels over

many orders of magnitude. We will develop increasingly complex configurations of these simple elements, culminating in complete neural subsystems for vision and hearing.

REFERENCES

Lilienfeld, J.E. Method and apparatus for controlling electric currents. Patent (1,745,175: January 28, 1930): October 8, 1926.

Lilienfeld, J.E. Device for controlling electric current. Patent (1,900,018: March 7, 1933): March 28, 1928.