# 11 Semiconductor Materials and Devices

This chapter is the heart of the book. We've learned about how physical phenomena can represent and communicate information, and will learn about how it can be input, stored, and output, but here we turn to the essential electronic devices that transform it. To understand these devices we need to understand how electrons behave in materials. We will start by reviewing the statistical mechanics of quantum systems, and then solve a simplified model of a particle in a periodic potential to introduce the idea of band structure. Based on this quantization of the available electronic states we will study junction and field-effect semiconductor devices, leading up to an introduction to digital logic. The chapter will close by considering some of the fundamental physical limits on making and using these devices.

## 11.1 QUANTUM STATISTICAL MECHANICS

When statistical mechanics was introduced in Section 3.4 we did not worry about the role of quantum mechanics. Now we must: quantum mechanics is essential in explaining the states available to electrons in a semiconductor. The statistics will need to account for the allowed occupancy, and the variable number of electrons as a function of external fields or internal doping.

Remember that statistical mechanical distributions are found by maximizing their entropy subject to external constraints expressed by Lagrange multipliers. In equation (3.60) we saw that fixing the average energy

$$\sum_i E_i p_i = E \tag{11.1}$$

introduced the temperature $\beta = 1/kT$ and gave the partition function for the canonical distribution

$$\mathcal{Z} = \sum_i e^{-\beta E_i} \quad . \tag{11.2}$$

Varying the number of particles introduces another constraint,

$$\sum_i N_i p_i = N \quad , \tag{11.3}$$

where the sum is over the possible number of particles $N_i$ in each state, and $N$ is the expected total number of particles in the system. Adding this to equation (3.49) and

repeating the derivation of the maximum entropy distribution gives the partition function for the *grand canonical distribution*

$$\mathcal{Z} = \sum_i e^{-\beta(E_i - \mu N_i)} \quad , \tag{11.4}$$

where the sum is now over the available states and their possible occupancies and $E_i$ is the total energy of a given configuration. The expected value of any quantity $f_i$ that depends on the state of the system is then found from

$$\langle f \rangle = \frac{1}{\mathcal{Z}} \sum_i f_i \; e^{-\beta(E_i - \mu N_i)} \quad . \tag{11.5}$$

The new parameter that arises from the constraint on the particle number is $\mu$, the *chemical potential*. Once again comparing the microscopic and macroscopic thermodynamic definitions shows that the chemical potential is equal to the rate at which the free energy $F$ grows as particles are added to the sytem [Reif, 1965],

$$\mu = \frac{\partial F}{\partial N} \quad . \tag{11.6}$$

Remember that there are two kinds of quantum systems: *fermions*, which have 1/2-integer spin and which cannot be in the same state because of the *Pauli exclusion principle*, and *bosons*, which have integer spin and can share the same state. Since electrons are fermions, here we will need to solve equation (11.5) subject to the condition that each state can have only one electron. When we consider photons, which are bosons, in Chapter 12 we'll want solutions that allow an unlimited number of particles per state.

The most important statistical quantity will be the expected occupancy of the available quantum states as a function of their energy and the temperature. If there are $N_i$ particles in the $i$th quantum state, with a single-particle energy $E_i$, we'll neglect interactions and take the total energy to be $E_i N_i$. Then the expected number of particles in state $s$ is found by summing over all possible configurations,

$$\begin{aligned} \langle N_s \rangle &= \frac{\sum_{N_1} \sum_{N_2} \cdots \; N_s e^{-\beta(E_1 N_1 + E_2 N_2 + \cdots) + \beta\mu(N_1 + N_2 + \cdots)}}{\sum_{N_1} \sum_{N_2} \cdots \; e^{-\beta(E_1 N_1 + E_2 N_2 + \cdots) + \beta\mu(N_1 + N_2 + \cdots)}} \\ &= \frac{\sum_{N_s} N_s e^{-\beta E_s N_s + \beta\mu N_s}}{\sum_{N_s} e^{-\beta E_s N_s + \beta\mu N_s}} \quad . \end{aligned} \tag{11.7}$$

The sum over $N_s$ can be pulled out of the other sums, and except for it the numerator and denominator cancel. Equation (11.7) can now be evaluted for the two cases of interest.

- *Fermions*
  Here $N_s$ must be either 0 or 1 because there can't be more than one fermion in a single state:

$$\begin{aligned} \langle N_s \rangle &= \frac{\sum_{N_s=0}^{1} N_s e^{-\beta E_s N_s + \beta\mu N_s}}{\sum_{N_s=0}^{1} e^{-\beta E_s N_s + \beta\mu N_s}} \\ &= \frac{0 + e^{-\beta E_s + \beta\mu}}{1 + e^{-\beta E_s + \beta\mu}} \\ &= \frac{1}{e^{\beta(E_s - \mu)} + 1} \quad . \end{aligned} \tag{11.8}$$

This is called the *Fermi–Dirac distribution.*

- *Bosons*
  For bosons, the sum over $N_s$ runs from 0 to $\infty$:

$$\begin{aligned}
\langle N_s \rangle &= \frac{\sum_{N_s=0}^{\infty} N_s e^{-\beta E_s N_s + \beta\mu N_s}}{\sum_{N_s=0}^{\infty} e^{-\beta E_s N_s + \beta\mu N_s}} \\
&= \frac{\sum_{N_s=0}^{\infty} N_s C^{N_s}}{\sum_{N_s=0}^{\infty} C^{N_s}} \qquad (C \equiv e^{-\beta E_s + \beta\mu}) \\
&= \frac{C\frac{d}{dC}\sum_{N_s=0}^{\infty} C^{N_s}}{\sum_{N_s=0}^{\infty} C^{N_s}} \\
&= \frac{C\frac{d}{dC}(1-C)^{-1}}{(1-C)^{-1}} \\
&= \frac{C(1-C)^{-2}}{(1-C)^{-1}} \\
&= \frac{1}{C^{-1}-1} \\
&= \frac{1}{e^{\beta(E_s-\mu)}-1} \quad . \qquad (11.9)
\end{aligned}$$

  This is the *Bose–Einstein* distribution. It differs from the Fermi–Dirac distribution only by the minus sign in the denominator, but we will see that this sign difference makes all the difference in the world.

## 11.2 ELECTRONIC STRUCTURE

In an atom that is part of a crystal, there are *core* electrons that remain tightly bound to the nucleus unless they are knocked out by energetic particles such as photons in *X-ray Photoemission Spectroscopy* (*XPS*) or electrons in *Auger spectroscopy*. There are also less weakly bound *outer-shell* electrons that can move around the crystal. A surprisingly good approximation is to ignore the inner-shell electrons, and treat the outer-shell electrons as a non-interacting ideal gas of fermions traveling in a medium with a spatially periodic potential due to the atoms on the crystal lattice sites [Ashcroft & Mermin, 1976]. From this simple and historically important model we'll find that there are momentum bands of allowed and forbidden electron energies which we will then use to explain the basic features of semiconductors. This is a physicists' approach; similar but complementary insights come from viewing band structure as resulting from the electronic states available in a system with many bonds [Hoffmann, 1988].

Here we'll introduce just enough quantum mechanics to analyze a one-dimensional model of an electron in a periodic potential, saving the general structure of quantum mechanics for Chapter 16. The spatial state of the electron will be described by its *wave function* $\psi(x)$, with the probability of seeing the electron at a point given by $|\psi(x)|^2$. Since the electron must be somewhere the wave function is normalized,

$$\int_{-\infty}^{\infty} |\psi(x)|^2 \, dx = 1 \quad . \qquad (11.10)$$

Measurable quantities are given by differential operators that act on the wave function; the most important one being the total energy, called the *Hamiltonian*

$$\mathcal{H}[\psi(x)] = \left[-\frac{\hbar^2}{2m}\frac{d^2}{dx^2} + V(x)\right]\psi(x) \quad , \tag{11.11}$$

where $V(x)$ is the electron's potential energy, and $(-\hbar^2/2m)\ d^2/dx^2$ is the operator corresponding to the kinetic energy. The expected value of the energy associated with the wave function is

$$\langle\psi|\mathcal{H}|\psi\rangle = \int_{-\infty}^{\infty}\psi^*(x)\mathcal{H}[\psi(x)]\ dx \quad . \tag{11.12}$$

The evolution of a wave function is given by the *time-dependent Schrödinger equation*

$$\mathcal{H}[\psi(x)] = i\hbar\frac{\partial\psi}{\partial t} \quad . \tag{11.13}$$

If the Hamiltonian does not depend on time, the allowable energy states are given by wave functions that satisfy the *time-independent Schrödinger equation*

$$\mathcal{H}[\psi_E(x)] = E\psi_E(x) \quad , \tag{11.14}$$

where the $\psi_E(x)$ are the *eigenfunctions* of the Hamiltonian, and the possible values of $E$ are the corresponding *eigenvalues*. If the potential vanishes, this is easily solved to find for free space

$$-\frac{\hbar^2}{2m}\frac{d^2\psi_E(x)}{dx^2} = E\psi_e(x) \quad \Rightarrow \quad \psi(x) = Ae^{ikx} + Be^{-ikx} \quad , \tag{11.15}$$

where $A$ and $B$ are constants that depend on the boundary conditions, and the energy $E$ and wave vector $k$ are related by $E = \hbar^2k^2/2m$.

In Chapter 16 we'll see that if another operator $O$ *commutes* with the Hamiltonian so that

$$\mathcal{H}\{O[\psi(x)]\} = O\{\mathcal{H}[\psi(x)]\} \quad , \tag{11.16}$$

then wave functions can be chosen to simultaneously be eigenfunctions of both operators. In particular, if the potential is periodic so that $V(x + \Delta) = V(\Delta)$, the Hamiltonian will be unchanged by an operator $T_\Delta[\psi(x)] = \psi(x + \Delta)$ that translates the wavefunction by this distance, so $T_\Delta$ commutes with $H$. This means that an energy eigenfunction will also be an eigenfunction of the translation operator, with an eigenvalue $\lambda$ that can depend on the energy

$$T_\Delta[\psi(x)] = \lambda_\Delta(E)\psi(x) \quad . \tag{11.17}$$

If we compose two translations by two multiples of the period of the potential,

$$T_{\Delta'}\{T_\Delta[\psi(x)]\} = \lambda_{\Delta'}\lambda_\Delta\psi(x) \quad , \tag{11.18}$$

but by definition the translations add so that

$$T_{\Delta'}\{T_\Delta[\psi(x)]\} = \lambda_{\Delta'+\Delta}\psi(x) \quad . \tag{11.19}$$

These equations will be consistent if $\lambda_\Delta(E) = e^{\alpha(E)\Delta}$, where $\alpha$ is a constant that depends

on $E$. A further constraint comes from requiring that after translation the wave function remain normalized

$$\int_{-\infty}^{\infty} |T_\Delta[\psi(x)]|^2 \, dx = \int_{-\infty}^{\infty} \left| e^{\alpha\Delta}\psi(x) \right|^2 dx \quad \Rightarrow \quad |e^{\alpha\Delta}|^2 = 1 \quad . \tag{11.20}$$

This means that the exponent is imaginary, so the translation operator is a multiplication by a phase factor that depends on the energy $T_\Delta = e^{ik(E)\Delta}$, or

$$\psi_E(x + \Delta) = e^{ik(E)\Delta}\psi_E(x) \quad . \tag{11.21}$$

If we write the wave function as a product of two terms $\psi(x) = A_k(x)u_k(x)$, where $u_k(x + \Delta) = u_k(x)$ reflects the periodicity of the potential, then

$$\begin{aligned} \psi(x + \Delta) &= A_k(x + \Delta)u_k(x + \Delta) \\ &= A_k(x + \Delta)u_k(x) \\ &= T_\Delta[\psi_k(x)] \\ &= e^{ik\Delta}A_k(x)u_k(x) \quad , \end{aligned} \tag{11.22}$$

which will hold if $A_k(x) = e^{ikx}$, therefore

$$\psi(x) = e^{ikx}u_k(x) \quad . \tag{11.23}$$

This is *Bloch's Theorem*.

The simplest periodic potential ignores the size of the atoms and models them as a sum of delta functions,

$$V(x) = \sum_{n=-\infty}^{\infty} V_0 \, \delta(x - n\Delta) \quad , \tag{11.24}$$

called the *Kronig–Penney model*. The atoms will be assumed to be fixed; in a real crystal there are quantized displacement waves called *phonons* [Ashcroft & Mermin, 1976]. In the intervals between these idealized atoms the potential vanishes, therefore the wave function there is just that of a free plane wave

$$\begin{aligned} \psi(x) &= Ae^{iqx} + Be^{-iqx} \\ \Rightarrow u_k(x) &= Ae^{i(q-k)x} + Be^{-i(q+k)x} \quad , \end{aligned} \tag{11.25}$$

where $q\hbar = \sqrt{2mE}$, and $A$ and $B$ are unknown constants.

We will now find how $q$ and hence $E$ are related to the $k$ that was introduced by Bloch's Theorem. Requiring that $u_k(0) = u_k(\Delta)$ gives

$$A + B = Ae^{i(q-k)\Delta} + Be^{-i(q+k)\Delta} \quad . \tag{11.26}$$

A second relationship comes from Schrödinger's equation

$$\begin{aligned} \left[ -\frac{\hbar^2}{2m}\frac{d^2}{dx^2} + V \right] \psi &= E\psi \\ \left[ E + \frac{\hbar^2}{2m}\frac{d^2}{dx^2} \right] \psi(x) &= \sum_{n=-\infty}^{\infty} V_0\delta(x - n\Delta)\psi(x) \quad . \end{aligned} \tag{11.27}$$

The wave function must be continuous across the delta functions, but its slope can change

at them. Integrating from $x = -\epsilon$ to $\epsilon$, taking the limit $\epsilon \to 0$, and using the periodicity of $u$ shows that (Problem 11.1)

$$\frac{\hbar^2}{2m} iq \left( A - B - Ae^{i(q-k)\Delta} + Be^{-i(q+k)\Delta} \right) = V_0(A + B) \quad . \tag{11.28}$$

Now we have two equations in the two unknowns $A$ and $B$. Eliminating $B$ gives

$$\left[ \cos(k\Delta) - \cos(q\Delta) - \frac{mV_0}{q\hbar^2} \sin(q\Delta) \right] A = 0$$

$$\Rightarrow \quad \cos(k\Delta) = \cos(q\Delta) + \frac{mV_0\Delta}{\hbar^2} \frac{\sin(q\Delta)}{q\Delta} \quad . \tag{11.29}$$

In the limit of infinite lattice spacing $\Delta \to \infty$ this reduces to $k = q \Rightarrow E = \hbar^2 k^2/2m$, the free electron case. For non-infinite spacing the relationship is more complicated: since $|\cos(k\Delta)| \leq 1$, there will be bands of $q$ values for which there is no $k$ that solves this equation, and hence gaps in the allowable energy $E = \hbar^2 q^2/2m$. The relationship between $k$ and $E(q)$ is plotted in Figure 11.1, with successive bands shifted back to the origin by multiples of $2\pi/\Delta$ (which can be done without changing the value of $\cos(k\Delta)$). Each of the regions shifted back is called a *Brillouin zone*. The symmetries of a real crystal lead to a much more complicated three-dimensional *band structure*, but the basic features are similar. For a free particle the bands are just sections of a parabola, which are bent by the crystal periodicity near the gaps at the zone boundaries. $k$ is called the *crystal momentum*. It indexes the eigenstates, playing a role that is analogous but no longer equal to the real momentum.

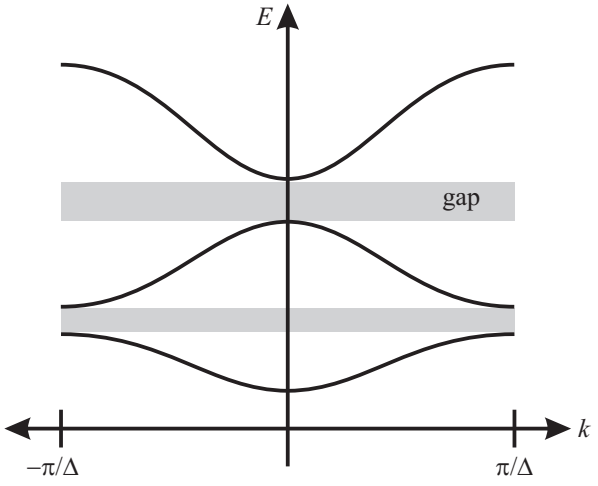


Figure 11.1. Band structure for the Kronig–Penney model.

Next, assume that the crystal has a finite length of $L = N\Delta$, and to avoid end effects assume periodic boundary conditions $\psi(0) = \psi(L)$. This implies that

$$u_k(0) = e^{ikL} u_k(L) \quad \Rightarrow \quad e^{ikL} = 1 \tag{11.30}$$

because of the periodicity of $u_k$. This will hold if

$$kL = 2\pi n \quad \Rightarrow \quad k = \frac{2\pi}{L} n = \frac{2\pi}{N\Delta} n \tag{11.31}$$

for integers $n$. Since these crystal momentum states are separated by a difference of $2\pi/(N\Delta)$ and each band is $2\pi/\Delta$ wide, there are

$$\frac{2\pi}{\Delta} \frac{N\Delta}{2\pi} = N \tag{11.32}$$

states per band. Because electrons are spin-1/2 fermions, each of these momentum states can hold two electrons, one "up" and one "down", and the occupation probability as a function of temperature is given by the Fermi–Dirac distribution

$$f(E) = \frac{1}{1 + e^{(E-\mu)/kT}} \quad . \tag{11.33}$$

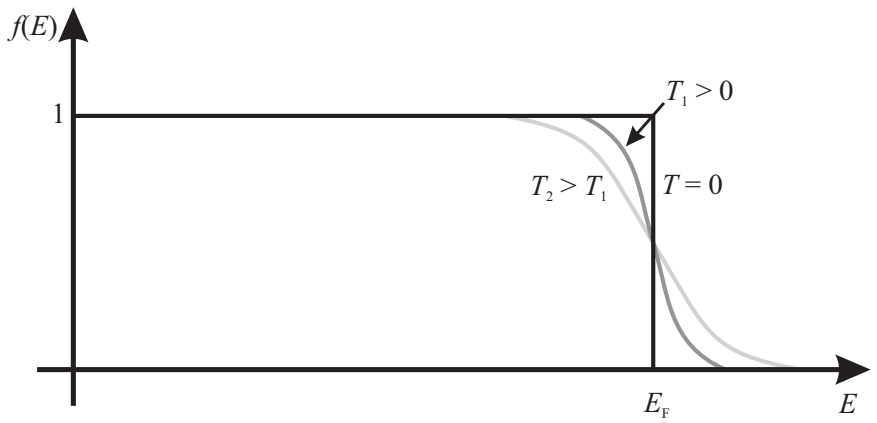


Figure 11.2. The Fermi–Dirac distribution.

The Fermi–Dirac distribution is shown in Figure 11.2. The chemical potential $\mu$ is the change in the free energy when one electron is added. The *Fermi energy* or *Fermi level* $E_F$ is the chemical potential at $T = 0$ K, and if it lies in a band it gives the highest filled state at $T = 0$ K. The Fermi level will be a function of the number of electrons in the crystal and hence the number of states that can be filled. As the temperature is raised, electrons in states below the Fermi energy will be excited above it, which will move the chemical potential relative to the Fermi energy. Because this difference is small at room temperature, we will follow the common (but not quite correct) practice of using them interchangeably.

An applied voltage can move an electron only if there is an electron to be moved, and a state for it to go into. Therefore, for conduction the states far below the Fermi energy don't matter because there are no available nearby states for an electron to move into, and states far above the Fermi energy don't matter because there is little probability of them being occupied. The uppermost filled band is called the *valence band*, and the lowest unfilled band is called the *conduction band*. In an insulator, the valence band is completely full, hence it is not possible for electrons to move unless they are excited

over the band gap. The chemical potential for an insulator lies in the middle of the gap because each carrier excited out of the valence band appears in the conduction band. In a metal the chemical potential lies within the conduction band and so there are plenty of conduction states available (Figure 11.3). The energy difference between a conduction electron and a free electron removed from the metal is called the *work function*.

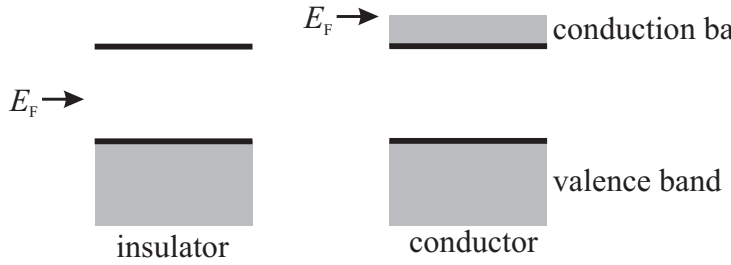


Figure 11.3. Band structure for an insulator and conductor.

A *semiconductor* is just an insulator that has an energy gap small enough for there to be an appreciable probability for an electron to be thermally excited across it at room temperature. For example, Ge, Si, and diamond all have full valence bands, but the gap energy of Ge is 0.67 eV, for Si it is 1.11 eV, and for diamond it is 5 eV ($kT$ at room temperature is 0.026 eV). The room temperature resistivity of a good insulator is $\sim 10^{10}\ \Omega \cdot$ cm, that of a good metal is $\sim 10^{-6}$, and for a semiconductor it is $\sim 10^{6}$.

When a full valence band has a single electron removed it leaves behind one available state. This single state can be viewed as a positively charged particle called a *hole* that moves under an applied field. Actually, all of the electrons in the band are moving in the opposite direction, much like the motion of a bubble trapped in a glass of water. The effective mass $m^*$ of a hole is given by finding the change in its momentum under an applied force, and is equal to $m_p^* = 0.56m_0$ for Si (where $m_0$ is the free electron mass). Electrons in a crystal also have an effective mass different from that of a free electron because of the curvature of the bands; in Si $m_n^* = 1.1m_0$.

The conduction properties of a material depend sensitively on the location of the Fermi energy relative to the nearest energy gap. Adding doping atoms can add or remove electrons, moving the Fermi energy and changing the character of the material. Such *extrinsic* materials are produced by adding *donor* atoms, such as P or As which have one extra outer electron compared to Si, or *acceptor* atoms such as Al which has one less electron. A donor will raise the Fermi level by giving electrons up to the conduction band, and an acceptor will lower the Fermi level by trapping valence band electrons, thereby producing holes. This ability to selectively move the Fermi level relative to the undoped level in the *intrinsic* material is the key to making semiconductor devices. Materials which are doped so that the dominant conduction is by electrons are called *n-type* (n for negative), and those doped by holes *p-type* (Figure 11.4).

The density of carriers $n$ of the conduction band is found by integrating up from the conduction band edge $E_c$ the product of the density of states $N(E)$, which gives the number of available states per volume in an energy range, times the Fermi distribution

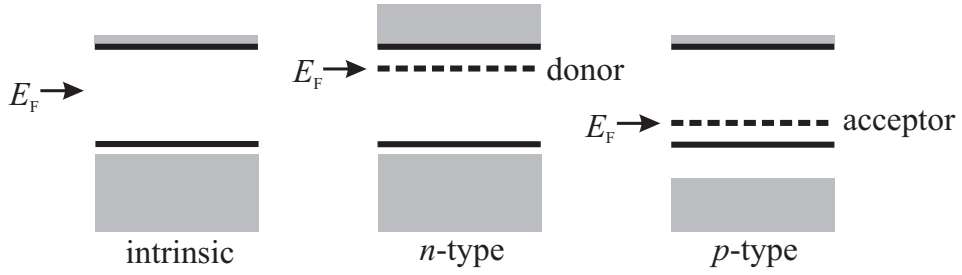


Figure 11.4. Doping of a semiconductor.

$f(E)$, which is their thermodynamic occupancy

$$n = \int_{E_c}^{\infty} f(E)N(E)\,dE \quad . \tag{11.34}$$

Since at room temperature $kT$ is much smaller than gap energies, a good approximation for the Fermi distribution is

$$f(E) = \frac{1}{1 + e^{(E-E_F)/kT}} \simeq e^{-(E-E_F)/kT} \quad . \tag{11.35}$$

The density of states can be approximated with the free-electron distribution. In 3D, equation (11.31) becomes

$$\vec{k} = \frac{2\pi}{L}\left(n_x\hat{x} + n_y\hat{y} + n_z\hat{z}\right) \quad , \tag{11.36}$$

hence

$$\begin{aligned} E &= \frac{\hbar^2}{2m_n^*}|k|^2 \\ &= \frac{\hbar^2}{2m_n^*}\left(\frac{2\pi}{L}\right)^2\left(n_x^2 + n_y^2 + n_z^2\right) \\ &\equiv \frac{\hbar^2}{2m_n^*}\left(\frac{2\pi}{L}\right)^2 r^2 \\ &= \frac{h^2}{2m_n^* V^{2/3}} r^2 \end{aligned} \tag{11.37}$$

or

$$dE = \frac{h^2}{m_n^* V^{2/3}}\, r\, dr \quad . \tag{11.38}$$

$m_n^*$ is the electron's effective mass, $V = L^3$ is the volume of the crystal, and we're assuming that because there are so many electron states, the sum of the squares of the integers $n_x^2 + n_y^2 + n_z^2$ can be taken to be a continuous variable $r^2$. The density of states $N(E)$ can equivalently be taken to be a function of this variable $N(r)$. In $r$-space, states occupy cubes of unit volume, therefore the total number of states in the crystal $VN(r)\,dr$ in an infinitesimal shell around $r$ is given by the area of the shell times its thickness

$$VN(r)\,dr = 2 \cdot 4\pi r^2\,dr \quad , \tag{11.39}$$

with the factor of 2 coming from spin-up and -down occupancy. Substituting in equations (11.37) and (11.38),

$$\begin{aligned} VN(r)\,dr &= 8\pi r^2\,dr \\ &= 8\pi r\cdot r\,dr \\ VN(E)\,dE &= 8\pi\frac{V^{1/3}\sqrt{2m_n^*E}}{h}\frac{V^{2/3}m_n^*}{h^2}\,dE \\ N(E)\,dE &= 8\pi\frac{m_n^{*3/2}\sqrt{2E}}{h^3}\,dE \\ &= \frac{1}{2\pi^2}\left(\frac{2m_n^*}{\hbar^2}\right)^{3/2}\sqrt{E}\,dE \quad . \end{aligned} \tag{11.40}$$

Then

$$\begin{aligned} n &= \int_{E_c}^{\infty} f(E)N(E)\,dE \\ &= \int_{E_c}^{\infty} e^{-(E-E_F)/kT}\frac{1}{2\pi^2}\left(\frac{2m_n^*}{\hbar^2}\right)^{3/2}\sqrt{E}\,dE \\ &= \frac{1}{2\pi^2}\left(\frac{2m_n^*}{\hbar^2}\right)^{3/2}e^{E_F/kT}\int_{E_c}^{\infty} e^{-E/kT}\sqrt{E}\,dE \quad . \end{aligned} \tag{11.41}$$

In doing this integration we're free to choose any energy scale we wish, and so can simplify it by taking $E_c = 0$

$$\begin{aligned} n &= \frac{1}{2\pi^2}\left(\frac{2m_n^*}{\hbar^2}\right)^{3/2}e^{E_F/kT}\int_0^{\infty} e^{-E/kT}\sqrt{E}\,dE \\ &= \frac{1}{2\pi^2}\left(\frac{2m_n^*}{\hbar^2}\right)^{3/2}e^{E_F/kT}\frac{\sqrt{\pi}}{2}(kT)^{3/2} \\ &= 2\left(\frac{m_n^*kT}{2\pi\hbar^2}\right)^{3/2}e^{E_F/kT} \quad . \end{aligned} \tag{11.42}$$

Going back to units in which $E_c \neq 0$ requires subtracting the difference off from $E_F$,

$$\begin{aligned} n &= 2\left(\frac{m_n^*kT}{2\pi\hbar^2}\right)^{3/2}e^{-(E_c-E_F)/kT} \\ &\equiv N_n e^{-(E_c-E_F)/kT} \quad . \end{aligned} \tag{11.43}$$

Similarly, the hole occupancy in the valence band is given by $1 - f(E)$; integrating this from the valence band edge $E_v$ gives the symmetrical relationship

$$\begin{aligned} p &= 2\left(\frac{m_p^*kT}{2\pi\hbar^2}\right)^{3/2}e^{-(E_F-E_v)/kT} \\ &\equiv N_p e^{-(E_F-E_v)/kT} \quad . \end{aligned} \tag{11.44}$$

For an intrinsic semiconductor the Fermi level will be in the middle of the gap, and the hole and electron concentrations will be equal $n = p = n_i$. The Fermi energy will move depending on the doping, but the product of the occupancies will be a constant

that depends only on the gap energy $E_g$:

$$np = N_n N_p e^{-(E_c - E_v)/kT} = N_n N_p e^{-E_g/kT} = n_i^2 \quad . \tag{11.45}$$

The carrier densities can be rewritten in terms of the intrinsic density $n_i$ and the intrinsic Fermi energy $E_i$ as

$$n = n_i e^{(E_F - E_i)/kT} \qquad p = n_i e^{(E_i - E_F)/kT} \quad . \tag{11.46}$$

If there is no doping then these will be equal.

Now consider what happens to one of these electrons in response to the force of an external electric field

$$F = \frac{dp}{dt} = qE \quad . \tag{11.47}$$

In free space this causes a steady increase in the velocity

$$dp = m\, dv = qE\, dt \quad , \tag{11.48}$$

but in a material the collisions with the lattice and defects will slow it down. *Kinetic theory* [Balian, 1991] makes the rough approximation that after a characteristic time $\tau$ a collision occurs which randomizes the electron's velocity, so that the average drift velocity is the expected value of the non-random contribution

$$\langle v \rangle = \frac{q\tau}{m} E \equiv \mu E \quad . \tag{11.49}$$

$\mu$ is the *mobility*. In terms of it, the conductivity is

$$J = nq\langle v \rangle = \sigma E \quad \Rightarrow \quad \sigma = nq\mu \quad . \tag{11.50}$$

This relates Ohm's Law to microscopic material properties. The linear relationship between drift velocity and applied field holds only at sufficiently low fields, limited ultimately by the dielectric breakdown voltage of the material (which $\approx 5{\times}10^5$ V/cm for Si).

For undoped Si at room temperature, the electrons have a mobility of about 1350 $\text{cm}^2/(\text{V}\cdot\text{s})$, and for GaAs it is 8500 $\text{cm}^2/(\text{V}\cdot\text{s})$, which is why GaAs is used for high-speed devices. Note that this is still *much* slower than the propagation velocities that we found for electromagnetic waves. If Si is doped at $10^{17}/\text{cm}^3$ the mobility falls to 800 $\text{cm}^2/(\text{V}\cdot\text{s})$ because of the extra scattering from the dopant atoms, and at $10^{19}/\text{cm}^3$ it is only 90. In a *High-Electron-Mobility Transistor* (*HEMT*) the doping material is confined to layers separated from where conduction takes place [Pavlidis, 1999]. This can be accomplished by using an alloy such as $\text{Al}_x\text{Ga}_{1-x}\text{As}$ that lets the band gap be tuned as a function of the composition $x$, ranging up to a few eV from GaAs to AlAs depending on the crystal orientation. These are examples of binary III–V semiconductors formed from elements in those columns of the periodic table; II–VI semiconductors such as CdSe are also useful, particularly for optoelectronics (Chapter 12).

## 11.3 JUNCTIONS, DIODES, AND TRANSISTORS

Fortified with the basic ideas of energy bands and the Fermi–Dirac distribution, we are now ready to tackle devices made out of semiconductors. These will rely on the properties

of junctions between materials. The Fermi energy (or, to be correct at $T \neq 0$ K, the chemical potential) in a material can be thought of as the height of an energy hill that the electrons are traveling on. If a drop of water is added to a bucket, its energy depends on how much water is already in the bucket. If two buckets filled with differing amounts of water are brought into contact and a partition between them is removed, then water will spill from one bucket to the other in order to equalize the energy difference. Similarly, when two semiconductors are brought into contact as shown in Figure 11.5, there is a transient current from the material with the higher Fermi level to the lower one until the Fermi levels are aligned, removing the energy gradient that is driving the current. A potential difference then appears between the bands that is equal to the potential difference $\Delta V$ between the two Fermi energies. This energy gradient is associated with a local electric field that is produced in the *transition* (or *depletion* or *space-charge*) region by the charge that moved between the materials. Once the bands have "bent" at the interface no average current will flow. Even though electrons and holes will recombine if they are given a chance because that lowers their energy, the electrons on the $n$ side cannot climb up the potential hill to reach the $p$ side. Similarly, holes behave oppositely from electrons, and so they cannot climb down the hill to reach the electrons.

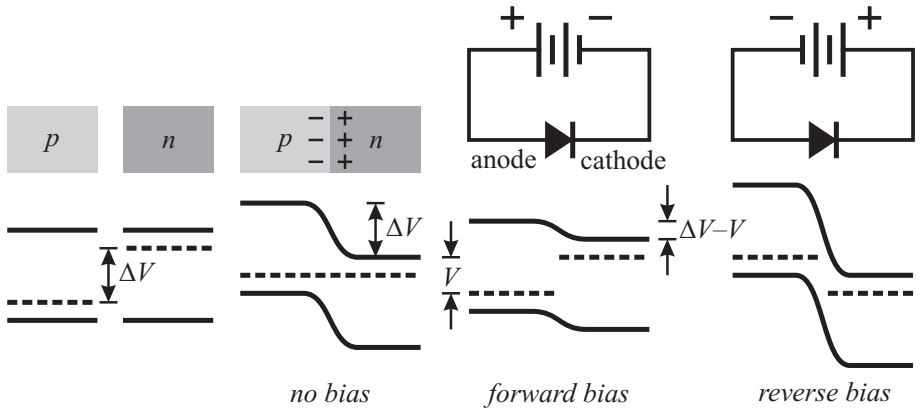


Figure 11.5. Biasing a $p$–$n$ diode.

If an electron is thermally excited from the valence band to the conduction band it leaves a hole behind, creating an *Electron–Hole Pair* (*EHP*). Normally these will quickly recombine, but if they form near the interface then the junction field will sweep the electron to the $n$ side and the hole to the $p$ side. This results in a *generation current*. In addition, there is a probability proportional to $e^{-\Delta E/kT} = e^{-q\Delta V/kT}$ for a carrier to be thermally excited over the energy barrier at the junction and then diffuse out, creating a *diffusion current*.

If a bias potential $V$ is applied across the junction in a *p–n diode*, it will split the Fermi energies, reducing or increasing the size of the barrier depending on the polarity. The diffusion current will then be

$$I_{\text{diffusion}} = Ae^{-q(\Delta V - V)/kT} = Ae^{-q\Delta V/kT}e^{qV/kT} \quad , \tag{11.51}$$

where $A$ is a constant that depends on device details including the junction geometry

and the amount of doping. It can be found by recognizing the the generation current is independent of the bias voltage (until that becomes large enough to eliminate the band bending), and that at zero bias the two currents must cancel, so that their sum is

$$I = I_{\text{generation}} \left( e^{qV/kT} - 1 \right) \quad . \tag{11.52}$$

This characteristic $I$–$V$ curve is shown in Figure 11.6. As the forward bias is increased, the current quickly increases and the diode turns on. The bias voltage appears as a *diode drop* potential in conduction across the diode; 0.6 V is a typical value. The diode blocks current in the opposite direction, letting through just the small generation current that is independent of bias but does depend on temperature (providing a useful thermometer).

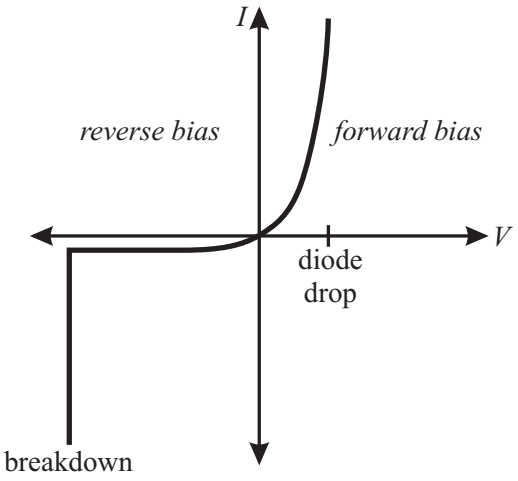


Figure 11.6. $I$–$V$ curve for a $p$–$n$ diode.

This is a DC relationship, which will roll off at higher frequencies because of the capacitance associated with the charge induced at the junction. It also breaks down at higher fields through two important mechanisms. If the bands are bent far enough, the valence and conduction bands come so close at the junction that the wave function for carriers overlaps between them, creating a probability for them to *tunnel* between the bands. Because the voltage at which this *Zener breakdown* occurs can be selected by the doping, it is useful for providing voltage references. And if the field is strong is enough to accelerate a carrier so fast that it excites more carriers when it scatters, *avalance breakdown* occurs. The ensuing cascade makes is possible to detect very small numbers of electrons or photons.

Something similar happens at the interface between a semiconductor and a metal, shown in Figure 11.7. When the materials are brought together, a current must initially flow to equalize the Fermi levels. But because the metal cannot have an internal electric field the band bending occurs entirely on the semiconductor side. This is called a *Schottky barrier*, and it once again rectifies the current, which can be either a bug or a feature. It's a simpler way to fabricate a diode because all that's needed is a metallization, but it also means that any lead attached to a semiconductor becomes a diode. Creating linear *ohmic*

*contact* requires extra effort, such as heavily doping the semiconductor at the interface to keep the transition region thin enough to permit tunnelling.

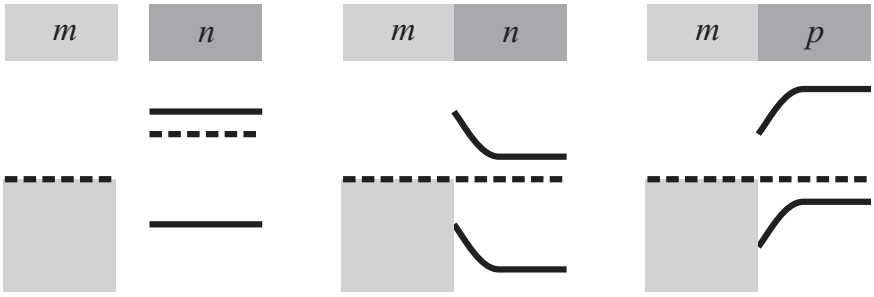


Figure 11.7. A Schottky barrier between a semiconductor and a metal.

If one junction is good, two are even better. Consider the back-to-back diodes shown in Figure 11.8, forming a *bipolar transistor*. The center region is called the *base* and the sides are the *emitter* and *collector*. In the absence of any bias, current cannot flow into either the emitter or collector. But if the emitter–base junction is forward-biased and the collector–base junction is reverse-biased then a current $I_{CE}$ can flow through the collector–emitter circuit. As before, the emitter–base junction will have an $I$–$V$ curve of the form of equation (11.52), but now in addition to determining its own current flow $V_{BE}$ will set that between the emitter and the collector if a voltage source is connected across them,

$$\begin{aligned} I_{CE} &= I_S \left( e^{qV_{BE}/kT} - 1 \right) \\ &= \beta I_{BE} \quad . \end{aligned} \tag{11.53}$$

This is called the *Ebers–Moll model* of a transistor, with a *saturation current* $I_S$. Because a voltage determines a current this is a *transconductance* device, but since $V_{BE}$ also produces a current $I_{BE}$ it's simpler to understand the transistor as a current amplifier. What makes this device so useful is that a small current between the emitter and base can control a much larger current between the emitter and collector; the proportionality factor $\beta$ is typically on the order of 100.

Bipolar transistors do have a significant liability: keeping them turned on draws a steady current through the base. This limits their use in applications for which power consumption or heat dissipation need to be minimized (i.e., almost all of them). Figure 11.9 show how a *Field-Effect Transistor* (*FET*) cures this by using an electric field rather than a current as the control input. A carrier source and drain are separated by a semiconducting channel, covered by a thin insulating layer and a metallic gate. One of silicon's great virtues is that it readily grows a tough $SiO_2$ oxide that is a good insulator, making this a Metal-Oxide-Semiconductor FET or *MOSFET*.

If the channel is $p$-type, and the source and drain are $n$-type, this is an *NMOS* transistor because electrons are the current carriers. Figure 11.9 plots the band structure along a vertical slice through the gate, oxide, and substrate. The overlap across the thin oxide aligns the Fermi levels of the gate and the substrate. As the gate is biased relative to the substrate the Fermi levels split, but the position of the bands at the interface is

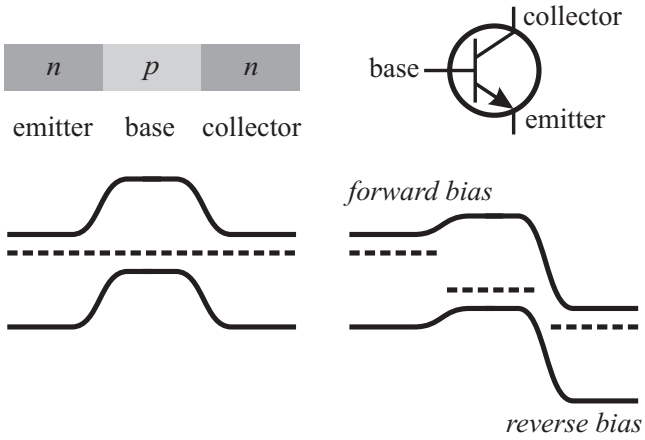


Figure 11.8. An $n$–$p$–$n$ transistor.

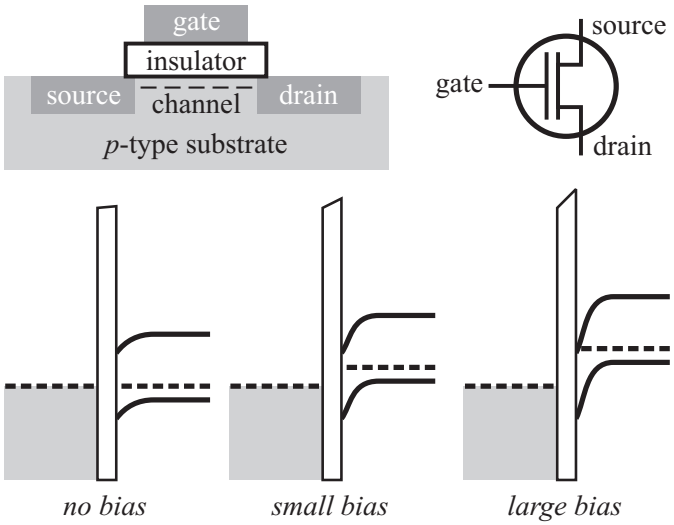


Figure 11.9. An NMOS FET.

fixed by the material properties. This is accomplished once again by the formation of a field gradient in a transition region. As the substrate Fermi level gets closer to the conduction band at the surface, excess electrons begin to appear in the channel. The gate–oxide–substrate combination can be thought of as a capacitor with a semiconductor for one plate. Charge on the gate must be matched by image charge in the substrate, but because it is semiconducting this image charge also changes the conductivity. Unlike the continuous base–emitter current drawn by a bipolar transistor, a MOSFET is a voltage-controlled device that dissipates control current only when the gate voltage is changing and hence charging or discharging this capacitance.

Figure 11.10 plots the current $I_{DS}$ between the drain and source as a function of the voltage $V_{GS}$ between the gate and source, for a fixed voltage $V_{DS}$ between the drain and source. An *enhancement mode* NMOS MOSFET is doped so that no current will flow

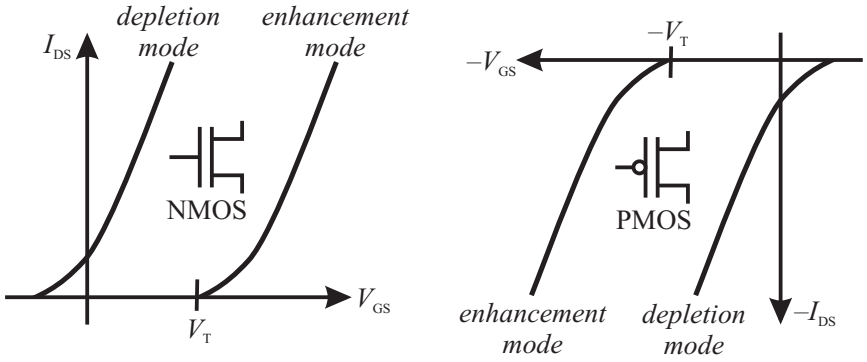


Figure 11.10. Threshold currents in MOSFETs, shown for a fixed $V_{DS}$.

for $V_{GS} = 0$. As $V_{GS}$ is increased it reaches the threshold voltage $V_T$ that brings the Fermi level between the valence and conduction bands so that electrons start to appear in the channel. Further increasing $V_{GS}$ increases the number of carriers, reducing the channel resistance. A *depletion* mode device is doped so that the Fermi level starts out high enough for there to be some conduction carriers; a negative threshold voltage is needed to turn this kind of transistor off. In a *PMOS* MOSFET the channel is $n$-type and the source and drain are $p$-type, the current is carried by holes, and decreasing the gate voltage increases their concentration.

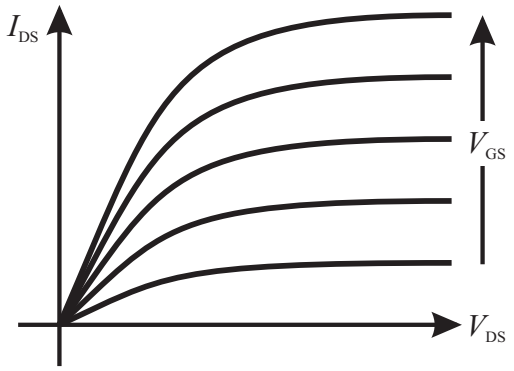


Figure 11.11. $I$–$V$ curves for an NMOS FET as a function of the gate voltage.

For small values of $V_{DS}$ the current $I_{DS}$ will be linear (ohmic), as shown in Figure 11.11. Increasing $V_{GS}$ decreases the resistance, thereby increasing the slope. But as $V_{DS}$ is increased the electrons in the channel are also pulled towards the source, eventually pinching off the channel and saturating the current.

## 11.4 LOGIC

*TTL* (*Transistor–Transistor Logic*) integrates bipolar transistors on a semiconducting

substrate to implement logical functions. While historically significant, its use is limited by the static current drawn by a gate when it is turned on. MOSFETs are an attractive alternative, but the asymmetry shown in Figure 11.10 presents a serious obstacle to their use.

Consider the cases shown in Figure 11.12. An NMOS and a PMOS enhancement-mode FET are being used to drive another FET, represented by its gate capacitance. In case (a), an NMOS FET is turned on by applying the supply voltage to the gate. For MOSFETs this voltage is called $V_{DD}$ (for historical reasons, the D is for drain; for bipolar transistors the supply is usually labeled $V_{CC}$, C for collector). Assume that the capacitor is initially charged to $V_{DD}$ and that the input to the FET is grounded. Because electrons are the charge carriers in an NMOS FET, they must flow from the source at ground to the positive capacitor at the drain to discharge it. Since $V_{GS}$ remains at $V_{DD} > V_T$ the FET stays on and the capacitor is fully discharged. Compare this to case (b), with the capacitor starting out grounded and $V_{DD}$ applied to the input. This is a problem. For the capacitor to charge up, electrons must flow from it, so it is the source. But as its voltage rises, $V_{GS}$ will eventually drop below $V_T$, shutting of the FET with $V_{DD} - V_T$ left on the capacitor instead of the desired $V_{DD}$. Because an NMOS FET can discharge a capacitor to ground, it can output a logical 0, but it can't output a logical 1 because it can't charge a capacitor up to the positive supply. Likewise, a PMOS FET can output a 1 but not a 0.

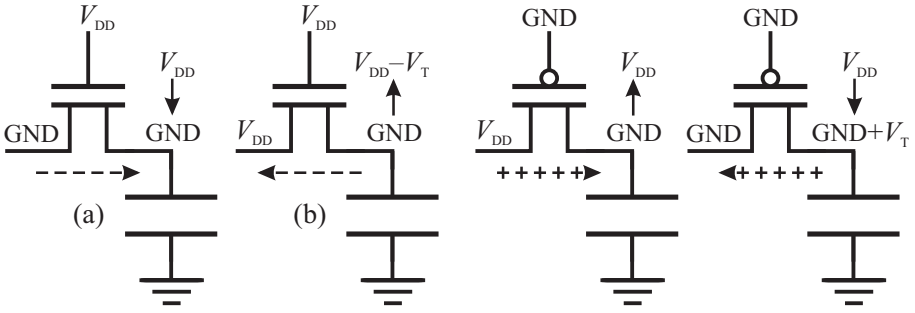


Figure 11.12. Charging and discharging capacitors through MOSFETs.

In integrated electronics, as in life, the solution to shared imperfections is a relationship based on complementary strengths. *CMOS* (*Complementary Metal Oxide Semiconductor*) logic uses pairs of MOSFETs, as shown in Figure 11.13 for the simplest circuit of all, an inverter. If the input is grounded, the PMOS transistor is on and the NMOS transistor is off, therefore the output is pulled up to $V_{DD}$, which the PMOS transistor can do well. If $V_{DD}$ is input, the PMOS transistor turns off and the NMOS transistor turns on, bringing the output to ground which it can do well. We have inverted the input. For either state only one transistor is turned on, it is used in the mode in which it works best, and current is drawn only during state changes. In practice, it is important that the PMOS and NMOS threshold voltages be well matched, otherwise during transitions there may be a period when they are both turned on and a *crowbar current* will flow from $V_{DD}$ to ground.

A gate with two inputs is shown in Figure 11.14. Now two NMOS transistors are

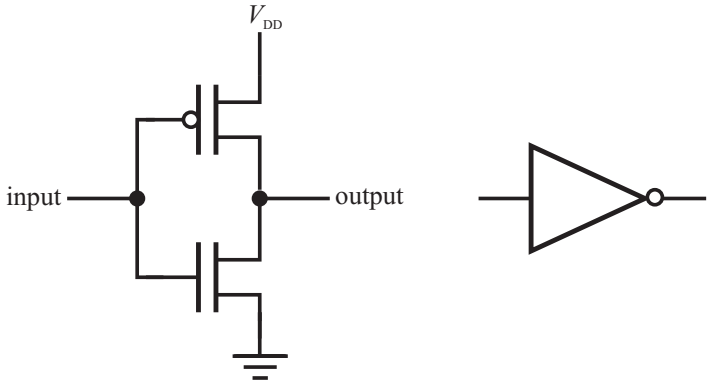


Figure 11.13. A CMOS NOT gate and its circuit symbol.

connected in parallel to ground, and two PMOS transistors are connected in series to $V_{DD}$. If $A = B$ = GND then the output will be pulled up to $V_{DD}$, and in all other cases it will be pulled down to ground. This is the NOR (not-or) function. Similarly simple circuits can implement the other basic logical functions. From these two examples the essential principle of CMOS design should be clear: NMOS FETs are used only to pull outputs to ground, and PMOS FETs are used only to pull them to $V_{DD}$.

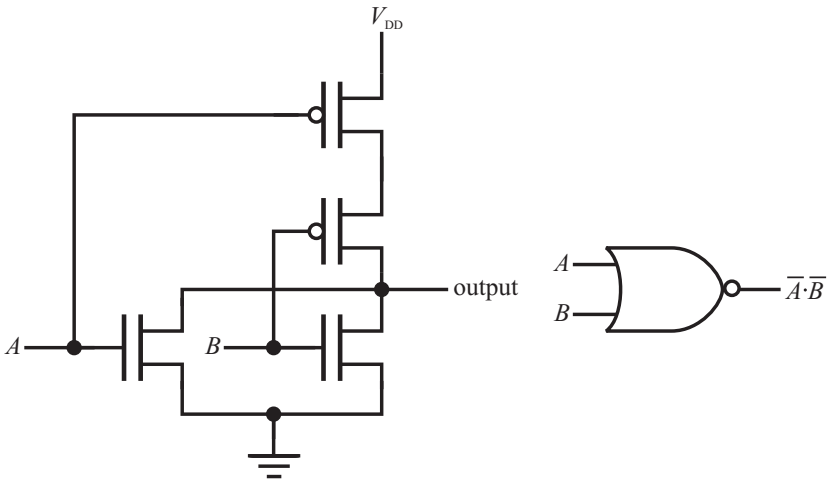


Figure 11.14. A CMOS NOR gate.

Table 11.1. *Linear* (XOR) *and nonlinear* (NOR) *logical functions.*

| $x$ | $1+x$ | $y$ | XOR($x$,$y$) | XOR($1+x$,$y$) | NOR($x$,$y$) | NOR($1+x$,$y$) |
|---|---|---|---|---|---|---|
| 0 | 1 | 0 | 0 | 1 | 1 | 0 |
| 0 | 1 | 1 | 1 | 0 | 0 | 0 |
| 1 | 0 | 0 | 1 | 0 | 0 | 1 |
| 1 | 0 | 1 | 0 | 1 | 0 | 0 |

Because the NOR gate is a nonlinear function of its arguments (Table 11.1), it is possible to obtain any logical function by combining it with NOT gates [Hill & Peterson, 1993]. A different nonlinear gate such as AND could be used as a primitive instead, but it is not possible with a linear gate such as XOR (exclusive-or) . An arbitrary logical function can in fact be realized in a *two-level* implementation using just a layer of NOT gates connected to a layer of NOR gates, so that the propagation delay of a signal through the circuit is fixed. This configuration is available packaged in a *Programmable Logic Array* (*PLA*) that can be used as a universal logical element. The technique most commonly used to reduce an arbitrary logical function to the smallest two-level implementation is the *Quine–McCluskey algorithm*, first developed by the philosopher W.V. Quine long before integrated circuits existed, in order to solve a puzzle in mathematical logic [Quine, 1952; McCluskey, 1956].

So far we've been discussing *combinatorial logic*, in which the output is determined by the instantaneous inputs. *Sequential logic* adds memory and a clock signal to drive transitions, so that the output can depend on the past as well as present values of the input. Clocks will be covered in Chapter 14; the last circuits to be considered here are *Random Access Memories* (*RAM*), starting with the *Static RAM* (*SRAM*) cell shown in Figure 11.15. The bit there is stored in a bistable configuration of two coupled inverters. If the input to one of the inverters is a logical 1 its output will be a 0, and this input to the other inverter will produce an output of 1 from it, agreeing with the input to the first inverter. The two inverters will also be in a stable configuration if the output of the first one is 1 and that of the second one is 0. To make this into a memory, transistors are connected between the outputs of the inverters and bit read/write lines. These pass transistors are turned on by row enable lines, letting a particular combination of row and bit lines address a unique bit. If sense amplifiers are connected to the bit lines, they can measure the state of the inverters and read out the bit, and if drive amplifiers are connected to the bit lines they can write a bit by forcing the inverters into a desired state. Two bit lines (the bit and its complement) are needed to make sure that both inverters end up in the desired state.

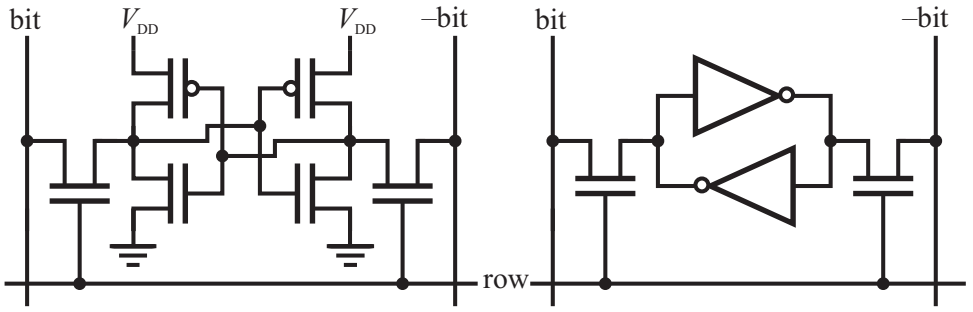


Figure 11.15. A CMOS SRAM cell.

The basic SRAM cell requires six transistors. In 1966 Bob Dennard at IBM realized that it is possible to make a memory with just one transistor and one capacitor per bit, significantly increasing the bit density [Dennard, 1968]. In such a *Dynamic RAM* (*DRAM*) cell the bit is stored as charge on a capacitor, as shown in Figure 11.16. When

the row and bit enable lines are turned on, charge can be written into the capacitor to store a bit, or an amplifier can detect the charge to read the bit. Unlike an SRAM cell, this read operation is destructive because the capacitor is used to charge up the bit line while it is being read, and even if a bit isn't read the charge will eventually leak away from the capacitor, therefore DRAM cells require complex refresh circuits. However, the space saving from having 1 transistor per bit much more than makes up for this extra complexity at the periphery of the memory. DRAM is also slower than SRAM because the bit line is passively driven by a capacitor rather than actively driven by an inverter. Because of this, less dense SRAM is used for fast cache memory, and denser DRAM is used for larger slower primary memories.

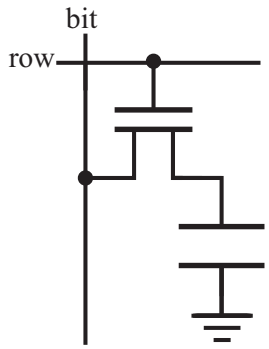


Figure 11.16. A CMOS DRAM cell.

Both SRAM and DRAM are *volatile* memories that must be powered to maintain their data. Since power can run out long before the value of data does, particularly in mobile or embedded applications, *non-volatile* memories are needed. The most common solution is to add a *floating gate* to the MOSFET structure shown in Figure 11.9. This is an electrode between the gate and the channel, completely surrounded by the insulating oxide. Charge stored on the floating gate can be read through the image charge it induces in the substrate changing the conductivity of the channel, and because it is completely isolated the charge retention times can be very long (many years). In *EPROM* (*Erasable Programmable Read-Only Memory*), charge is deposited on the floating gate by using a write voltage large enough to excite high-energy "hot" electrons over the gate's barrier, and the entire memory is erased by exposing the die to ultraviolet light with enough energy to knock the electrons back out. Because of the tens of volts and ultraviolet light needed for writing and reading, dedicated programmers are used for changing EPROM. In *EEPROM* (*Electrically Erasable Programmable Read-Only Memory*), the oxide thickness is reduced from $\sim$100 nm to $\sim$10 nm, making it possible for electrons to tunnel onto and off of the floating gate [Fowler & Nordheim, 1928]. This requires an extra transistor per cell to control the charging, and can reduce the charge storage time and reliability of the device, but it permits in-circuit access to read and write arbitrary bits. *Flash memory* is a compromise that writes with hot electrons and erases with tunneling, permitting in-circuit programming, using just one transistor per cell at the expense of restricting erasure to memory sectors rather than individual bits.

Because of its relative ease of fabrication, low power consumption, and high pack-

ing density, CMOS dominates integrated circuit production. This has historically been accompanied by slower switching speeds because of the $RC$ charging time associated with gate transitions. This is why higher-frequency applications have used higher-power bipolar TTL or *ECL* (*Emitter-Coupled Logic*) families, or higher-mobility materials such as GaAs. But the speed of CMOS has increased beyond 1 GHz because of beneficial features of device scaling. Once the materials become pure enough, and the channel drops well below 1 μm, an electron can travel through a transistor without scattering. Such *ballistic* or *hot electron* devices can operate at speeds far beyond what the bulk mobility would suggest. CMOS is still limited in its ability to source or sink current; for this reason *BiCMOS* processes marry the best of both worlds by using MOSFETs for on-chip logic and bipolar transistors for driving external signals.

## 11.5 BEYOND CMOS

### 11.5.1 printed organic

Brütting, Wolfgang. "Introduction to the physics of organic semiconductors." Physics of organic semiconductors (2005): 1-14.

Berggren, Magnus, David Nilsson, and Nathaniel D. Robinson. "Organic materials for printed electronics." Nature materials 6, no. 1 (2007): 3-5.

### 11.5.2 printed inorganic

Ridley, Brent A., Babak Nivi, and Joseph M. Jacobson. "All-inorganic field effect transistors fabricated by printing." Science 286, no. 5440 (1999): 746-749.

Sun, Yugang, and John A. Rogers. "Inorganic semiconductors for flexible electronics." Advanced materials 19, no. 15 (2007): 1897-1916.

### 11.5.3 graphene

Novoselov, Kostya S., Andre K. Geim, Sergei V. Morozov, D. Jiang, Y_ Zhang, Sergey V. Dubonos, Irina V. Grigorieva, and Alexandr A. Firsov. "Electric field effect in atomically thin carbon films." science 306, no. 5696 (2004): 666-669.

Schwierz, Frank. "Graphene transistors." Nature nanotechnology 5, no. 7 (2010): 487.

Mogera, Umesha, and Giridhar U. Kulkarni. "A new twist in graphene research: Twisted graphene." Carbon 156 (2020): 470-487.

Cao, Yuan, Valla Fatemi, Shiang Fang, Kenji Watanabe, Takashi Taniguchi, Efthimios Kaxiras, and Pablo Jarillo-Herrero. "Unconventional superconductivity in magic-angle graphene superlattices." Nature 556, no. 7699 (2018): 43-50.

Moysidis, Savvas, Ioannis G. Karafyllidis, and Panagiotis Dimitrakis. "Graphene logic gates." IEEE Transactions on Nanotechnology 17, no. 4 (2018): 852-859.

### 11.5.4 CNT

Bachtold, Adrian, Peter Hadley, Takeshi Nakanishi, and Cees Dekker. "Logic circuits with carbon nanotube transistors." Science 294, no. 5545 (2001): 1317-1320.

McEuen, Paul L., Michael S. Fuhrer, and Hongkun Park. "Single-walled carbon nanotube electronics." IEEE transactions on nanotechnology 99, no. 1 (2002): 78-85.

Chen, Zhihong, Joerg Appenzeller, Yu-Ming Lin, Jennifer Sippel-Oakley, Andrew G. Rinzler, Jinyao Tang, Shalom J. Wind, Paul M. Solomon, and Phaedon Avouris. "An integrated logic circuit assembled on a single carbon nanotube." Science 311, no. 5768 (2006): 1735-1735.

### 11.5.5 SET

Chen, R. H., A. N. Korotkov, and K. K. Likharev. "Single-electron transistor logic." Applied Physics Letters 68, no. 14 (1996): 1954-1956.

Kastner, Marc A. "The single-electron transistor." Reviews of modern physics 64, no. 3 (1992): 849.

Guo, Lingjie, Effendi Leobandung, and Stephen Y. Chou. "A silicon single-electron transistor memory operating at room temperature." Science 275, no. 5300 (1997): 649-651.

### 11.5.6 mechanical

Bromley, Allan G. "Charles babbage's analytical engine, 1838." Annals of the History of Computing 4, no. 3 (1982): 196-217.

Menabrea, Luigi. "Sketch of the Analytical Engine. Invented by Charles Babbage. By LF Menabrea of Turin, Officer of the Military Engineers. With notes upon the Memoir by the Translator Ada Augusta, Countess of Lovelace." Bibliothèque Universelle de Genève (1942).

Dewdney, A. "A tinkertoy computer that plays tic-tac-toe." Scientific American 261, no. 4 (1989): 120-123.

Blick, Robert H., Hua Qin, Hyun-Seok Kim, and Robert Marsland. "A nanomechanical computer—exploring new avenues of computing." New Journal of Physics 9, no. 7 (2007): 241.

Wenzler, Josef-Stefan, Tyler Dunn, Tommaso Toffoli, and Pritiraj Mohanty. "A nanomechanical Fredkin gate." Nano letters 14, no. 1 (2013): 89-93.

Hafiz, Md Abdullah Al, Lakshmoji Kosuru, and Mohammad I. Younis. "Microelectromechanical reprogrammable logic device." Nature communications 7 (2016): 11137.

Masmanidis, Sotiris C., and Rassul B. Karabalin. "Iwijn De Vlaminck, Gustaaf Borghs, Mark R." Freeman, and Michael L. Roukes,"Multifunctional Nanomechanical Systems via Tunably Coupled Piezoelectric Actuation", Science 317 (2007): 780.

### 11.5.7 memristors, neural networks

Strukov, Dmitri B., Gregory S. Snider, Duncan R. Stewart, and R. Stanley Williams. "The missing memristor found." nature 453, no. 7191 (2008): 80-83.

Vongehr, Sascha, and Xiangkang Meng. "The missing memristor has not been found." Scientific reports 5, no. 1 (2015): 1-7.

Zidan, Mohammed Affan, Hossam Aly Hassan Fahmy, Muhammad Mustafa Hussain,

and Khaled Nabil Salama. "Memristor-based memory: The sneak paths problem and solutions." Microelectronics journal 44, no. 2 (2013): 176-183.

Merced-Grafals, Emmanuelle J., Noraica Dávila, Ning Ge, R. Stanley Williams, and John Paul Strachan. "Repeatable, accurate, and high speed multi-level programming of memristor 1T1R arrays for power efficient analog computing applications." Nanotechnology 27, no. 36 (2016): 365202.

Thomas, Andy. "Memristor-based neural networks." Journal of Physics D: Applied Physics 46, no. 9 (2013): 093001.

Yao, Peng, Huaqiang Wu, Bin Gao, Jianshi Tang, Qingtian Zhang, Wenqiang Zhang, J. Joshua Yang, and He Qian. "Fully hardware-implemented memristor convolutional neural network." Nature 577, no. 7792 (2020): 641-646.

### 11.5.8 defect tolerance

Heath, James R., Philip J. Kuekes, Gregory S. Snider, and R. Stanley Williams. "A defect-tolerant computer architecture: Opportunities for nanotechnology." Science 280, no. 5370 (1998): 1716-1721.

### 11.5.9 spatial

Gershenfeld, Neil. "Aligning the representation and reality of computation with asynchronous logic automata." Computing 93, no. 2-4 (2011): 91-102.

## 11.6 Limits

In the 1960s, Gordon Moore of Intel noticed that the number of transistors on a chip, along with almost every other specification, was doubling every one and a half years. This exponential growth has come be known as *Moore's Law* [Moore, 1979]. It is not a law of nature; it is an observation about exceptional engineering efforts. The many decades over which it has applied have made it possible to foretell the future of *Very-Large-Scale Integrated circuits* (*VLSI*) with surprising prescience. On the one hand, seemingly insurmountable obstacles have been overcome each year to continue the scaling; on the other hand, sometime around 2020–2040 most devices parameters will simultaneously reach fundamental physical limits [Keyes, 1987]. Wires as we know them cannot be thinner than one atom, memories cannot have fewer than one electron, and to be viable the chip fab plants must cost something less than the GDP of the planet.

While these impending limits are a cause for concern about continued progress in improving the performance of electronics, the reality is that gate speeds or bit densities are already no longer the serious constraints they once were for many applications. Some of the most interesting challenges now lie beyond processing with detecting, communicating, and presenting electronic information. Nevertheless, because each decade of device improvement has led to corresponding new and generally unforeseen opportunities, it's still worth asking how to maintain and extend this pace.

The present battleground is the physics of *microfabrication* [Brodie & Muray, 1982]. Chips are currently produced using *lithographic* techniques to optically define device

features. The diffraction limit has pushed the optical systems to shorter and shorter wavelengths, but this is still well above atomic sizes. True atomic-scale patterning has been accomplished using *Atomic Force Microscopes* (*AFMs*) [Cooper *et al*., 1999] and *Scanning Tunneling Miscroscopes* (*STMs*) [Stroscio & Eigler, 1991] , which piezoelectrically scan a sharp tip above a sample and follow the atomic topography by measuring either the cantilever deflection or the electron tunneling current. While these *scanning probe* systems are slow, lithographically-produced parallel arrays of tips promise to yield commercially useful writing speeds.

The billions of dollars that must be invested in the machinery to deposit, expose, etch, implant, dope, diffuse, dice, and test chips in a single fab line are quickly becoming the more serious scaling constraint. An alternative is to eliminate the fab line entirely and use table-top printing processes, which can attain nanometer features [Jackman *et al*., 1998; Ridley *et al*., 1999]. A related limit that receives less attention but may become even more economically significant is the lowest cost per packaged part, which unlike most other specifications has remained relatively constant over the VLSI scaling era at around 10 cents. An alternative approach to bring this down below a cent for applications such as electronic tagging of commodity objects is to remotely interrogate natural materials [Fletcher *et al*., 1997].

One of the reasons chip fabrication is so expensive is that as the size of chips grows while their minimum feature size shrinks, the impact of a single speck of dust is magnified. A very small defect can doom an entire part. The conventional response has been to use ultrapure materials in ultraclean rooms, but even so the yields of state-of-the-art chips are usually so bad that they are considered a sensitive trade secret (on the order of a few percent). A radically different approach is to design machines with the expectation that most of their parts will be faulty [Heath *et al*., 1998]. Components can be hierarchically packaged in modules of greater and greater complexity, which can then be adaptively rewired based on self-testing. There is some empirical basis for this kind of partitioning, through *Rent's rule*, the observation that many engineered (and biological) systems have a power-law scaling relationship between the number of connections to a subsystem and the number of functional units in that subsystem, with an exponent typically between 1/2 and 1 [Landman & Russo, 1971; Vilkelis, 1982].

Thermodynamics presents profound limitations that are also of great short-term significance [Gershenfeld, 1996]. 10 W laptops run out of power before airplane trips end, the 100 W desktop computers in a building taken together can consume more power than air conditioning systems use or can handle, and it's a challenge to keep a 100 kW supercomputer from melting through the floor. As we saw in Section 4.5, the roots of information theory grew out of the study of the efficiency of steam engines, and are now returning to help optimize the thermodynamic performance of computing machines [Leff & Rex, 1990]. Rolf Landauer resolved a long-standing puzzle by showing that erasure is where computation necessarily incurs a thermodynamic cost, because the heat associated with the change in entropy that follows from resetting an unknown bit to a known state is $Q = TdS = kT \log 2$ [Landauer, 1961]. Charles Bennett went further to unexpectedly demonstrate that universal computation is possible without any erasure by reversibly rearranging inputs and outputs [Bennett, 1973]. Since $kT \log 2 \approx 10^{-21}$ J, it was originally thought that these limits were remote. More recently, it's been appreciated that the design guidance they provide is applicable at much higher energy scales. *Reversible*

*logic* seeks to recover rather than dissipate the energy associated with bits being erased [Merkle, 1993; Younis & Knight, 1993], and *adiabatic logic* makes changes no faster than they are needed (Problem 11.5) [Athas *et al*., 1994; Dickinson & Denker, 1995]. Both principles have been used in fabricating circuits that show promising reductions in power consumption.

Even more fundamental are limits associated with physical constants. One is the speed of light. Synchronous logic requires distributing the clock over an entire chip each cycle; aside from the charging energy this entails, it also limits the cycle time to the chip size divided by the speed of light. One response is to eliminate clock delivery by using *asynchronous logic*, in which gates assert their output when they receive valid inputs rather than a global clock signal [Birtwistle & Davis, 1995]. This can also be beneficial for reducing dissipation and wiring complexity, although the ultimate limit on clock speed comes from the quantum energy–time uncertainty relationship, which argues for using the maximum available energy in the minimum possible number of gates in order to minimize the communication time [Lloyd, 2000].

Another is the size of atoms. As feature sizes drop below 0.1 μm, continuum approximations can no longer be made. This shows up in *electromigration*, the motion of individual atoms due to the momentum transported by the electronic current, which leads to wiring failures that must be prevented through careful attention to the metallurgy and current density. The discreteness of current is turned from a bug into a feature in a *Single-Electron Transistor* (*SET*) [Likharev & Claeson, 1992; Grabert & Devoret, 1992]. An electron can tunnel across an insulator onto a conducting island only if states are available to it on both sides, creating a periodic modulation called the *Coulomb blockade* in the charging current due to the integer number of electrons allowed on the island. Among other applications, this can be used to create a memory cell that stores a single electron [Durrani *et al*., 1999].

Once devices reach these limits, further advances are possible only by finding new degrees of freedom to represent and manipulate information. One option is to recognize that analog nonlinear systems can be used for far more than binary logic; some examples will be seen in Chapter 15. Another is to retain the notion of bits, but use quantum mechanics to describe their logical as well as physical states. The remarkable implications of this will be explored in Chapter 16.

## 11.7 SELECTED REFERENCES

[Ashcroft & Mermin, 1976] Ashcroft, N., & Mermin, N.D. (1976). *Solid State Physics*. New York: Holt, Rinehart and Winston.

A very readable introduction to solid state physics. The index deserves special attention.

[Sze, 2006] Sze, S.M. (ed). (2006). *Modern Semiconductor Device Physics*. 3rd edn. New York: Wiley-Interscience.

Definitive device physics text.

[Streetman & Banerjee, 2014] Streetman, Ben, & Banerjee, Sanjay. (2014). *Solid State Electronic Devices*. 7th edn. Pearson.

This is a more accessible introduction to device physics.

## 11.8 Problems

(11.1) (*a*) Derive equation (11.28) by taking the integral and limit of equation (11.27).
(*b*) Show that equation (11.29) follows.

(11.2) What is the expected occupancy of a state at the conduction band edge for Ge, Si, and diamond at room temperature (300 K)?

(11.3) Consider Si doped with $10^{17}$ As atoms/cm$^3$.
(*a*) What is the equilibrium hole concentration at 300 K?
(*b*) How much does this move $E_F$ relative to its intrinsic value?

(11.4) Design a *tristate* CMOS inverter by adding a control input to a conventional inverter that can force the output to a high impedance (disconnected) state. These are useful for allowing multiple gates to share a single wire.

(11.5) Let the output of a logic circuit be connected by a wire of resistance $R$ to a load of capacitance $C$ (i.e., the gate of the next FET). The load capacitor is initially discharged, then when the gate is turned on it is charged up to the supply voltage $V$. Assume that the output is turned on instantly, and take the supply voltage to be 1.8 V and the gate capacitance to be 1 fF.
(*a*) How much energy is stored in the capacitor?
(*b*) How much energy was dissipated in the wire?
(*c*) Approximately how much energy is dissipated in the wire if the supply voltage is linearly ramped from 0 to 1.8 V during a long time $\tau$?
(*d*) How often must the capacitor be charged and discharged for it to draw 1 W from the power supply?
(*e*) If an IC has $10^9$ transistors, each charging and discharging this gate capacitance once every cycle of a 1 GHz clock, how much power would be consumed in this worst-case estimate?
(*f*) How many electrons are stored in the capacitor?