1
Introduction and Review of Electronic Technology

Electronic devices are central to modern technology. Silicon chips are everywhere, including cars and appliances, and have transformed computation, information processing, and communications, culminating in the modern Internet. The silicon revolution started with the transistor, leading to the integrated circuit, and to the Pentium chip. A related semiconductor device, the solid-state junction laser, in conjunction with the optical fiber, has led to cheap, reliable virtually instantaneous worldwide communication. Outsourcing, globalization, and the "flat world" have been enabled by these technical advances.

The assumption of this book is that this revolution is not over, but rather is entering a new phase. The era of microelectronics is opening to a future of nanoelectronic technology.

The central feature of the silicon revolution has been the miniaturization of transistors and their grouping into integrated circuits, which now contain billions of identical transistor elements. In very large scale integration (VLSI), a whole computer can exist on a square centimeter of silicon. Smaller transistors are cheaper, more are available on a single chip, and operate more quickly, now as fast as 3 billion steps per second. Gordon Moore, a founder of the Intel Corporation, noted long ago that the number of transistors per chip, roughly one square centimeter, tended to double every 18 months or so as the technology improved and larger markets appeared. "Moore's law" has seen the transistor count increase from hundreds to hundreds of millions! Chips containing 0.8 billion transistors on roughly one square centimeter are being produced as described in Chapter 7 [1].

The key to this advance has been the "scaling" to smaller size of the active cells, containing field effect transistors (FETs) and other devices. Scaling has taken silicon electronics into the nanometer domain, where it now is approaching its limit, set by the size of atoms. The smallest dimension in the FET has been the thickness of the thermally grown silicon dioxide insulator for the gate electrode. It has long been recognized that scaling will work only down to thicknesses large compared to the silicon and oxygen atomic radii in the SiO_2, needed to preserve the desired insulating property. Intel Corporation [1] has abandoned the scaled silicon dioxide to insulate the gate electrode, in favor of deposited "high dielectric constant" oxides based on the heavy metals hafnium and zirconium. Literally, the thermally grown silica, forced

Quantum Nanoelectronics. Second Edition. Edward L. Wolf

ISBN: 978-3-527-40749-1

thinner and thinner by the scaling formula, contained only a handful of silicon atoms across its thickness, allowing electrons to “leak” through by the quantum mechanical tunneling effect.

A second imperative for a new era in nanoelectronics comes from the limit of patterning resolution, limited by wavelength of the light used to imprint patterned features onto the chip. The smallest feature size in the newest generation of silicon devices is 45 nm, achieved by artful use of light of 193 nm wavelength.

The equipment for producing and applying the patterning light is a leading cost in a fabrication facility, and reducing the wavelength of patterning light has become increasingly difficult and costly. Energy conservation is also a driving force for technology change. Large computing installations consume megawatts of power. Laptop computers run hot and their batteries frequently need recharging.

A third indicator, an opportunity for change, comes from new computing technologies. One of these, the Josephson junction-based “rapid single flux quantum” technology, has logic circuits that are wholly superconducting and thus use less energy. An entirely new concept is “quantum computing,” in which the binary bit is replaced by a “qubit” that can take on more than two values, which is inherently faster in solving certain types of computations, including factoring of large numbers, of great interest in cryptography. The first realization of this new class of “quantum computers” has also been based on Josephson junctions, using superconducting niobium at very low temperature. A large installation of this type would use much less power than today’s supercomputer.

The traditional scaling to smaller sizes in the silicon technology has come to an end. As noted, the Intel Corporation has abandoned the thermal silicon gate oxide, and the photolithography cannot proceed further without abandoning lenses of any sort. Size reduction will continue for several more generations, but increasingly requires further essential innovation. Nonetheless, it is entirely clear that the present silicon computer technology has essentially reached its final form.

The several aspects of this new situation are the topics of this book. A new era in electronics is opening, with opportunities for new ideas, new companies, and new jobs for technologists.

Of course, the present state of computing technology is highly advanced, delivering huge performance at still decreasing cost. This technology will not disappear. The new evolution may resemble that of automobiles. One can argue that automobiles have only incrementally changed since the introduction of the automatic transmission and air conditioning, and a similar outcome is not unreasonable for electronics. It is unlikely that there will be a change away from silicon electronics as dramatic and complete as the recent tipping point in which digital photography replaced silver halide photography. It does seem likely, though, that in some application areas there will be reason enough to have major innovation.

The important point is that still there is room for much improvement. As we will see, the laws of physics still allow huge advances, perhaps a factor of 10 000, in device density and computer performance beyond the present levels in the semiconductor technology. For example, the area of the repeating unit cell in the most recent [1] 45 nm Intel CMOS (complementary metal oxide silicon) technology is $0.346\,\mu\text{m}^2$, corresponding to device density of $2.89 \times 10^{12}\,\mathrm{m}^{-2}$ or $2.89 \times 10^{8}\,\mathrm{cm}^{-2}$. ("45 nm" refers to the smallest feature size *F*, but the size of the repeating cell of the 45 nm technology is much larger, about 590 nm. For this chip, if we express the area as a multiple of the square of the feature size $F = 45\,\mathrm{nm}$, the multiple is 170.9.) But analysis of a proposed hybrid molecular/silicon technology, "CMOL," predicts a device density of $3 \times 10^{12}\,\mathrm{cm}^{-2}$, an increase by a factor (10 380.6) exceeding 10^4! (The corresponding equivalent linear sizes of the repeated cell are 588 and 5.77 nm.)

There remains a huge opportunity for improvement, quite allowed by laws of physics. But most of the physically allowed improvement is unlikely to be achieved by continued scaling the present silicon technology.

How will breakout into a new technology occur? Applications able to pay a premium for improved performance may justify new research, new approaches, and new fabrication facilities. The traditional opportunities lie in military applications and supercomputing, but the rapidly rising need for fast computation in "cloud computing" may stimulate an earlier response from the private sector.

Military applications favor nonvolatile memory and radiation hardness, for example, neither a strong point of the silicon technology. Present large-scale supercomputing installations waste energy, expending megawatts for running and cooling the computers and also for cooling the rooms that contain them. For example, the *Wall Street Journal* on October 30, 1997 described the computer system that had handled 1.2 billion shares on a previous day on the New York Stock Exchange. The Exchange was proud to have spent $2 billion on the new system. The system was described as having 450 Tandem and Hewlett-Packard computers, "refrigerator-sized boxes," operating on a parallel network synchronized by two atomic clocks and using 200 miles of fiber optic cables (the computers are partly in Manhattan and partly in Brooklyn). According to the *Wall Street Journal*, this New York Stock Exchange computer system operates at "3500 kilowatts of power, 8000 tons of air conditioning, 8000 phone circuits, 5000 electronic devices, and 300 data-traffic routers." A typical chip in a laptop computer dissipates 100 W, and high-performance chips use more power and are often water cooled. Energy costs are widely expected to rise. Computers, "server farms," and other telecommunications installations are noticeable contributors to the total electric power usage and the "heat island" effect in large cities.

There are several possibilities for postsilicon and hybrid silicon technologies that may allow the spirit of Moore's law to continue. At least two of the new technologies, involving superconductors, potentially have markedly smaller energy costs.

The start-up company, D-Wave Systems, Inc., has demonstrated [2] its Orion quantum computer that substitutes the cost of cryogenic cooling in place of Pentium chip cooling and air conditioning. (This approach is not at all proven, but is an example of potential.) This "quantum computer" [3] also involves a total change in computing method, in which the "on/off" binary bit is replaced by a variable based on quantum mechanics. It is known that this approach is superior in solving certain classes of computational problems, for example, those that are known as NP-complete. This will be investigated in Chapter 11. An alternative low dissipation cryogenic computing technology, called RSFQ (rapid single flux quantum), is well studied [4] but presently applied only to certain specialized applications. Compared to quantum computing, this approach is at an engineering stage.

The material of choice in these cryogenic computers is likely to be superconducting niobium, which creates zero heat as it conducts an electrical current. For a large installation, “quantum computing,” which is the approach of D-Wave Systems, Inc., or the well-known RSFQ superconducting technology could eventually take the lead. For large-scale supercomputers, including those used in stock exchanges, it is conceivable that a tipping point toward cryogenic supercomputing will be reached, driven by superior performance at lower energy costs and lower overall size and cost of the supercomputer installation.

The “45 nm node” is the present stage [1] of conventional semiconductor technology, and further 32 and 22 nm nodes are projected in the “Roadmap” of the semiconductor industry [5]. 22 nm, or 220 Å, is roughly 200 atom diameters, and this dimension might represent the width of a wire interconnect, perhaps of copper or aluminum. (A size to remember is the radius of the electron “orbit” in the simplest atom, hydrogen. The Bohr radius is 0.0529 nm.) At a size less than roughly 100 atoms on a side, one can no longer expect a piece of copper or aluminum to act like a good metal, and the same thing is true of a semiconductor.

As a first effect in making a wire smaller, notice that the typical electrical carrier is closer to a surface, where scattering will occur, leading to electrical resistance and heating. The properties of atoms are of course very well understood, and the properties of “quantum dots” of dimensions 4 or 5 nm (40 or 50 atoms across) can be confidently estimated. But the important thing is that new rules, those of quantum physics, are needed to predict the properties of matter in the atomic range of sizes and in the transition up to “mesoscale” blocks of matter more than 100 atoms on a side. The worker in nanoelectronics will need to know when the traditional classical physics rules of “semiconductor VLSI” technology will work; otherwise, how to replace these rules using quantum physics. (In Chapter 11, we survey the miniaturization of the transistor and compare it finally to a conceptual transistor based on a benzene ring.)

This will be more important of course if one of the inherent quantum technologies takes the lead. Already workers in semiconductor technology are typically members of interdisciplinary teams involving electrical and computer engineers, chemists and chemical engineers, materials scientists, nanotechnologists, and physicists. This will continue and the trend will be toward a larger role for chemists and physicists who can improvise new approaches when the conventional VLSI approaches no longer apply.

It seems reasonable to use the term “nanoelectronics” to include the ongoing effort in the semiconductor industry and not to restrict its meaning to a possible nonsilicon computing technology. The definition of “nanotechnology,” as employed by the National Science Foundation (NSF) in the United States, is a technology involving controlled manufacture of a device having at least one dimension in the range below 100 nm. By this definition, the silicon technology, in which the gate oxide thickness has long been below 100 nm, is an example of nanotechnology. (In the most recent generation of MOSFET transistors, the “equivalent silicon dioxide” thickness is 1 nm.

The gate insulator in the actual device, however, is thicker, to avoid leakage by quantum mechanical tunneling, but still maintaining the scheduled capacitance by use of a much larger dielectric constant, near 22.)

The nanometer thick oxides that must be accurately defined in the Josephson junction RSFQ technology also fall within the NSF definition of nanotechnology. The "45 nm node" in the silicon technology puts two controlled dimensions, namely the insulator thickness and the wiring linewidth, below 100 nm. A second reason for a broader definition of nanoelectronics is that the winning technology may be hybrid, for example, a molecular logic layer built on top of a silicon device.

The goal of this textbook is to prepare the serious reader to participate in the quest for new approaches in nanoelectronics, starting from a bachelor's degree. (Introductory material may be skipped over by the advanced reader, of course.) The end of traditional scaling in the silicon technology in 2008 is associated with size scales approaching atomic sizes. In this limit, the rules of physics change to quantum mechanics, thus a key to the new era. Continuum descriptions of matter and charge have to be replaced by discrete pictures of charge as electrons and matter as atoms held into molecules and solids by covalent bonds. Quantum mechanics is needed to understand the new aspects. The future technologist should also be aware that molecules may in future play the role of individual transistors (see Chapter 7) and chemical covalent bonds of different types may serve either as insulators (tunnel barriers) or as conducting wires (see Chapter 9).

Since new nanoelectronic approaches may be "hybrid," in which a new element is added onto basic units of the silicon technology, the future technologist should be completely conversant in the silicon technology, including its "post-Moore's law" variations. The sophisticated methods of semiconductor technology are applicable to some nonsilicon approaches, for example, to the quantum computing technology of D-Wave Systems' Orion computer. Here, the basic units are nearly dissipation-free Josephson junctions rather than transistors, but these devices are still fabricated using photolithographic methods.

This book is written for a motivated reader who has an undergraduate training in science and mathematics, but not necessarily in electronics. Introductory physics and chemistry are assumed, but quantum physics and electrical engineering are not assumed. To equip a serious student to participate in a new industry is a challenging and perhaps too difficult task to accomplish using a single new text! This remains the goal, and the approach is to emphasize basic concepts, particularly those of quantum nanophysics. Quantum physics is doubly selected by the failure of classical thinking at the nanometer scale and also by the appearance of explicitly quantum technologies for computing, as described in Chapter 11.

To establish the vocabulary and context, a review of electronics is our starting point. Having laid out the outlines of the very broad technology that nanoelectronics may eventually replace, at least in part, we then turn in the bulk of this text to the central and absolutely necessary concepts of quantum mechanics, as applied to electrons, atoms, molecules, metals, semiconductors, and semiconductor junction devices. Finally, the state of the art in several of the most promising areas of nanoelectronic innovation is addressed.

To assist the serious reader to gain a working knowledge in this challenging field, three further and more advanced sources might be useful [6–8]. The reader is expected to have his pencil, calculator, and introductory physics and chemistry books close at hand.

1.1
Introduction: Functions of Electronic Technology

To review electronic technology, a rather broad area, first the discussion is limited to discrete devices. Then, a classification scheme has been chosen according to function. First we survey sources of current and voltage, DC (direct current) and AC (alternating current), followed by detectors. We then arbitrarily adopt a classification of further devices based on the number of terminals.

1.1.1
Review of Electronic Devices

Electronic devices are selected here for modern and future relevance and are described operationally. Voltage is measured in volts, electrical nominal positive current in amperes, and power $P = IV$ is in watts. Emphasis is placed on devices of current interest, including some large-scale but pivotal devices that perhaps can be significantly improved by application of nanotechnology.

1.1.2
Sources of Current and Voltage: DC

In technical practice, a "current source" is a device that will maintain a constant current through an external circuit even as the load resistance varies, and similarly for a voltage source. We are generally concerned with sources of energy that are efficient and long lasting, with interest frequently in energy per unit mass or energy per unit volume.

1.1.2.1 Batteries: Lithium Ion, Ni–Cd, NiMH, and "Supercapacitors"

Rechargeable batteries reversibly interchange chemical and electrical energy. These are categorized by open-circuit voltage (V) and electrical energy (W h kg^{-1} or W h l^{-1}). The equivalent circuit is an electromotive force voltage (EMF) arising in the underlying chemical reaction, with a series resistance.

A battery consists of an array of individual cells that are connected in series to obtain a desired total voltage and in parallel to reach a necessary current capacity. Often, the peak rate of discharge and maximum power output are important.

Energy density is at a premium in modern applications, for example, in cellular telephones, and also in larger scale applications such as hybrid automobiles (Figure 1.1). Li ion batteries are used in consumer electronics, notably cellular telephones and laptop computers, because of their high energy density, in the range

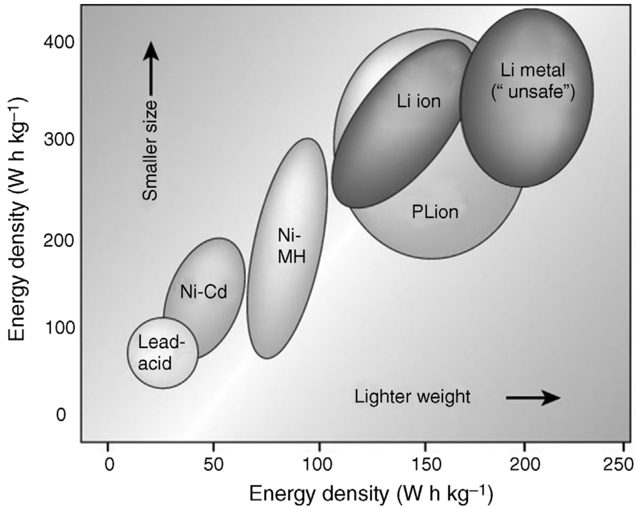

Figure 1.1 Comparison [9] of energy densities for several types of batteries. Units are $\mathrm{W\,h\,l^{-1}}$ (ordinate) and $\mathrm{W\,h\,kg^{-1}}$ (abscissa). Here the label "PLiON" refers to an important class of rechargeable plastic Li ion batteries.

$160\,\mathrm{W\,h\,kg^{-1}}$ and $350\,\mathrm{W\,h\,l^{-1}}$. The Li ion battery has the largest portion of the portable battery market, about 63%, mostly in electronics, but not yet in hybrid cars. Ni–Cd batteries are used in power tools.

The Prius hybrid car uses NiMH (nickel metal hydride) batteries, presently considered a more conservative choice than the Li ion battery, for which, as shown in Figure 1.2, the graphite electrode may present a fire hazard. Nonetheless, an expensive roadster is being sold by Tesla, which is purely electric and powered by 6831 Li ion cells of the type that are used in laptop computers. This vehicle accelerates to 60 mph in less than 4 s and has a range of 210 miles [10].

There is at present an intense competition to finalize the battery design for a mass-production all-electric car, called Volt, planned by General Motors. Innovations being considered for [10] the battery (changes from that shown in Figure 1.2) are a cathode of iron phosphate $LiFePO_4$ (LFP) [11] (manufactured by Lithium Technology Corp. and by A123Sysems, which may involve nanoparticles) and possibly an anode consisting of lithium titanate nanoparticles (Altairnano, Inc.) that will not burn.

It appears [11] that the electrical conductivity of the iron phosphate $LiFePO_4$ cathode, which had been low, was increased by a factor of 10 million by doping with metals such as aluminum, niobium, and zirconium, but also probably involving nanoscopic carbon particles. These advances now make the iron phosphate cathode workable and allow a higher discharge current, fast charging time, and stability under extreme conditions. The basic advantages of LFP cathodes over the cobalt cathodes are low cost, high abundance of iron, and freedom from overheating. Another advantage is that in the chemical reaction in the LFP case, relative to that shown in

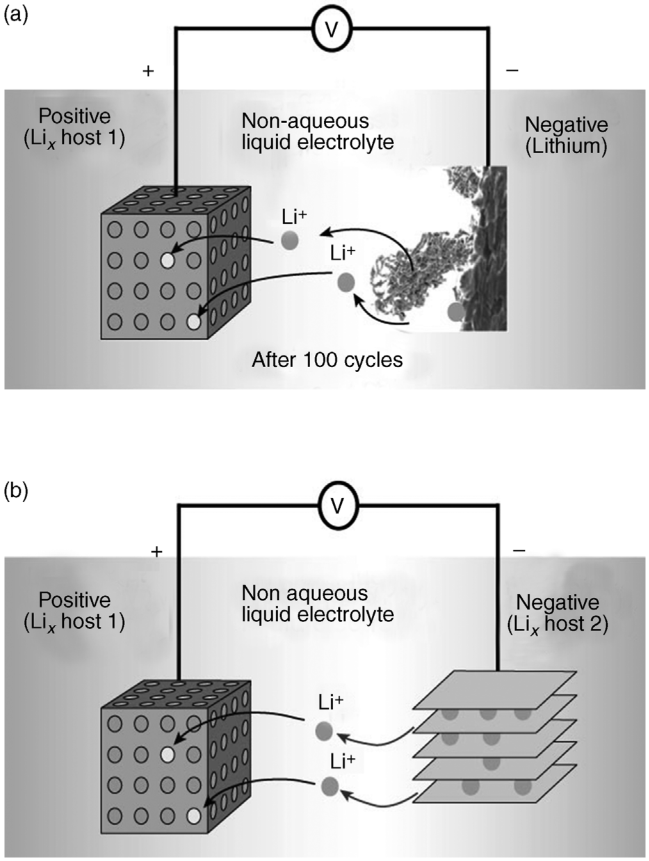

Figure 1.2 Operating principle [9] of rechargeable Li ion battery. Li ions diffuse from high-energy sites in the graphite anode to low-energy sites in the cathode, driving charge around the external circuit. Cells using $Li_{1-x}CoO_2$ (left) and Li_x intercalated into graphite (right) provide 3.6 V; energy densities of 120–150 W h kg^{-1} are widely used in portable electronic devices. Li ions reversibly enter (intercalate) and leave weakly bound positions between the graphene carbon layers of graphite.

Equation 1.1, the value of x goes completely to zero, leaving no residual Li in the cathode when fully charged.

The traditional lead–acid automobile battery consists of dense and heavy materials, lead and sulfuric acid, and is used primarily in high-current applications such as starting conventional automobile engines.

The chemical reaction for one form of Li ion cell is

$$Li_{1-x}CoO_2 + Li_xC_6 = C_6 + LiCoO_2, \qquad (1.1)$$

where the carbon is a graphite electrode. Ions, not electrons, are the current carriers. Note that lithium ions are not oxidized. In a lithium ion battery, the positive lithium ions flow internally from the graphite anode to the cathode, with the transition metal, cobalt, in $Li_{1-x}CoO_2$ being reduced from Co^{4+} to Co^{3+} during discharge. The performance may be as high as $160\,W\,h\,kg^{-1}$.

The performance of Li ion batteries can be improved by nanostructuring the electrodes, to enlarge the effective surface area. For example, it is stated that both higher power and higher storage capacity can be obtained by applying the active anode and cathode materials in a very thin film to copper nanorods anchored to sheets of copper foil. Enlarged electrode surface area by nanostructuring can increase the short-circuit current of a battery, assuming good electrical conduction to the nanostructured areas.

Nickel Metal Hydride A NiMH battery is a type of rechargeable battery, with 1.2 V nominal voltage, similar to a nickel–cadmium battery, but has an anode of hydrogen-absorbing alloy, instead of cadmium. The anode reaction occurring in a NiMH battery is

$$H_2O + M + e^- \leftrightarrow OH^- + MH. \tag{1.2}$$

The battery is charged in the forward direction of this equation and discharged in the reverse direction. Nickel(II) hydroxide forms the cathode. The metal M in the anode of a NiMH battery is typically an intermetallic compound. Several different intermetallic compounds have been developed, which mainly fall into two classes. The most common form of anode metal M is AB_5, where A is a rare earth mixture of lanthanum, cerium, neodymium, and praseodymium and B is nickel, cobalt, manganese, and/or aluminum. Other batteries use higher capacity negative electrodes based on AB_2 compounds, where A is titanium and/or vanadium and B is zirconium or nickel, modified with chromium, cobalt, iron, and/or manganese.

In any of these compounds, M serves the same purpose, reversibly forming a mixture of metal hydride compounds. When hydrogen ions are forced out of the potassium hydroxide electrolyte solution by the charging voltage, it is essential that hydride formation is more favorable than forming a gas, allowing a low pressure and volume to be maintained. As the battery is discharged, these same ions are released to participate in the reverse reaction.

NiMH batteries have an alkaline electrolyte, usually potassium hydroxide. A NiMH battery can have two to three times the capacity of an equivalent size NiCd battery. However, compared to the lithium ion battery, the volumetric energy density is lower. The specific energy density for the NiMH battery is approximately $70\,W\,h\,kg^{-1}$. An advantage of the NiCd battery over the NiMH is typically higher short-circuit current.

Supercapacitor, Ultracapacitor, and Double-Layer Capacitor A capacitor stores charge proportional to applied potential, $Q = CV$. Unlike a battery, there is no characteristic charging or discharging voltage V, and the voltage can be increased to a voltage limited by breakdown in the dielectric layer. The corresponding energy is

$$U = \frac{1}{2}CV^2 = \frac{Q^2}{2C}. \tag{1.3}$$

The familiar formula for a parallel-plate capacitor is

$$C = \frac{\kappa \varepsilon_0 A}{t}, \tag{1.4}$$

where A is the area, ε_0 is the permittivity of free space, and κ is the permittivity (relative dielectric constant) of the insulator, whose thickness is t.

Large-area supercapacitors. Within this framework, a large surface area leads to a large capacitance, and this effect is the basis for the most common form of supercapacitor [12]. One electrode is coated with a porous conducting medium of extremely high surface area. A nonaqueous electrolyte then fills in the interstices and makes contact with the counter electrode. The effective area is the area of the nanoporous electrode, which can be as much as 100 000 times the plate area [12]. This is a commercial product, usually based on carbon nanoparticles. The effective dielectric thickness is set by the interaction of the nonaqueous electrolyte with the porous electrode, the electric double layer. This thin dielectric will not allow a large voltage to be applied. According to Ref. [12], a capacitance of 5000 F can be obtained in a cell of dimensions 5 cm × 5 cm × 15 cm, similar to a D-cell battery. Banks of such capacitors have been used to run demonstration buses, trucks, and construction cranes. Recharging through braking of motion has also been demonstrated. A strong point of a supercapacitor is that it can be charged and recharged almost indefinitely, and it can discharge quickly, providing high power for a limited amount of time.

For a conceptual version of a high-performance supercapacitor, imagine that the nanotube "forest" shown in Figure 6.7 is immersed in a liquid electrolyte that fully penetrates the array of tubes, connecting to a second electrode. The high electrical conductivity and the connectivity of the tubes to the growth surface would make this device capable of high discharge rate, compared, for example, to a nanoporous array as shown in Figure 12.17 and schematically in Figure 12.18.

High-permittivity supercapacitors. A second approach [12] to a supercapacitor is through a dielectric of large permittivity, κ. As noted, the stored charge and the related energy, $Q^2/2C$, can be increased in proportion with κ.

Insulators such as the ferroelectric barium titanate ($BaTiO_3$) can have large permittivity κ, possibly even exceeding 10 000, while values for "high-κ" gate oxides as in the new Intel 45 nm node devices are in the range 20–30 [1]. The high permittivity of $BaTiO_3$ occurs at temperatures close to the critical temperature of a ferroelectric phase transition, and the value obtained may be sensitive to the composition and to the temperature. (A ferroelectric, analogous to a ferromagnet, which has a spontaneous internal magnetic field, exhibits a spontaneous internal electric field and surface charge density, opposite on opposing surfaces, as a result of distortion of the ion positions in the material.) The supercapacitor or ultracapacitor may be used as an adjunct to a battery, because, for a short time, it can provide a large short-circuit current, to provide high bursts of power.

1.1.2.2 Thermionic Emitters

Thermionic emission (electron escape from a metal) can occur for the fraction (usually very small) of free electrons in a metal whose energy E exceeds the electronic binding energy ($E_F + \varphi$, where φ is the work function) of that metal. The highest energy available to most electrons is the Fermi energy, E_F (or μ_F), and those electrons are contained inside the metal by the barrier energy $\Delta E = \varphi$, called the work function. At high temperature T, the chance f of an electron in the metal to momentarily occupy a state at energy $E_F + \varphi$, and thus to escape, is about

$$f \approx \exp\left(-\frac{\varphi}{k_B T}\right), \tag{1.5}$$

This is a simple estimate based on the Boltzmann classical distribution. (Recall that this distribution $f_{MB} = \exp(-E/k_B T)$ gives the distribution of speeds v in an equilibrium gas, where $E = mv^2/2$.) Work functions for metals lie in the range of a few electron volts, say 2–5 eV. At room temperature, $k_B T \approx 0.025$ eV. Here k_B, Boltzmann's constant, has the value $1.381 \times 10^{-23}\,\mathrm{J\,K^{-1}}$ or $8.63 \times 10^{-5}\,\mathrm{eV\,K^{-1}}$. Thus, the chance f for an electron in the metal to escape is roughly in the range $f = 1.8 \times 10^{-35}$ to $f = 1.38 \times 10^{-87}$ relative to the chance of finding that electron at the Fermi energy. This probability is exceedingly small, but strongly dependent on temperature and on the value of φ. At room temperature, a change in work function by $2.303 \times 0.025\,\mathrm{eV} = 0.057\,\mathrm{eV}$ changes the probability f (Equation 1.5) by a factor of 10.

The thermionic current density J_{th} leaving the metal is the product of the number R of impacts per second per square meter on the barrier and f. The rate R of electron impacts ($\mathrm{m^{-2}\,s^{-1}}$) from the interior of the metal onto the work function barrier is large, depending basically on how many electrons are present at the Fermi energy and how quickly they move. Accurate calculation of the emission current density requires concepts that will be developed later in Chapter 3. However, to get an idea of what is involved, we can use a classical estimate from the kinetic theory of gases for the rate R (units of $\mathrm{m^{-2}\,s^{-1}}$) of molecules crossing an arbitrary surface. The classical formula is

$$R = \frac{Nv}{4}, \tag{1.6}$$

where we will interpret N as the volume density of free electrons and v as the typical free electron speed. We can make a first rough estimate of the impact rate R for electrons on a metal's surface, even though, as we will see, electrons in a metal behave differently from molecules in a gas. For a typical metal, N is of order $10^{28}\,\mathrm{m^{-3}}$ and the typical velocity of electrons, the Fermi velocity

$$v_F = \left(\frac{2E_F}{m_e}\right)^{1/2}, \tag{1.7}$$

is $1.3 \times 10^6\,\mathrm{m\,s^{-1}}$, for $E_F = 5$ eV. Using Equation 1.6, we estimate $R = 3.3 \times 10^{33}\,\mathrm{m^{-2}\,s^{-1}}$, a very high rate of impacts on the work function barrier. As a reality check, we can apply our rough approach to electrons in a hypothetical metal similar to gold, at room temperature. These electrons certainly do not leave the metal. For this case, the work function φ is about 5 eV, so that the fraction of electrons energetically able to leave is about $\exp(-\varphi/k_B T) = 1.38 \times 10^{-87}$. The predicted rate of emission is $J_{th} \sim R\exp(-\varphi/k_B T) = 4.5 \times 10^{-54}$, consistent with the reality that electrons do not leave an ambient gold surface. The factor $\exp(-\varphi/k_{\mathrm{B}}T)$ increases rapidly with increasing temperature and with decreasing work function.

A typical operating temperature for a *thermionic* (electron emitting) filament or cathode is 2500 K, and for this temperature (forgetting that gold, unlike tungsten, would melt) our simple approach for the hypothetical metal would give a thermionic current density

$$J_{\mathrm{th}} \sim e\left(\frac{Nv}{4}\right)\exp\left(-\frac{\varphi}{k_{\mathrm{B}}T}\right), \tag{1.8}$$

which evaluates as $(1.6 \times 10^{-19}) \times (3.3 \times 10^{33}) \times [\exp(-5.0/0.2083)] = 1.98 \times 10^{4}\,\mathrm{A\,m^{-2}}$. (The symbol $\sim$ reminds us that this is a very rough, back-of-the-envelope estimate. Here the number 0.2083 eV represents the value of $k_{\mathrm{B}}T$at 2500 K.) But it turns out that this estimate is not too far off! This value for the current density is about seven times larger than a reported value, $0.3\,\mathrm{A\,cm^{-2}}$, for tungsten at 2500 K. (The discrepancy would be larger if we had used the work function, 4.54 eV, for tungsten.)

The collision rate R expression is changed by a more proper quantum approach, as we will see in Chapter 4. The *density of electron states* $g(E_{\mathrm{F}})$ (per unit volume and per unit energy) at the Fermi energy in a metal is $g(E_{\mathrm{F}}) = 3N/2E_{\mathrm{F}}$. This would then be multiplied by an energy range ΔE about $2k_{\mathrm{B}}T$ to give an effective number of electrons N'. Thus, $N' = g(E_{\mathrm{F}})\Delta E = N \times 3kT/E_{\mathrm{F}} = N \times 3 \times 0.2083/5 = 0.125N$, or a reduction by a factor of 8, and this removes the discrepancy mentioned. This change also brings in the temperature Tas a part of the rate, which is a step toward the correct expression (which is proportional to T^2). Still lacking is treatment of the fact that electrons at different energies above and below the Fermi energy see different barrier heights, smaller and larger, respectively, than the work function φ. So, in the end, an integration over energy is needed, with the exponential factor in the integrand. In fact, the angle of incidence of a given electron on the work function barrier has an effect, and these angles have to be integrated as well.

All such estimates are useful, in the opinion of the author, and a working technologist should not be afraid of making simple estimates to guide his thinking as well as to predict the size of an effect of interest.

Finally, full [13] analysis gives for the thermionic emission current density

$$J = e\left[\frac{4\pi m k_{\mathrm{B}}^2(1-r)}{h^3}\right] T^2 \exp\left(-\frac{\varphi}{k_{\mathrm{B}}T}\right), \tag{1.8a}$$

where r is a reflection probability of the electron at the metal surface. An evaluation of the correct prefactor gives $[e4\pi m k_{\mathrm{B}}^2/h^3] = 1.2 \times 10^{6}\,\mathrm{A\,m^{-2}\,K^{-2}}$. This formula (with $r = 0$) gives $0.25\,\mathrm{A\,cm^{-2}}$ if evaluated at 2500 K using the work function for tungsten, 4.54 eV, close to the experimental value. Here h, Planck's constant, is 6.6×10^{-34} J s.

One can see that thermionic emission of electrons is important for metals of small work function, at high temperature. The cathode of a diode vacuum tube is heated to produce a density of electrons in vacuum, which may be drawn to the positively charged plate electrode, and is usually large, favoring the emission of electrons. Typical vacuum tubes have a tungsten heater (high work function and high melting temperature) and a cathode (thermionic electron source) of a lower work function.

The hot cathode provides electrons by thermionic emission, which are then drawn to the high positive voltage of the plate to provide the forward current of the diode. The flow can be moderated by the potential of intervening grid electrodes in triode or screen grid electron tubes.

Another use of metals of high work function and high melting temperature, such as tungsten, is as an incandescent light source. The spectrum of light emitted is called the blackbody spectrum (see Equations 2.95–2.97) and an issue of current interest is the energy efficiency of generating light. (Why are "halogen" headlights, offered on expensive cars, brighter and more efficient? The answer is basically that a higher operating temperature is possible if the tungsten filament is enclosed in an iodine or bromine atmosphere. This has to do with another effect of high temperature, namely the chance for the metal itself to evaporate, which can cause the light to fail. It turns out that iodine or bromine vapor stabilizes the surface of tungsten against evaporation, which allows a higher operating temperature.) We will return to these issues.

1.1.2.3 **Field Emitters**

A metal, as we have seen, has an energy barrier φ that keeps electrons inside. If an electric field $E = V/t$ is applied, suppose a voltage $V = -50\,\text{V}$ is applied to a gold electrode, relative to a second gold electrode at $t = 10\,\text{nm}$ spacing, then the electric field is $5\,\text{V}\,\text{nm}^{-1}$. If the work function is 5 V, then an electron in the metal will see an energy barrier $\varphi = V_\text{B}$ of height 5 eV, which decreases to zero in a distance 1 nm. The electric field we have chosen is then $5\,\text{V}\,\text{nm}^{-1}$, a huge value. This 1 nm thick triangular barrier is so thin that the electron has a good chance to escape by *quantum mechanical tunneling*. Considering a simpler case, the probability T^2 for tunneling transmission through a square barrier (constant potential energy height V_B) is

$$T^2 = \exp\left[-\frac{2(2mV_\text{B})^{1/2}t}{\hbar}\right], \tag{1.9}$$

where $\hbar$ is Planck's constant, $h/2\pi = 1.1\times 10^{-34}\,\text{J s}$. In our example, we have $V_\text{B} = 5\,\text{eV} = \ 8\times 10^{-19}\,\text{J}$, and $m = 9.1\times 10^{-31}\,\text{kg}$ is the mass of the electron. For $t = 1\,\text{nm}$, for the square barrier the transmission probability is $T^2 = 1.1\times 10^{-10}$. Returning to the triangular barrier for field emission, we can see that the thickness of this electric field produced barrier, t, is determined by the energy barrier φ and the electric field strength, since $tEe = \varphi$, so $t = \varphi/eE$. If we put this into the formula (1.9) for T^2, and set $2\phi = V_\text{B}$, we get, as a back-of-the-envelope estimate,

$$T^2 \sim \exp\left[-\frac{2(2m)^{1/2}(\varphi/2)^{3/2}}{\hbar eE}\right]. \tag{1.10}$$

The corresponding field emission current density J will be approximately the collision rate R (calculated above) times the electron charge e times T^2. So the field emission current decreases exponentially as the 3/2 power of the work function, half power of the electron mass, and inversely with the electric field! (Here we have divided the work function by 2 as a rough approach to applying the square barrier formula to a triangular barrier.)

This formula predicts the correct dependence on work function and electric field E for the field emission current, but the numerical factor in the exponent needs to be changed when the triangular barrier is correctly treated. We can also imagine that a huge electric field will be needed, roughly on the scale of one Fermi energy μ_F per nanometer, since a nanometer is a typical distance to tunnel. So perhaps 5 V nm^{-1} or 5 GV m^{-1} will be needed.

For the actual current density, a first approximation might be $J = eRT^2$, where R is the rate of impacts mentioned above. However, this prefactor eR does not depend on electric field E and is incorrect. A more realistic approach [14] gives the field-emitted current density as

$$J = AE^2T^2, \tag{1.10a}$$

where T^2 is close to Equation 1.10, E is the surface field, and $A = [4(\mu_F\varphi)^{1/2}e^3]/[(\mu_F + \varphi)8\pi h\varphi]$.

Electron Field Emission from a Nanoscale Tip From the experimental point of view, it is relatively easy to get such a large electric field, by applying a modest voltage to a sharp tip. From elementary electrostatics, the potential V at a radius $r \geq a$, from sphere of radius a, at potential $V = V_0$, is

$$V = V_0\left(\frac{a}{r}\right), \quad r \geq a, \tag{1.11a}$$

corresponding to

$$E = -\frac{\mathrm{d}V}{\mathrm{d}r} = -\frac{V_0 a}{r^2}. \tag{1.11b}$$

For example, a potential of 50 V applied to a tip of radius $a = 10$ nm produces a field of 5 V nm^{-1} (5 GV m^{-1}) at the surface of the tip. (This field decreases as $(r/a)^{-2}$ for $r > a$.) The chance of tunneling for an electron from this tip is large, as suggested by the above calculation for the one-dimensional case. Since there are a lot of electrons in a typical metal, a large current will flow, maybe even enough, in this set of parameters, to melt the tip by Joule heating!

Field emission tips with radii on the 100 nm scale are used in high-resolution transmission electron microscopes (TEMs), to provide point sources of electrons of small *de Broglie wavelength*

$$\lambda = \frac{h}{p} = \frac{h}{mv} = \frac{h}{(2mV)^{1/2}}, \tag{1.12}$$

where m is the electron mass, 9.1×10^{-31} kg, h is Planck's constant 6.6×10^{-34} J s, and V is the voltage of the TEM. The de Broglie wavelength of the electron is a quantum property that tells how accurately the electron can be localized. The good thing about electrons is that this wavelength, for large voltage V, can be made much smaller than the wavelength of light, allowing a sharper image to form. Electron vacuum triode devices using field emission tips have been discussed, although not commercially produced. Arrays of tiny tips, as may be realized by a forest of *carbon*

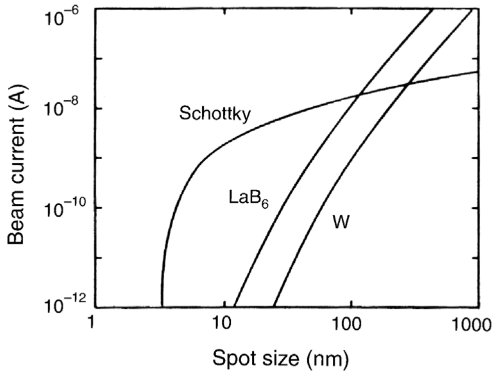

Figure 1.3 Performance [15] of point current sources for transmission electron microscopes. The spot size of the electron beam ranges from around 3 nm for a thermally assisted field emission (Schottky) source at 1 pA to around 1000 nm for a tungsten point providing a microampere. LaB_6 has a low work function, near 2.5 eV. A Schottky emitter is typically a sharpened W tip, perhaps coated with ZrO_2 to get a smaller work function. The Schottky source is run at an elevated temperature, which assists tunneling and also tends to remove undesired adsorbates from its surface.

nanotubes, are also discussed as a broad area source of electrons to illuminate a screen as in a television set (Figure 1.3).

A carbon nanotube provides about the sharpest tip that one can imagine because the tube has a diameter as small as a few nanometers. The nanotube conceptually results if a single sheet of graphite (pure covalently bonded carbon, in the form of hexagonal benzene rings, called a graphene sheet) is rolled to a radius of a few nanometers. An endcap for the cylindrical rolled graphite sheet can occur by a mixture of six- and five-membered carbon rings, leading to a closed tube whose diameter and tip radius are only a few nanometers. The electrical conductivity of these tubes depends on the choice of axis about which the underlying graphene sheet has been rolled; in many cases, the tubes are metallic conductors. Field emission from such a tip, thus also from an array of such tips, can produce useful field emission current densities even at a small applied tip voltage and low temperature.

1.1.2.4 **Ferroelectric and Pyroelectric Devices**

A crystalline electrical insulator such as quartz, SiO_2, or sapphire, Al_2O_3, is composed of repeating unit cells, each of which is electrically neutral. Of course, a neutral unit cell can still have an electric dipole moment, **p**, if the charge is not distributed evenly. In more detail, we can ask if the centroid (average position) of positive charge coincides with the centroid of negative charge. If not, the unit cell will have an *electric dipole moment.* (From elementary electrostatics, an electric dipole is defined as

$$\mathbf{p} = q\mathbf{l}, \tag{1.13}$$

where q is the charge and $\mathbf{l}$ is the vector displacement of opposite charges. The electric field from such a dipole is similar to the magnetic field from a bar magnet.) In a ferroelectric insulating crystal, such as $BaTiO_3$, below its nominal critical temperature of 132 °C, the unit cell has a distorted charge distribution leading to a nonzero dipole moment. Moreover, the distortions of the unit cells are all in the same direction, so that the "ferroelectric transition" is a cooperative distortion of all the unit cells in the crystalline sample (or in a domain of the sample). The result is a macroscopic polarization **P** defined as electric dipoles per unit volume. **P** is equivalent to a surface polarization charge density, $P = \sigma\,\mathrm{C\,m^{-2}}$.

The ferroelectric polarization **P** is closely analogous to the ferromagnetic magnetization **M.** It can be switched and the direction of polarization can be used to store information. A plot of P versus E (E is applied electric field) exhibits hysteresis. The acronym FeRAM is used to describe ferroelectric random access memory, in which information is stored in the polarization directions of tiny lithographically formed ferroelectric elements.

Materials in a class of ferroelectrics referred to as PZT are candidates for FeRAM cells. Here PZT stands for $\mathrm{PbZr_{1-x}Ti_xO_3}$ (lead zirconium titanate). (These PZT materials are also widely used for their related piezoelectric properties.) Second, for temperatures approaching the critical temperature from above, the permittivity (relative dielectric constant) increases, peaking at T_c. (The value of the permittivity for a crystal of $\mathrm{BaTiO_3}$ can exceed 10^4 at T_c, but will generally be lower for a thin film.)

The surface charge density $P = \sigma$ varies with temperature (this is the *pyroelectric effect*) according to the relation

$$\Delta P = p_{\mathrm{Py}} \Delta T, \tag{1.14}$$

where p_{Py} is the pyroelectric coefficient. The ferroelectric thus can serve as a source of charge and current, as its temperature is changed in the vicinity of its critical temperature.

This is the basis for one type of imaging infrared sensor array that provides night vision. The thermal electromagnetic waves (blackbody radiation) emitted by the object are imaged onto the sensor array, and the intensity on a given element (pixel) proportionally changes its temperature, and hence its surface charge. The surface charge image is read out as a voltage array and a grayscale image is formed.

1.1.3 Generators of Alternating Current and Voltage: AC

Sources of alternating voltage and current can be classified according to frequency, and familiar examples range from 60 Hz to radio and optical frequencies. In all cases, electromagnetic waves are generated. An important range for telecommunications is the THz range of light, which is carried along optical fibers.

1.1.3.1 Faraday Effect Devices

Faraday's law states that

$$\mathcal{E} = -\frac{\mathrm{d}\Phi_{\mathrm{M}}}{\mathrm{d}t}, \tag{1.15}$$

where Φ_{M} is the magnetic flux and $\mathcal{E}$ is the voltage appearing at the terminals of a loop enclosing the magnetic flux. This is the basis for electrical power generation.

Faraday's law is also the basis for the electron synchrotron, which accelerates charged particles around a path enclosing a changing magnetic flux, and in the presence of a static magnetic field perpendicular to the path, to keep the charged particle in orbit.

This law also applies to the generation of a voltage pulse in the RSFQ computer technology [4], where a single quantum of magnetic flux,

$$\Phi_0 = \frac{hc}{2e} = 2.07 \times 10^{-15}\,\text{Wb} \tag{1.16}$$

(the Weber is the SI unit of magnetic flux) trapped in a loop containing one or two Josephson junctions, disappears. Here h is Planck's constant, c the speed of light, and e the electron charge. The resulting voltage pulse, which carries the information in the RSFQ computer technology, can be described as

$$\int V\,\mathrm{d}t = 2.07 \times 10^{-15}\,\text{Wb} = 2.07\,\text{mV ps}. \tag{1.17}$$

Typically, the pulse will be a few millivolts in amplitude and last for about a picosecond.

1.1.3.2 **Crystal Oscillators**

A high-frequency voltage source of particular importance in computer technology is the crystal oscillator, in which the frequency determining element is a crystalline slab of piezoelectric quartz. The piezoelectric effect produces an internal electric field proportional to the deformation, or strain, of the material, which varies with time as it is put into oscillation. Electrodes on the broad faces of the slab acquire charge as the crystal distorts (its thickness changes) and the resulting voltage is amplified and fed back to sustain the oscillation. The mechanical oscillation of the quartz crystal has a high quality factor

$$Q = \frac{\omega_0}{\Delta\omega}, \tag{1.18}$$

where $\Delta\omega$ is the width of the peak and the center frequency ω_0 is stable. The oscillation is conceptually identical to (it is a distributed version of) a mass on a spring, whose frequency is

$$\omega_0 = \left(\frac{K}{m}\right)^{1/2}, \tag{1.19}$$

where K is the spring constant in N m^{-1} and m is the mass in kg. In the quartz crystal case, the effective spring constant is related to the Young's modulus, and the mass is distributed over the slab. The Q is high because the distortion of the quartz is purely elastic. Such oscillators are not integrated onto the chip, however. The typical oscillator frequency is in the MHz range and may be multiplied up to the GHz range by circuitry on the chip.

An exposed quartz crystal is used as a "thickness monitor" in vacuum deposition systems. The deposit increases the mass of the oscillating crystal, reducing the oscillation frequency and allowing the thickness of the deposit to be inferred from the small frequency shift, which can be accurately measured because of the high *Q*. Deposits down to 0.1 nm thickness can be measured. A frequency counter can measure a MHz signal down to 0.1 Hz resolution, if the signal is stable in time.

1.1.3.3 Gunn Diode Oscillators

A Gunn diode exhibits a "negative differential resistance," with $\mathrm{d}I/\mathrm{d}V < 0$ in a range of applied voltages *V*, forming the basis for an oscillator. The device operation depends on a sophisticated property of "conduction bands" in certain "indirect bandgap" semiconductors that contain bands of free electrons of differing "mobilities," related to their differing "effective masses." (The mobility, μ, of a carrier in a semiconductor is defined by the relation

$$v_\mathrm{D} = \mu E, \tag{1.20}$$

where v_D is the average velocity acquired by the carrier in an applied electric field *E*. The basic formula for mobility is

$$\mu = \frac{e\tau}{m^*}, \tag{1.21}$$

where m^* is the electron effective mass and τ is the time between scattering collisions of the carrier. For GaAs, the low mass group of carriers, which are electrons at the conduction band minima, have $m^* = 0.068 m_\mathrm{e}$, where $m_\mathrm{e} = 9.1 \times 10^{-31}$ kg.)

In general, the current density *J* in a semiconductor, as will be covered in Chapter 4, is the sum of electron and hole currents,

$$J = (N_\mathrm{e}\mu_\mathrm{e}e + N_\mathrm{h}\mu_\mathrm{h}e)E = \sigma E. \tag{1.22}$$

Here N_e and μ_e, are, respectively, the number density and mobility of electrons (similarly for holes), *e* is the electron charge, $e = 1.6 \times 10^{-19}$ C, and σ is the resultant electrical conductivity expressed in Siemens ($\Omega^{-1}\,\mathrm{m}^{-1}$). This is an expression of Ohm's law. The electrons move in the conduction band whose energy is raised by the bandgap energy E_g with respect to the valence band. In an n-type semiconductor of heavy doping N_D, the free electron density may approach N_D, and the free hole density will be negligible. The opposite is true of a p-type semiconductor of heavy acceptor doping N_A: the hole density may approach N_A, and the electron density will be negligible. A hole is created when an electron in a filled covalent bond is promoted to the *conduction band*, which costs energy E_g, on the order of 1 eV, which is closely related to a covalent bond energy.

The hole can propagate, since a nearby electron can jump into the vacant position. The charge of the hole is positive, therefore, and its mobility is the mobility associated with an electron in the band in question. Typically, hole masses are large and hole mobilities are low. A donor impurity has one more electron than the host semiconductor, and this electron typically is excited to a conduction band leaving behind positively charge donor ions of density N_D^+. An acceptor impurity has one fewer electron than the host. Typically, the acceptor impurity takes an electron from the host, leaving a hole, excited into the valence band, leaving negatively charged acceptor ions of density N_A^+. In case of very large doping of either sign, "metallic conduction" occurs in which the carriers leave the source impurities and move about as much as they would in a metal.

Gunn diodes are typically made of n-type GaAs, usually as a sandwich of lightly doped n-GaAs between two layers of heavily doped n-GaAs. Such diodes, with a suitable driver circuit, can oscillate up to about $f = 200\,\text{GHz}$, corresponding to a vacuum electromagnetic wavelength $\lambda = c/f = 1.5\,\text{mm}$ or $1500\,\mu\text{m}$. Here c is the vacuum speed of light, $3 \times 10^8\,\text{m}\,\text{s}^{-1}$. Gunn diodes similarly made of GaN can support oscillations up to 3 THz, corresponding to vacuum wavelength as small as $0.1\,\text{mm} = 100\,\mu\text{m}$. GaN is a wide bandgap semiconductor that has become important as an element of blue light emitting lasers.

1.1.3.4 **Esaki Diodes**

The Esaki diode is a "heavily doped" semiconductor pn junction in which the n- and p-regions have very high densities of electrons and holes, respectively. In the language of the previous section, a metallic p-region in the Esaki diode directly abuts a metallic n-region. At the actual discontinuity, the junction, the result is an electrical dipole layer, with a corresponding jump V_J in electrostatic potential across the "depletion region" whose width is

$$W = \left[\frac{2\varepsilon\varepsilon_0 V_J(N_D + N_A)}{e(N_D N_A)}\right]^{1/2}. \qquad (1.23)$$

In this formula, ε is the relative permittivity of the semiconductor (11.8 for Si) and ε_0 is the permittivity of vacuum, $\varepsilon_0 = 8.85 \times 10^{-12}\,\text{F}\,\text{m}^{-1}$. In the Esaki diode, the barrier junction width w is so small that electrons in the valence band of the p-region can easily tunnel into empty states in the conduction band of the n-region. This process produces an anomalous "hump current" at low forward bias in the I–V relation, which is followed by a region of "negative differential resistance," $\mathrm{d}I/\mathrm{d}V < 0$. This forms the basis for an oscillator. The junction voltage, V_J, is related to the energy shift ΔE of the semiconductor bands across the junction, $V_J = \Delta E/e$. ΔE is usually smaller than the semiconductor bandgap E_g, unless the semiconductors are heavily doped and metallic. In that case, ΔE is E_g plus the sum of the Fermi degeneracies in the n- and p-regions. The available values of the width w are of order 10 nm, and Esaki diode oscillators can operate in the GHz range. It is not possible to integrate these devices into the planar silicon technology, so they have been superseded by devices simpler to construct and amenable to large-scale integration.

A device that is similar to the Esaki diode in its operating principle and that has been implemented in the planar silicon technology is the Si/SiGe "resonant interband tunnel diode" (RITD), described in Chapter 8. This *heterojunction* between Si and $\text{Si}_{1-x}\text{Ge}_x$, which can be grown on an Si wafer and for which the energy bands across the junction can be arranged similarly as in the Esaki diode, allows tunneling transport. It is important in a useful heterojunction (between chemically different crystals) to match the dimensions of the unit cells, so that the same crystal structure occurs on either side of the junction. This situation is described as "heteroepitaxy" and can occur as the second layer is carefully deposited on the first, with the atoms finding the same lattice positions on either side.

1.1.3.5 Injection Lasers

An injection laser converts DC electrical power into monochromatic light, and the familiar form is the laser pointer. An important injection laser is used to generate light of approximately 1000 nm wavelength to illuminate optical fibers in telecommunications. Similar devices are used in compact disk (CD) and digital video disk (DVD) players to read and to store information. An injection laser, briefly, requires a pn junction biased at a large positive voltage, $V \approx E_g/e$, so that large electron and hole currents flow oppositely through the device, meeting at the junction. The injection laser is designed to encourage recombination of the electrons and holes (producing light) in a zone centered on, but wider than, the depletion region (whose width w was quoted above). The desired result is emission of light of frequency f approximately $f = E_g/h$. The frequency f exactly matches a particular resonant frequency mode,

$$f_{m'} = \frac{2c^*}{m'L}, \quad \text{for integer } m', \tag{1.24}$$

of a mirror cavity of length L centered on the junction. Here $c^* = c/n$, with n the refractive index. m' is an integer representing the number of light half-wavelengths that fit inside the partially reflecting walls of the cavity. The phenomenon of stimulated emission of radiation can allow a single electromagnetic mode to be excited, so that the device is a laser (light amplification by stimulated emission of radiation) leading to coherent radiation at a single wavelength (frequency). Coherent light leaking out from the partially reflective mirrors has a well-defined direction and frequency, coming from a single electromagnetic mode in the cavity.

Strong light can also be emitted under less stringent conditions, the device in this case will be called an LED (light emitting diode). The electrical efficiency of the conversion and the minimum electrical current density (laser threshold) are important to applications. Great effort has been expended to extend laser action to shorter wavelengths, mainly by finding new semiconductors of larger bandgap, such as InGaN and GaN, and finding ways to make high-quality pn junctions in the new materials. The advent of the “Blu-ray” blue–violet laser of the short wavelength (405 nm) has made possible DVDs of smaller pixel size and higher information density. Sony Corp. in an April 23, 2007 news release announced cumulative sale of 2 billion laser diodes for CD and DVD applications. The company also announced new facilities that can manufacture millions of the new Blu-ray devices each month. The power of these devices can be in the 100 mW range.

Intensive research and development in injection lasers and LEDs has been directed to lower the current threshold (to improve battery life) and increase the conversion efficiency of electrical to light energy, and also to extend the wavelength range from the infrared and red to the blue region of the visible spectrum. This corresponds to lowering the wavelength of the emitted light, which allows the light to be focused on a smaller area, allowing for increased information density.

Incandescent electric lights provide about 20 lm W^{-1}, while compact fluorescent bulbs provide about 60 lm W^{-1}. Present commercial LEDs provide around 30 lm W^{-1}, but 100 lm W^{-1} is achieved in laboratory tests. (The actual energy efficiency, photon energy per second in the visible range divided by power from the plug, is not revealed by these figures. The energy efficiency of incandescent lights is so low, perhaps 5%, that they have been banned in several countries.)

Arrays of LED devices are familiar, for example, in automobile taillights, portable flashlights and electric lanterns, and digital displays. Dense arrays of LED devices, patterned lithographically, are widely used in digital displays, providing a less expensive technology than the color TV vacuum tube.

In the color TV vacuum tube, a single electron beam scans across an array of three colored phosphor pixels, and the beam current is synchronously modulated to create light and dark regions. The difficulty is that the scanned electron beam requires the bulky and expensive vacuum tube. Flat screen displays are much preferred, and presently the leading technology is the LCD (liquid crystal display). The modulation of the light is accomplished by electric fields that align the liquid crystal domains. This is a large specialized area, which we will not dwell on.

Cost and efficiency factors have also driven development of organic light emitting diodes (OLEDs), which are competitors for displays such as the LCD television and also for illumination, where efficiency is important.

1.1.3.6 **Organic Light Emitting Diodes**

Dye molecules have a large literature and many uses. The basic effect is strong absorption of light of a particular wavelength $\lambda = hc/\Delta E$, where ΔE is the energy difference between two electronic states of the molecule. After absorbing a photon and changing to the excited electronic state, the molecule typically relaxes slightly to a lower energy form of its excited state, from which it can return to the ground state by emitting a longer wavelength photon $\lambda' = hc/\Delta E'$. This emission process is the basis for generating light in the OLED. The device requires two electrodes that promote excitation of dye-like molecules, which then emit light. Instead of exciting the dye-like molecule with shorter wavelength light, the excitation is arranged by removing an electron from the lower energy state by having it tunnel to the anode (it may be said that a hole tunnels from the anode to the molecule) and also an electron transfers to the higher energy level of the molecule from the cathode. Then the molecule emits light as the electron in the excited level falls into the vacant electron state (hole) in the ground state of the molecule. Recent important advances in this area are described in Chapter 5.

Two OLED applications are high-power devices to be used for illumination, with interest in having a higher efficiency of illumination than in incandescent light bulbs and in making arrays that serve the same purpose as the color television tube or LCD television display.

In the first application, it appears that OLED devices are capable of efficiencies in the range 40 lm W^{-1}, which exceeds incandescent light efficiencies, which are in the range below 20 lm W^{-1}. (The lumen is a unit of luminous flux, a measure of the light perception of the eye. The energy flux, different from the luminous flux, is termed the radiant flux.) A difficulty, now overcome, has been in finding reliable OLED devices to operate in the blue region, so as to provide white light in combination with light from earlier developed OLED devices in the red and green regions. (It appears that the power per OLED device is lower than available in semiconductor pn junction devices as are used for CD and DVD players, and also lower than incandescent and the more efficient compact fluorescent light bulbs.)

To create a color display, three different molecules may be chosen to emit light in the red, green, and blue regions, and an array of pixels can be addressed by evaporated patterned metal "wires" in a crossbar array. One set of wires is evaporated on a semitransparent oxide glass electrode, the three different molecules can be evaporated onto pixel locations, and the upper array of wires is then deposited by evaporation from a source. Addressing one upper wire and one lower wire of the crossbar array will allow a single pixel to be illuminated.

The fabrication processes used here are quite similar to, but cheaper than, those used in silicon device technology. The size of the individual pixel, the basic light generating diode element, can be very small, limited by the lithographic process, which, as we have seen, is now capable of wires of 45 nm width. This type of display can be arranged on a flexible substrate.

1.1.3.7 **Blackbody Emission of Radiation**

A surface at absolute temperature T radiates electromagnetic radiation with total power, per unit "black" ($e = 1$) surface area,

$$P = e\sigma_{\mathrm{SB}}T^4, \tag{1.25}$$

where the Stefan–Boltzmann constant $\sigma_{\mathrm{SB}} = 5.67 \times 10^{-8}\,\mathrm{W\,m^{-2}\,K^{-4}}$ and the emissivity e can be less than 1.0 for reflective surfaces. (The same surface, as we have seen, may also have tendencies to emit electrons and even atoms, but these effects are usually negligible.)

Blackbody radiation is an important effect in cooling a hot surface, even at modest everyday temperatures where electron and atom emissions are absolutely negligible.

(For example, an unclothed 310 K Eskimo of area 1 m^2 in an igloo of temperature 273 K will suffer by radiation a loss of energy $\Delta P = \sigma_{\mathrm{SB}}(310^4 - 273^4) = 5.67 \times 10^{-8} \times (9.23 - 5.55) \times 10^9 = 208.6\,\mathrm{W}$, assuming emissivities are unity. (If the igloo were sufficiently reflective, the life-threatening radiation loss could be avoided.))

The wavelength distribution of blackbody radiation (see Equation 2.95) has a peak at wavelength λ_{m} such that

$$\lambda_{\mathrm{m}}T = \text{constant} = 2.9 \times 10^6\ \mathrm{nm\ K}. \tag{1.26}$$

The value of $\lambda_{\mathrm{m}} = 486$ nm for the solar spectrum is in the visible range corresponding to $T \approx 5973$ K, while $\lambda_{\mathrm{m}} = 9.67\,\mu\mathrm{m}$ is in the infrared range for $T = 300$ K. This thermal radiation, which corresponds to $E = hf = hc/\lambda_{\mathrm{m}} = 0.128\,\mathrm{eV}$, can be imaged by night vision systems, see Equation 1.14.

The spectrum in Figure 1.4 has closely the shape of the Planck blackbody radiation spectrum, plotted versus wavelength, for 5973 K. This spectrum was measured in vacuum above the Earth's atmosphere and directly measures the huge amount of

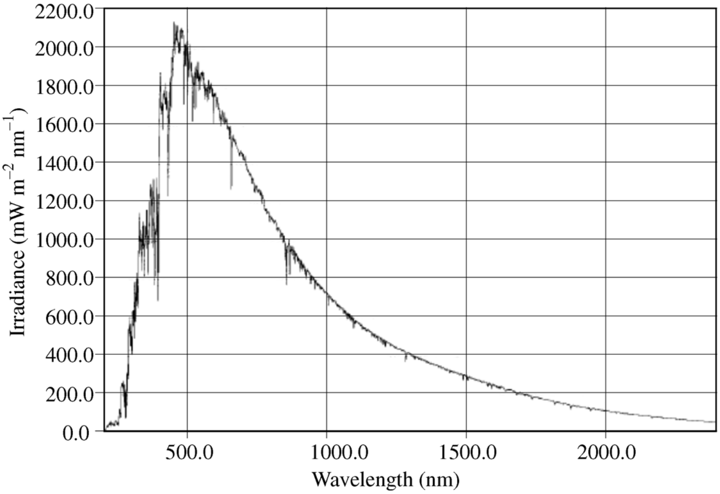

Figure 1.4 Directly measured solar energy spectrum, from 200 to 2400 nm, from a satellite-carried spectrometer [16]. The units are related to energy, mW m^{-2} nm^{-1}, and the area under this curve should be close to 1366 W m^{-2}. The sharp dips are atomic absorption lines (see Chapter 3) presumably from atoms in the atmosphere surrounding the Sun. An Earth's surface spectrum of solar light is shown in Chapter 12. Note that the peak here is close to 486 nm, corresponding (see the text) to a blackbody at 5973 K. The portion of this spectrum beyond about 700 nm cannot be seen, but represents infrared heat radiation. In a similar spectrum from a light bulb, the portion beyond 700 nm is wasted.

energy perpetually falling on the Earth from the Sun, quoted as 1366 W m^{-2}. This spectrum was measured by an automated spectrometer carried in a satellite well beyond the Earth's atmosphere. The sharp dips in this spectrum are atomic absorption lines, the sort of feature that can be understood only within quantum mechanics, as will be described in Chapters 2 and 3. The atoms in question are presumably in the Sun's atmosphere.

The incandescent light bulb is approximately a blackbody emitter, and as we have mentioned the rating is about 20 lm W^{-1}. The lumen is a unit of luminous flux, which differs from radiant energy flux, because the sensitivity of the human eye is included in the definition. If the incandescent light emits photons in the infrared range (heat, not light), these are part of the radiant flux but are not considered as part of the luminous flux as they are not detected by the eye.

1.1.4 Detectors

Sensors or detectors of electrical phenomena include single devices like a Geiger counter for ionizing radiation, or a voltmeter, and also arrays of detector devices, like the retina of the eye or the array of charge-coupled detector (CCD) devices in a digital camera. Such array devices may be ranked according to the spatial resolution or image detail available. All detectors are usefully compared in their sensitivity to the fundamental limits that may apply. For example, the eye can detect a few photons, the quantum limit for light, while a photomultiplier can detect a single photon with certainty.

1.1.4.1 Photomultiplier and Geiger Counter

A photomultiplier vacuum tube detects light by the photoelectric effect, in which absorption of a single quantum of light ejects an electron from a metal. The condition is that the quantum energy of the light, hf (where h is Planck's constant and f the frequency of the light), must exceed the work function φ of the metal. In the vacuum photomultiplier tube, a cascade of metallic electrodes is placed past the initial photodetecting surface, with a positive bias voltage V between succeeding multiplier electrodes.

The initial photoelectron is accelerated to energy $eV \gg \varphi$, so that its impact onto the first multiplier electrode will release several, say n, secondary electrons. The secondary electrons in turn are accelerated to the second multiplier electrode where a further multiplication occurs.

If there are N multiplier electrodes (these are called dynodes), each with a multiplier factor n, the single photon can lead to n^N electrons being collected! So we can define the photomultiplier gain as

$$G = n^N, \tag{1.27}$$

with n the number of secondary electrons and N the number of dynodes.

This can produce from a single photon a measurable electron charge in a short time interval, to give a current pulse. For example, if $n = 3$ and $N = 12$, the multiplication factor is about half a million, 0.531×10^6.

The "drifted germanium detector" is an elegant solid-state device that can accurately measure the energy of a single absorbed high-energy photon, called a gamma ray. A very pure single crystal of germanium is treated in a special way with lithium, so that the mean free paths of electrons and holes are larger than the dimensions of the device. The germanium is pure and is cooled to 77 K, the temperature of liquid nitrogen, so that the number of thermally generated free electrons and free holes is low, and only a small background current flows. Metallic electrodes apply a bias voltage across the germanium, and a sensitive detector of the charge flow in the circuit is provided. A gamma ray, that is, a photon of high energy (MeV) $E_\gamma = hf \gg E_g$, is absorbed, creating initially an electron in the conduction band and a hole in the valence band, whose kinetic energies add up to $hf - E_g$. The final result from the single absorbed gamma ray photon, as a result of internal processes in the germanium, is n electrons and n holes flowing oppositely across the crystal and into an external charge meter, such that the total measured charge

$$Q = n2e = 2e\left(\frac{hf}{E_g}\right) = \int I(t)\,\mathrm{d}t. \tag{1.28}$$

From this relation, the gamma ray energy hf is deduced from the measured charge Q and the known value of E_g. So

$$E_\gamma = \left(\frac{Q}{2e}\right) E_g. \tag{1.29}$$

The carrier multiplication process that occurs within the semiconductor, usually extremely pure germanium, finally creates $n = hf/E_g$ electron–hole pairs from the initial high-energy electron–hole pair. The high-energy carriers reduce their kinetic energy by collisions creating additional electron–hole pairs of lower energy, and in this process loss of energy to lattice vibrations (phonons) apparently is not an important effect. A similar process would be beneficial in the efficiency of solar cells, to capture energy from the portion of the solar spectrum in the range $hf > E_g$.

1.1.4.2 **Photodetector, Solar Cell, and pn Junction**

A common photodetector is called a photodiode, which may be called a pin diode or a pn diode. The active element is a semiconductor (or insulator, in the pin case), and the sensitivity is to light whose quantum energy hf is larger than the bandgap of the semiconductor or insulator. Small bandgap materials, for example, InSb, are needed for sensitivity to infrared light. These devices are operated with an applied bias voltage, and the absorption of a quantum of light produces electrons and holes. The photocarriers flow through the circuit and measurement of the current is the means of detection of light.

The solar cell is a pn or metal–semiconductor junction, which is optimized to extract electrical energy with high efficiency by absorbing sunlight (see Chapter 12). In this mode, there is no external bias, and the pn junction is connected to a load, for example, a resistive element or a battery to be charged. Recall the description above of a pn junction, which has a depletion region of width w sandwiched between n- and p-type semiconductors. The depletion region is an electric dipole layer, with positive charge on the n-type side (produced as the donor impurities release their electrons to fall into the acceptor impurity sites on the p-side) and thus negative charges on the p-side. The result of the electric dipole layer is a voltage drop V_J across the depletion layer, and consequently, a large internal electric field V_J/w inside the width w.

The solar cell is designed so that one of the electrodes is thin, allowing sunlight to penetrate into the depletion layer. The action of light is to generate electron–hole pairs in and near the region of the junction. The action of the internal electric field is to accelerate electrons and holes in opposite directions and finally to drive electrons around the external circuit. It turns out that the active layer for light absorption is much wider than the depletion width w, because minority carriers (holes on the n-side and electrons on the p-side) can drift or diffuse a relatively large distance, called the recombination length L, and these carriers have a high chance of encountering the junction and falling down the potential barrier. So the active region has a width about the sum of the recombination length for electrons plus that for holes.

1.1.4.3 Imaging Detector, CCD Camera, and Channel Plate

A prototype imaging detector is the silver halide photographic film, which records a two-dimensional light image. This image is difficult to read out for later processing of any kind. A second prototype imaging detector is the retina of the eye, where time-dependent signals are transmitted by the optic nerve for processing in the brain. The optic nerve allows parallel readout from individual rod and cone cells.

Consumer photography has shifted from silver halide film, which requires a difficult chemical process before the image can be visualized, to digital cameras using arrays of CCDs. The charge-coupled detector is typically an array of metal–semiconductor contacts, on a silicon surface, whose metal electrodes are so thin as to be transparent. Exposure to a light image produces a charge image, namely a set of photoinduced charges behind each metal electrode. The quality of the silicon is high enough that the charge image is stable against recombination over a time sufficient to provide a complex sequential readout of the charge image. Using a complex electrode structure with a repeating sequence of electrical pulses, the spatial information is transformed into a sequential time signal on a single wire, or perhaps a few wires, rather than a parallel array of wires reaching each pixel. The essential part of this process is "charge coupling," which is repetitive transfer of photoinduced charges from one pixel to the next, until at the final pixel the charge is transformed to a sequence of voltage pulses on a readout electrode. All of these detectors are capable of detecting a small number of photons, close to the limit of sensitivity.

A channel plate is an amplifying device, in principle like an array of photomultiplier tubes. In addition, the device can be used to amplify an image constituted of electrons rather than photons. In one form, the device is an array of micrometer diameter holes etched into an insulating plate such as alumina, with a large voltage across the plate, which accelerates electrons that fall into the individual tubes. Secondary emission occurs as the electron cascades under the applied electric field down the tube, and the charge bundle exits the tube into vacuum. The output of the channel plate, which operates in high vacuum, is an electron density image, and this image may fall on a phosphor surface to turn the electron image back into an amplified optical image.

1.1.4.4 SQUID Detector of Magnetic Field and Other Quantities

This device is a closed loop containing one or two *Josephson tunnel junctions*. The junctions can be configured to measure changes in the magnetic flux intercepting the closed loop, with an extraordinary sensitivity, better than one flux quantum,

$$\Phi_0 = \frac{hc}{2e} = 2.07 \times 10^{-15}\,\mathrm{Wb} \tag{1.30}$$

(the Weber is the SI unit of magnetic flux). If, for example, the closed loop has 1000 turns of wire and the enclosed area is 1 cm × 1 cm, then the magnetic field intensity B to change the enclosed flux by one quantum is $B = 2.07 \times 10^{-15}\,\mathrm{Wb}/(1000 \times 10^{-2} \times 10^{-2}) = 2.04 \times 10^{-14}\,\mathrm{T} = 2.04 \times 10^{-10}\,\mathrm{G}$. This is a very small magnetic field; in comparison, the Earth's magnetic field is 0.3–0.6 G, in different locations. So the SQUID (superconducting quantum interference device) is the most sensitive detector of magnetic field and can be converted into the most sensitive detector of current by use of a coupling coil carrying the current, with a resistor added to the coupling coil to detect a small voltage applied across the resistor. All SQUID devices operate at cryogenic temperature, needed to provide the superconducting state of the SQUID loop.

1.1.5 Two-Terminal Devices

Continuing a survey of device types, next considering "two-terminal" devices.

1.1.5.1 Semiconductor pn Junction (Nonohmic)

The semiconductor pn junction is the most important device in semiconductor electronics. It has a nonlinear (nonohmic) current–voltage characteristic and acts similarly to a vacuum diode in basically passing a current in only one direction. So it is a rectifier, creating a DC voltage when impressed with an alternating voltage. It is, as described above, the result of joining an electron-conducting semiconductor to a hole-conducting semiconductor.

1.1.5.1.1 I(V) Characteristic and Temperature Dependence The pn junction, described above, exhibits the *I*–*V* relation

$$I = I_0\left[\exp\left(\frac{eV}{k_B T}\right) - 1\right], \tag{1.31}$$

where the constant I_0 (the reverse current) is strongly temperature dependent, varying approximately as $\exp(-eV_J/k_BT)$, where V_J is the voltage shift across the junction. The nominally positive or forward sign of V is chosen to reduce the band bending in the junction, that is, to raise the energy of the electrons in the n-side of the junction. The temperature dependence is the basis for the use of the junction as a thermistor, a temperature sensor. The band bending voltage V_J can approach E_g/e in magnitude, where E_g is the semiconductor bandgap energy. The value is typically in the range 0.5–1 V for common semiconductors. V_J is also approximately the maximum open-circuit voltage if the pn junction is used as a solar cell. A similar *I*–*V* relation is often observed for metal–semiconductor contacts, or Schottky diodes.

1.1.5.1.2 Action as Voltage Variable Capacitor: Varactor The pn junction causes a jump V_J in electrostatic potential across the "depletion region" whose width $w = [2\varepsilon_0 V_J(N_D + N_A)/e(N_D N_A)]^{1/2}$. In this formula (1.23), N_D and N_A are the dopant densities, ε is the relative permittivity of the semiconductor (11.8 for Si), and ε_0 is the permittivity of vacuum, $\varepsilon_0 = 8.85 \times 10^{-12}\,\mathrm{F\,m^{-1}}$. If the device is biased by a voltage V, the width w changes simply as

$$w(V) = \left[\frac{2\varepsilon\varepsilon_0(V_J - V)(N_D + N_A)}{e(N_D N_A)}\right]^{1/2}. \tag{1.32}$$

Here V is chosen as in the diode formula as the direction of "forward bias" can be seen to reduce the band bending and the width of the depletion region.

For use of the junction as a voltage variable capacitor or *varactor*, the useful voltage range is the reverse voltage range,

$$w(V_{\mathrm{Rev}}) = \left[\frac{2\varepsilon\varepsilon_0(V_{\mathrm{J}} + V_{\mathrm{Rev}})(N_{\mathrm{D}} + N_{\mathrm{A}})}{e(N_{\mathrm{D}}N_{\mathrm{A}})}\right]^{1/2}. \tag{1.33}$$

The charge Q stored by the capacitor, most simply expressed when the doping densities are equal, is $Q = AN_{\mathrm{D}}w/2$, with A the junction area, and formally leads to the capacitance as $C = \mathrm{d}Q/\mathrm{d}V$. The result can also be expressed in terms of the usual flat-plate capacitor formula

$$C = \frac{\varepsilon\varepsilon_0 A}{w(V_{\mathrm{Rev}})}. \tag{1.34}$$

The reverse voltage increases the width of the junction and also increases the electric field in the depletion region $E = V_{\mathrm{Rev}}/w(V_{\mathrm{Rev}})$.

1.1.5.1.3 **Action as Voltage Clamp: Zener Diode** In strong reverse bias $V_{\mathrm{Rev}} \gg k_{\mathrm{B}}T/e$, the nominal current is $-I_0$ and is the result of diffusion of minority carriers. Referring to the above, we can see that the junction electric field $E = V_{\mathrm{Rev}}/w(V_{\mathrm{Rev}}) \approx \mathrm{const} \times (V_{\mathrm{Rev}})^{1/2}$ for $V_{\mathrm{Rev}} \gg V_{\mathrm{J}}$. As first explained by Clarence Zener, a large "interband tunnel current" can abruptly (but reversibly) occur in reverse bias. The onset voltage for Zener breakdown is very reproducible and temperature independent for a given junction. This effect is used to provide a clamping action in circuit applications.

The simplest explanation is that electrons in the valence band of the p-material are able to tunnel across the depletion region to empty conduction band states on the n-side of the junction, when the junction electric field in reverse bias is large enough. This produces a large current and the resistance of the device abruptly becomes small for voltages exceeding the reverse breakdown voltage. The value of the breakdown voltage is controlled by the level of doping in the semiconductor.

1.1.5.1.4 **Action as Photodetector and Energy Transducer: Solar Cell** The pn junction is sensitive to light photons whose quantum energy hf exceeds the bandgap energy E_{g}, which must be absorbed within a diffusion length of the junction. So the effective photons are absorbed within a distance L_{p} of the junction in the p-region and within a distance L_{n} of the junction in the n-region. Recall that the reverse current I_0 is the sum of diffusion currents of minority carriers (electrons in the p-region, which can diffuse a length L_{p}, and holes in the n-region, which can diffuse a length L_{n}). The light-generated extra carriers then have a high probability of diffusing to the junction and being swept across by the junction electric field. If the device is connected to a short circuit, the result will be a current ideally of two electron charges per absorbed photon. This is the action of a pn junction photodetector.

If the pn junction is irradiated under open-circuit conditions, it will develop an open-circuit voltage because the photogenerated electrons and holes will occupy the ionized donor and acceptor states and thus neutralize the space charge region. The limiting open-circuit voltage will correspond to the shift of the Fermi levels between the n- and p-materials, which itself must be less than E_g/e. This process can be thought of as charging the capacitor represented by the pn junction. If the pn junction is connected to a resistive load, power will be delivered to the load. It turns out that the limiting efficiency of converting light energy to electrical energy is in the vicinity of 20% for crystalline silicon solar cells in sunlight. An important consideration in the efficiency is the relation of the bandgap energy to energy distribution of photons in sunlight. If the bandgap is too large, none of the light will be absorbed, and a solar cell will not be possible. It is also important that the semiconductor be of the type (indirect bandgap) where recombination of electrons and holes is relatively slow.

1.1.5.1.5 **Action as Source of Light: Injection Laser** A different application of the pn junction is to produce light. The semiconductor in this case must have a "direct bandgap" so that recombination of electrons and holes can occur quickly if they are nearby. In this case, a heavily doped pn junction is placed in strong forward bias, so that the height of the electrostatic barrier for electrons to cross into the junction region is reduced from V_J to $V_J - V$. This forward bias increases (by $\exp(eV/k_BT)$) the density of electrons and holes at the junction, and thus increases their rate of recombination back to photons.

To make a laser (light amplification by stimulated emission of radiation), the junction is enclosed by partially reflecting mirrors of spacing L, so that emitted light is trapped, especially if its wavelength $\lambda' = c/nf$ (n is the refractive index of the semiconductor) matches one of the geometric resonances $\lambda' = 2L/m'$, where m' is a positive integer. A variety of design features are employed to reduce the chance that the carriers cross the junction, that is, to increase the chance that they recombine radiatively in the junction region. An active field in semiconductor technology has been finding new semiconductors, such as GaN, suitable for injection laser operation in a wide range of wavelengths, especially to extend the range toward the blue in the optical spectrum.

1.1.5.2 Metal–Semiconductor Junction and Alternative Solar Cell

An abrupt metal–semiconductor contact is called a Schottky junction. In such a junction, a space charge region is usually present, as in the pn junction, and an electrostatic barrier V_J appears, leading to the same sort of rectifying $I(V)$ as was discussed for the pn junction. In this case, the origin of the electrostatic barrier is different (see Figure 4.13). At the metal–semiconductor interface, one has terminated the semiconductor on a plane surface, interrupting the periodic potential for electrons that occurs otherwise. Special electronic states, called "surface states" arise, with a spectrum of energies crossing the energy gap of the semiconductor. Typically, the spectrum of surface states of the terminated semiconductor will have a peak in the vicinity of the middle of the energy gap. For a surface on an n-type semiconductor with donor density N_D, the result is that a depth l_n of the semiconductor is depleted; that is, the donor impurity sites are empty (positively charged), the electrons having transferred to the surface states. This will leave an electrostatic barrier of height V_J and consequent width

$$w = \left[\frac{2\varepsilon\varepsilon_0 V_J}{eN_D}\right]^{1/2}. \tag{1.35}$$

Two comments: first, this formula describes an ohmic contact to a semiconductor, as a Schottky barrier contact (metal deposited directly on a clean but highly doped semiconductor) for which the depletion barrier width is so small that tunneling occurs rapidly. To make an ohmic contact, the semiconductor has to be heavily doped, at least on a local basis.

Second, the metal–semiconductor contact is a prototype for a solar cell. For a solar cell, one needs an electrostatic barrier V_J, which is accessible to the sunlight. Light absorbed more than a diffusion length away from the junction does no good, and in the wavelength ranges $\lambda < hc/E_g$ the light is heavily absorbed and does not penetrate far into the semiconductor. The metal layer in the Schottky barrier contact can be very thin, allowing light to enter, and at the same time the metal layer is highly conductive and can be relied upon to carry the light-induced current to the external circuit.

1.1.5.3 **Tunnel Junction (An Ohmic Device)**

A tunnel junction is conceptually a very thin capacitor, such that the electron wavefunctions from one metal electrode extend across the dielectric into the opposite electrode. The chance of an electron from one metal abruptly appearing on the opposite side becomes large and dominant for the current across the junction. The I (V) for this device is linear with small corrections quadratic in voltage, so the device is basically an ohmic resistor. The prototype system in which this understanding was achieved is the Al–Al oxide–Al system studied by Ivar Giaever. Aluminum metal quickly forms an oxide, but the depth to which the oxide grows is uniform, well defined, and small: a few angstrom (0.1 nm) units in a few minutes of atmospheric exposure. (The oxide growth, with zero applied potential, stops quickly as its thickness increases, which makes aluminum cans virtually indestructible and an ecological hazard.) Giaever studied this system and unequivocally demonstrated electron tunneling (as an elastic, loss-free, and very fast random process, see Section 8.2.4) with the extra result of confirming the Bardeen, Cooper, and Schrieffer theory of superconductivity, which led to a Nobel Prize in Physics, for Ivar Giaever, Leo Esaki, and Brian Josephson. The current across the tunnel junction has a particular frequency spectrum relating to uncorrelated random quantum jumps across (through) the tunnel barrier.

1.1.5.4 **Josephson Junction**

The Josephson junction (see Section 8.4 for details) is a tunnel junction with special properties because both its electrodes are superconducting. These topics are important because they underlie the most active area in quantum computing in 2008, which is the adiabatic quantum computer. This computer is based on superconducting flux quanta, as will be described in Chapter 11.

Superconductivity is described as a macroscopic quantum state, so that the rules of quantum mechanics, which usually work only for very small systems such as electrons orbiting in atoms, apply precisely to a big piece of superconductor, which may even be miles long.

A basic idea of quantum mechanics is that there is a wavelength λ associated with a matter particle. It was found that a beam of electrons, when passed through a double slit, exhibits interference fringes on a screen behind the slits. The behavior is exactly as when monochromatic light is passed through a double slit, to produce an interference pattern. Fitting the results for electrons indicates that the wavelength associated with a mass particle is $\lambda = h/p$, where p is the momentum mv and h is Planck's constant, $h = 6.6 \times 10^{-34}$ J s.

A plane-polarized light wave can be described by its electric field

$$\mathbf{E} = \mathbf{E}_0 \exp\left[i2\pi\left(\frac{x}{\lambda} - \frac{t}{T}\right)\right] = \mathbf{E}_0 \exp[i\theta(x,t)]. \tag{1.36}$$

The wave is coherent if it is periodic with wavelength λ over large distances Δx and also perfectly exhibits the period Tover large time intervals Δt. For many purposes, the time dependence can be factored out, so we can write

$$\mathbf{E} = \mathbf{E}_0 \exp[i\theta(x)]\exp(-i\omega t), \tag{1.37}$$

where $\omega = 2\pi/T$. Since the electron behavior is identical with that of light waves, regarding the interference effects, we are led to associate with an electron a wavefunction, called ψ, with the same dependence on position and time. We change the equation only by expressing the wavelength in terms of the momentum, $p = h/\lambda$. So

$$\psi = \psi_0 \exp\left[\frac{i2\pi p}{h}\right]\exp(-i\omega t) = \psi_0 \exp\left[\frac{ip}{\hbar}\right]\exp(-i\omega t), \tag{1.38}$$

where $\hbar = h/2\pi$.

The interference of electron waves, it turns out, is not precisely the same as for light waves, when there is a magnetic field. If a magnetic field is present in the region of the two slits, the interference pattern formed by electrons reaching the screen behind the slits is actually found to shift! The exact modification of the wavefunction of the electron is replacement of the momentum, $\mathbf{p}$, by $\mathbf{p} - e\mathbf{A}$, where $\mathbf{A}$ is the magnetic vector potential, defined by

$$\mathbf{B} = \Delta \times \mathbf{A}. \tag{1.39}$$

So the final form of the wavefunction for an electron of perfectly defined momentum $\mathbf{p}$ in the presence of a magnetic field (represented by a vector potential $\mathbf{A}$) is

$$\Psi = \Psi_0 \exp\left[i\left(\mathbf{p} - \frac{e\mathbf{A}}{c}\right)\cdot\frac{\mathbf{r}}{\hbar}\right] = \Psi_0 \exp[i\Theta]. \tag{1.40}$$

Superconductivity is a peculiar state of a metal where all of the N free electrons are grouped into $N/2$ pairs, each having charge $-2e$ and momentum $\mathbf{p}$ zero. The "pairs" are all in exactly the same quantum state "$\mathbf{p} = 0$". Thus, pairs of charge $2e$ are described by the pair wavefunction

$$\Psi = \Psi_0 \exp[i\Theta] = \Psi_0 \exp\left[i\left(-\frac{2e\mathbf{A}}{c}\right)\cdot\frac{\mathbf{r}}{\hbar}\right], \tag{1.41}$$

since the pair momentum is zero.

Superconducting Flux Quantum The correctness of this description is demonstrated by the observation of quantization of the magnetic flux. Consider a closed loop of superconductor and enforce the usual condition that the wavefunction $\Psi_0 \exp[i(-2e\mathbf{A}/c)\cdot\mathbf{r}/\hbar]$ shall be single valued on the closed loop. This means that

$$\int \left(-\frac{2e\mathbf{A}}{c}\right)\cdot\frac{\mathbf{r}}{\hbar}\,\mathrm{d}\mathbf{r} = n2\pi, \tag{1.42}$$

where n is an integer.

By evaluating this integral with use of the definition $\mathbf{B} = \Delta \times \mathbf{A}$ and Stokes' theorem, it follows that

$$\iint B\,\mathrm{d}^2\mathbf{r} = \Phi_\mathrm{M} = n\Phi_0 = n\left(\frac{hc}{2e}\right) = n \times 2.07 \times 10^{-15}\,\mathrm{W}. \tag{1.43}$$

This effect has been observed and provides strong evidence for the superconducting state as a coherent macroscopic quantum system. This effect and the very small value of the flux quantum, as was mentioned in connection with the SQUID, underlie extremely sensitive detection of magnetic field B.

The implied coupling between all of the pairs means that a current can flow only if all $N/2$ electron pairs in the metal do *exactly* the same thing. It is as if the pairs are in lockstep, like soldiers in parade. For a typical good metal, N is about $10^{28}\,\mathrm{m}^{-3} = 10^{22}\,\mathrm{cm}^{-3}$, so in the superconductor there are 10^{22} electrons per cubic centimeter doing exactly the same thing. The statement is that all pairs have the same phase Θ. Professor John Bardeen, who was one of the three persons to theoretically describe superconductivity, is quoted as saying that the superconducting phase can be "coherent over miles of ... wire."

If there is a current density J, say of 1 A cm^{-2}, the usual formula $J = Nev$, where v is the electron velocity, so $J = Nev = (10^{22}\,\mathrm{cm}^{-3})(1.6 \times 10^{-19})v$, implies $v = 0.63 \times 10^{-3}$ cm s^{-1}. This is a small velocity, applied to all electrons, and represents a small shift in the momentum P of the electron system from its original value, which was zero. Since all of the electrons do the same thing, there is no scattering and no electrical resistance (no voltage is needed to maintain this current density). The defining statement for a superconductor is that a current can flow at zero electric field, and if system is undisturbed the current will continue for a very long time.

Experiments have verified these statements. A current density for a coherent electron wave may be described as

$$J = J_0 \exp\left[i\left(\mathbf{p} - \frac{e\mathbf{A}}{c}\right)\cdot\left(\frac{\mathbf{r}}{\hbar}\right)\right] = J_0 \exp[i\Theta], \tag{1.44}$$

where Θ is the phase and $\mathbf{A}$ is the magnetic vector potential such that the magnetic field is $\mathbf{B} = \Delta \times \mathbf{A}$. The presence of the magnetic vector potential in the phase means that important interference effects will be governed by a magnetic field if present.

The Josephson junction is a thin tunnel barrier between two superconductors, such that a small supercurrent,

$$J = J_{0\mathrm{J}} \sin\theta \tag{1.45}$$

up to a limiting value $J = J_{0\mathrm{J}}$, can pass (no voltage across the barrier), where $\theta = \Delta\Theta$ is the jump of superconducting phase across the junction. So, no current flows unless there is a phase change θ across the junction. The prefactor $J_{0\mathrm{J}}$ is dependent on the tunnel barrier and contains the typical factor of the type $T^2 = \exp[-2(2mV_\mathrm{B})^{1/2}t/\hbar]$, where V_B is the tunnel barrier height, m the electron effective mass, and t the thickness of the barrier, as described in Section 2.6.4.

1.1.5.5 Resonant Tunnel Diode (RTD, RITD)

A resonant tunnel diode, see Figure 8.1, is a series combination of two tunnel barriers spaced by a nanoscale distance L, giving it a nonlinear current–voltage characteristic $I(V)$. The operation of this device depends upon transient trapping of an electron in one of a possible set of localized states between the two barriers. The transient states are defined essentially by the condition that n half-wavelengths for the electron fit into the barrier spacing L. The energies of the trapped states are approximated by a simple quantum formula, $E_n = n^2h^2/8mL^2$, with corresponding nonlinear features in the $I(V)$ characteristic at voltages $V_n = E_n/e$.

This is a quantum device depending on the fact that an electron has an associated wavelength $\lambda = h/p$. As is characteristic of quantum devices, the time scale for operation is short, so that it can be used as a high-frequency oscillator or in fast switching applications. In the most common case $n = 1$, there is one relevant trapped electron state, and a single anomalous peak (hump) in the $I(V)$ current is observed. The performance of the device is given by the ratio of the current at the peak of the hump divided by the lower current observed at the minimum or valley following the peak. This ratio is termed the PVCR.

A particular type of resonant interband tunnel diode has been successfully implemented by growing layers including $\mathrm{Si}_{1-x}\mathrm{Ge}_x$ with x about 0.4 on a p-conducting silicon substrate, see Figure 8.6. This produces a band structure that resembles that of the Esaki diode. For this type of device, a peak-to-valley ratio (PVCR) of 6 has been reported.

1.1.5.6 Spin-Valve and Tunnel-Valve GMR Magnetic Field Detectors

A change in electrical resistance with applied magnetic field, magnetoresistance, is one of the methods of measuring a magnetic field. The new spin-valve and tunnel-valve nanoelectronic devices depend upon the fact that the electron has an inherent spin, $S = 1/2$, which can be oriented in one of the two possible directions. One of the consequences of electron spin is ferromagnetism, the state of some metals in which all of the electron spins in a region called a domain are aligned parallel, either in the spin-up or in the spin-down direction, and this direction in a small film of ferromagnet will be sensitive to an ambient magnetic field, such as the one produced in the vicinity of magnetized domain written into the surface of a computer hard disk.

The new nanometer-scale devices are termed giant magnetoresistance (GMR) devices simply because the resistance changes that quantum effects produce in the nanometer scale exceed those available in conventional devices. The GMR detectors have been pivotal in allowing higher storage capacity on disk drives because these detectors can be made much smaller than the earlier detectors made with conventional pickup coils. The Nobel Prize in Physics in 2007 was awarded to Albert Fert and Peter Grunberg for their early work on the physics of the diffusive version of giant magnetoresistance effect.

The GMR spin-valve magnetic field detector device is actually a sandwich of a film of metal like copper, but having only nanometer-scale thickness, between two thin ferromagnetic films. The ferromagnetic outer films of the spin valve are engineered in such a way that one film will be easily switched in its magnetization direction by the ambient magnetic field, while the other film will have its magnetization direction fixed. The basic quantum effect in the GMR spin-valve detector is that the electrical resistance along the sandwich, provided by electrons moving in the center copper film, differs slightly if the magnetization directions of the two outer films of the sandwich are parallel or antiparallel. The detector works only if the sandwich spacing is on the nanometer scale, so that the spin directions of the electrons carrying the measured current are unchanged as they pass through the device.

The GMR tunnel-valve device, see Figure 8.12, is used in the most advanced disk drives such as those in the iPod and also in a form of memory known as magnetic random access memory (MRAM). This detector is simply a tunnel junction, as described above, having ferromagnetic electrodes. One is a "soft" ferromagnet (easily switched in its magnetization direction by the ambient magnetic field) and the other is a "hard" ferromagnet, whose magnetization direction is fixed. It is found that the tunnel current and therefore the resistance of the tunnel junction differs significantly, especially if the tunnel barrier is made of the insulator MgO, if the magnetization directions of the ferromagnetic electrodes are parallel or antiparallel. This is a pure quantum effect, arising from the quantum mechanical tunneling effect. It depends upon the fact that the electron states in a ferromagnetic metal are shifted for electrons whose spin directions are parallel versus antiparallel with respect to the electron spins of the ferromagnetic domain. So the electron density of states in the ferromagnet is different for spin-up versus spin-down electrons. The further effects are that (a) electron tunneling occurs without changing the spin direction of the tunneling electron and (b) the tunneling rate is proportional to the density of initial and final states, on opposite sides of the tunnel barrier.

The GMR tunnel-valve device is used for a new magnetic random access memory in computer technology. In this case, a binary bit of information is represented by the two states of the relative magnetization of the ferromagnetic electrodes, parallel versus antiparallel. One advantage of MRAM is that the information is preserved if the electrical power is removed, and such memory is called nonvolatile. Conventional dynamic random access memory (DRAM) in present computers has to be refreshed on a continuing basis, which increases energy consumption and means that all information is lost if the electrical power is shut off.

1.1.6 Three-Terminal Devices

A prototype three-terminal device is the vacuum triode, consisting of a cathode, an anode, and a control grid. Three electrodes are needed for any kind of control function, so this class of devices is extremely important. Transistors are control and storage devices in the semiconductor technology and are more important for our purposes than vacuum tube devices of any sort.

1.1.6.1 **Field Effect Transistor**

The field effect transistor consists of source drain and gate, which have the same functions, respectively, as the cathode, anode, and control grid of the vacuum triode. The FET is fabricated on the surface of silicon. An electron-conducting n-FET starts with a crystal of p-type silicon, provided with metallic N^{++} source and drain contacts, to define the ends of the "channel" and allow extraction of the device output current. The channel length l is basically the spacing between these two metallically doped contacts. An insulator is established on the top of the channel to insulate it from the gate electrode, which then covers the electron channel between the source and the drain. A positive gate voltage is applied to "invert" the surface of the p-type semiconductor, so that it (the conductive channel) is populated with electrons in number determined by the gate voltage V and gate capacitance C through a relation similar to $Q = CV$. A voltage applied between source and drain leads to the output current of the device, by electron flow along the channel. The smallest dimension in the FET has been the thickness of the insulation of the gate electrode. The gate insulator originally was thermally grown silicon dioxide.

The superiority of the field effect transistor to the vacuum triode has many aspects. The vacuum tube is bulky, fragile, and energy consuming. These were reasons that semiconductor research was pushed in the hope of providing miniaturized computers for the purposes of the space mission to the moon. Once the basic silicon transistors were invented, alleviating all of the above-mentioned difficulties of the vacuum tube, the real advantage of the semiconductor technology became apparent, which is its susceptibility to being miniaturized. The idea of "scaling" has guided the semiconductor industry for decades and has led to the cost per transistor falling exponentially with time.

The present price per transistor in large chips (Chapters 7 and 13) is probably less than 20 cents per million, a fact of great importance for the prospects of any successor to the silicon technology. (Such an estimate might attribute \$100 of the roughly \$1000 cost of a good laptop computer to the silicon chip, which contains about 0.5 billion devices.) The device density has risen as described by Moore's law, with continuous reductions in channel length and gate insulator thickness. It was long been recognized that scaling as originally conceived, using the thermal silicon dioxide insulator, can work only down to thicknesses large compared to the silicon and oxygen atomic radii in the SiO_2, needed to preserve the desired insulating property. As mentioned above, Intel has now abandoned the scaled silicon dioxide to insulate the gate electrode, in favor of artificially deposited "high dielectric constant" oxides based on the heavy metals hafnium and zirconium. Literally, the thermally grown silica, forced thinner and thinner by the scaling formula, contained only a handful of silicon atoms across its thickness, allowing electrons to "leak" through by the quantum mechanical tunneling effect.

The dominant computer logic technology continues to be “CMOS”: complementary metal oxide semiconductor field effect transistors. The technology at present has reached a size scale where the smallest features are 45 nm in width. The problem of tunneling through the gate oxide has been solved by making the insulator thicker, which has required choosing new materials with a larger permittivity than silicon dioxide. The capacitance of gate electrode to the silicon crystal has to be maintained sufficiently large, to maintain a sufficiently large number of carriers in the channel, and thus a sufficiently large output current, in on condition of the device with positive gate voltage. The length l of the channel has also fallen into the same 45 nm range, where quantum effects such as the longitudinal confinement described above in terms of the resonant tunneling diode come into play. The present situation with scaling is that the basic fabrication facility has been sufficient, without need to change the wavelength of the light used in the photolithography. The main change has been in replacing the thermal oxide with an atomic beam deposition process for the high-permittivity oxide dielectric.

The high electron mobility transistor (HEMT) is based on the FET as described above, but the channel between the source and the drain is characterized by high electron mobility. This means that carriers can transit more rapidly between source and drain. HEMT devices are usually fabricated in the GaAs system of semiconductors, which include GaAlAs and GaInAs, and are based on *heterojunctions* between these materials. This system is also favorable because the electron mobilities are much higher than in silicon, for the reason that the effective masses are smaller than in Si. The mobility of electrons at 300 K in GaAs is reported as $8500\,\mathrm{cm^2\,V^{-1}\,s^{-1}}$ compared with $1500\,\mathrm{cm^2\,V^{-1}\,s^{-1}}$ in Si. A heterojunction is a junction between two chemically distinct semiconductors, but usually great care is taken so that the crystal lattice is undisturbed at the junction. The growth of one layer on the other is said to be *epitaxial*, if the same crystal lattice type and dimension are reproduced in the deposited layer. This is desirable, because electron scattering at the interface is minimized.

Epitaxial heterojunctions are typically fabricated in an ultrahigh vacuum chamber, and the method is sometimes called MBE (molecular beam epitaxy). An HEMT based on GaAs may have a channel based on an epitaxial junction between heavily doped n-type AlGaAs (larger bandgap) on undoped GaAs. AlGaAs has a larger bandgap than GaAs, in such a way that at an interface the conduction band of GaAs occurs at a lower energy than in AlGaAs. For electrons in the GaAs, the doped layer acts as a barrier. (Figures 4.4b, 5.25, and 8.1 show aspects of this situation.) In this case, however, the AlGaAs has donor impurities. Thus, the donor electrons in the AlGaAs fall into the lower energy conduction band of the adjoining pure GaAs layer, where they move with high mobility because the GaAs is extremely pure. If the pure GaAs layer is also thin, then the electrons are said to form a two-dimensional electron gas (they can move freely parallel to the interface, but are trapped in the third direction by the edges of the GaAs). The HEMT transistor (also known as heterojunction FET or HFET) is a premium device typically used in very high frequency applications.

1.1.6.2 **Bipolar Junction Transistors: npn and pnp**

The original transistor was of the bipolar junction (BJT) type, rather than the FET type as described above. These devices are conceptually most easily described as a series connection of two oppositely oriented pn junctions, with the "base region" between the junctions being very thin. To make an npn transistor, two closely spaced n-regions could be diffused into the surface of a piece of p-type semiconductor, with corresponding electrodes for emitter, base, and collector. The width of the base region, which in an npn junction bipolar transistor is p-type material, is smaller than the diffusion length for minority carriers (electrons), so that electrons injected from the emitter largely diffuse across to the collector electrode. In other words, they are likely to diffuse across the base to the base/collector junction, where they are swept forward to its n-side by the internal electric field of the base/collector junction, which is enhanced by being reverse biased at voltage V_c.

The base electrode bias controls the rate at which electrons enter the base from the emitter, thinking of the emitter/base junction as a forward biased (base positive) diode whose current I into the base region will be $I = I_0 \exp(eV_B/k_BT)$. The output voltage, the product of the emitter current I (assuming it all crosses the base into the collector) and the collector resistance R_c, exceeds the emitter voltage IR_e in the ratio $R_c/R_e > 1$. In the common-base connection, here will also be power gain in the same ratio, $I^2R_c/I^2R_e > 1$. The device can also be used as a switch, in a common-emitter circuit: if the base electrode is reverse biased, the emitter collector current, which can be large, will be sharply cut. The base current is nearly zero.

The pnp bipolar transistor works in the same way, with the roles of electrons and holes being interchanged.

The heterojunction (HBT) bipolar junction transistor, which may be either npn or pnp, is a device whose principle is the same as described above, but is capable of higher frequency operation. It is typically fabricated using the methods of epitaxial growth in an ultrahigh vacuum chamber, sometimes called molecular beam epitaxy. ("Epitaxy" implies that the same crystal lattice type and dimension are reproduced in a deposited layer, so that electron scattering at the interface is minimized.) One consequence of epitaxial fabrication is that the thickness of the base region can be greatly reduced. It is also possible to have a different bandgap and different carrier mobilities in the base region. The resulting devices are typically used in radio frequency circuits. Aluminum gallium arsenide and germanium–silicon are two materials systems used for heterojunction bipolar transistors.

1.1.6.3 **Resonant Tunneling Hot-Electron Transistor (RHET)**

A resonant tunneling emitter has been incorporated in a class of junction transistors, which offer opportunities for multiple stable states, suitable for certain memory and logic applications (see Figure 8.4). If the emitter layer is replaced by a resonant tunneling diode with two accessible trapped states, then electrons of different energies may be injected into the base layer under different emitter–base bias conditions. Devices of this sort with multiple emitters have also been proposed and tested. These devices have several forms but have not found significant application.

1.1.7 Four-Terminal Devices

1.1.7.1 Thyristors: npnp and pnpn

Thyristors are large four-terminal high-power devices used for controlling electric motors, for example. A related set of devices is silicon controlled rectifiers. These devices have four alternating semiconductor layers and can be regarded as a series connection of three diodes of alternating polarity. A latching operation is achieved in these devices when avalanche breakdown occurs in one of the diodes, which must be reverse biased in this case.

1.1.7.2 Dynamic Random Access Memory

DRAM is the basic memory used in silicon computing electronics. The size of the repeating cell in memory chips has decreased with Moore's law scaling. A feature of this type of memory is that leakage occurs and the memory has to be rewritten on a short time scale, which is frequently 64 ms for a row of cells. The repeating cell in DRAM consists of one capacitor and one transistor, which may be compared to the requirement for six transistors per cell to make static (latching) random access memory.

1.1.7.3 Triple-Barrier RTD (TBRTD)

Devices with two quantum wells have been designed and tested. Such devices exhibit multivalued $I(V)$ curves, which may, for example, provide three different voltages at which the same current is measured. One such device is described in connection with Figure 8.4.

1.1.8 Data Storage Devices

Storage devices can be categorized broadly into optical storage devices, such as CD and DVD, and electrical storage devices, such as DRAM and SRAM (static random access memory). Electrical or computer storage devices are categorized as random access versus disk (magnetic storage) with disk memory offering larger storage but much longer access time. RAM is subdivided into volatile and nonvolatile, regarding retention of the information in the absence of power. Storage may also be categorized by the physical mechanism of storage. From a physical basis, the information may be represented by electric charge, by direction of ferromagnetic magnetization, by direction of ferroelectric polarization, and by the amorphous versus crystalline state of a small region in what is called phase change memory (PCM).

1.1.8.1 Optical Memory Devices

The principal types of interest are the CD and DVD. In both cases, information is read from the pixels of the optical medium by a focused laser beam. In the case of the

DVD-R (a digital video disk that can be rewritten), a material is employed that readily can be shifted from a crystalline to an amorphous state. The two states have a different optical reflectivity values, and this allows readout of information. In this case, the disk can be written and read by the same laser beam.

1.1.8.2 Electrical Computer Memory Devices

Two conventional forms of memory, DRAM and SRAM, have been defined above. DRAM operates by storage of electrical charge on a capacitor, which, because of leakage, has to be refreshed. SRAM is achieved in a variety of ways, but one basic type is a flip-flop arrangement of six transistors that have two latched overall states.

New forms of RAM under current development include

(a) MRAM, which is nonvolatile and is based on the magnetic tunnel junction. See Chapter 10.

(b) PRAM (phase change random access memory) based on the same type of alloys as used in the DVD-R disks, which can exist in crystalline (high electrical conductivity) and amorphous (low electrical conductivity) states and can switch from one to the other by electrical pulsing as well as by laser irradiation. This form of memory is nonvolatile. It appears that this type of memory is in the process of being miniaturized and may supplant flash memory, which is widely in use at present.

(c) FeRAM (ferroelectric random access memory) in which the direction of electrical polarization of a ferroelectric material, which can be switched, stores information. This is also nonvolatile.

Another important type of memory is called flash memory, which is also a form of charge storage memory. Here the information is permanent, but the number of erasures and rewrites is limited. In this type of memory, charge is stored in the gate oxide of a transistor. Erasure of the stored charge is a relatively slow process. Flash memory is used in convenient “thumb drives” that store 8 GB and cost less than $50, plug into USB ports on most computers. Flash drives are also used in the iPOD Nano personal recorder devices and in the iPhone. It appears that larger flash drives, in the range up to 128 GB, are being developed as replacements for traditional hard drives in small laptop computers.

The basic hard drive of the desktop and laptop computer is the magnetic disk memory. As mentioned above, the development of the GMR reading head has made larger capacity, smaller size, and lower cost available. The magnetic disk is used in the iPOD personal music players, but in iPOD Nano devices the magnetic disk is replaced by flash memory of up to 8 GB capacity. The access time of the magnetic disk for writing and reading is limited by the mechanical nature of the device, which involves a spinning disk and a reading head that moves to access the different data tracks at different radii on the spinning disk. The mechanical nature of the disk drive also makes it susceptible to error or damage by mechanical vibration or shock.

References

1 Mistry, K., Allen, C., Auth, C., Beattie, B., Bergstrom, D., Bost, M., Brazier, M., Buehler, M., Cappelani, A., Chau, R., Choi, C., Ding, G., Fischer, K., Ghani, T., Grover, R., Han, W., Hanken, D., Hattendorf, M., He, J., Hicks, J., Heussner, R., Ingerly, D., Jain, P., James, R., Jong, L., Joshi, S., Kenyon, C., Kuhn, K., Lee, K., Liu, H., Maiz, J., McIntyre, B., Moon, P., Neirynck, J., Pae, S., Parker, C., Parsons, D., Prasad, C., Pipes, L., Prince, M., Ranade, P., Reynolds, T., Sandford, J., Shifren, L., Sebastian, J., Seiple, J., Simon, D., Sivakumar, S., Smith, P., Thomas, C., Troeger, T., Vandervoorn, P., Williams, S. and Zawadzki, K. (2007) Electron Devices Meeting (IEDM 2007), December 10–12, 2007, IEEE International, pp. 247–250.

2 Colin Johnson, R. (2007) Quantum computer Orion debuts. EE Times, February 8, 2007. van der Ploeg, S., Izmalkov, A., Grajcar, M., Hubner, U., Linzen, S., Uchaikin, S., Wagner, Th., Smirnov, A., Van den Brink, A., Amin, M., Zagoskin, A., Il'ichev, E.and Meyer, H. (2007) *IEEE Transactions on Applied Superconductivity*, **17**, 113.

3 Grover, L.K. (1999) Quantum mechanics helps in searching for a needle in a haystack. *Physical Review Letters*, **79**, 325.

4 Likharev, K.K. and Semenov, V.K. (1991) *IEEE Transactions on Applied Superconductivity*, **1**, 3.

5 The International Technology Roadmap for Silicon, on the Web at www.public.itrs.net/.

6 Ning, T.H. and Taur, Y. (1998) *Fundamentals of Modern VLSI Devices*, Cambridge University Press, Cambridge.

7 Goser, K., Glosekotter, P.and Dienstuhl, J. (2004) *Nanoelectronics and Nanosystems: From Transistors to Molecular and Quantum Devices*, Springer, Berlin.

8 Waser, R. (ed.) (2005) *Nanoelectronics and Information Technology: Advanced Electronic Materials and Novel Devices*, 2nd edn, Wiley-VCH Verlag GmbH, Berlin.

9 Tarascon, J. and Armand, M. (2001) *Nature*, **414**, 359.

10 The end of the petrolhead: tomorrow's cars may just plug in. The Economist, June 28, 2008.

11 Ravet, N., Abouimrane, A. and Armand, M. (2003) *Nature Materials*, **2**, 702; See also Gorman, J. (2002) New material charges up lithium-ion battery work – bigger, cheaper, safer batteries. Science News, September 28, 2002.

12 Schindall, J. (November 2007) The charge of the ultra-capacitor. *IEEE Spectrum*, pp. 42–46.

13 Solymar, L. and Walsh, D. (2004) *Electrical Properties of Materials*, 7th edn, Oxford University Press, Oxford, p. 87.

14 Fowler, R. and Nordheim, L. (1928) *Proceedings of the Royal Society A*, **119**, 173.

15 Cleland, Andrew N. (2003) *Foundations of Nanomechanics: From Solid State Theory to Device Applications*, Springer, Berlin.

16 Thuillier, G., Herse, M., Labs, D., Foujols, T., Peetermans, W., Gillotay, D., Simon, P. and Mandel, H. (2003) *Solar Physics*, **214**, 1.