SEMICONDUCTORS, SILICON BASED

1. Introduction

With the invention of the transistor in 1947 semiconductors changed the basis of electronics from vacuum tubes to semiconductor devices. Some of the history of the development of transistor technology has been described (1–9). The invention of the integrated circuit in 1958 revolutionized electronic systems. The reason for this can be seen by considering the Intel 8086 (1978), the 16-bit microprocessor used in the first IBM PC. This chip used NMOS technology (MOS = metal/oxide/semiconductor) with 10-μm feature lengths and contained 20,000 transistors in a 25-mm^2 chip. Compare this with the Whirlwind computer (1950), the first large-scale, real-time control system, which contained 15,000 vacuum tubes and occupied a 2500-ft^2 room. Thus, a single semiconductor chip contained the complexity of a large system. As important, the failure rate of the Intel 8080 with 4500 transistors, 220 in 10^9 h, was comparable to the failure rates of reliable, single transistors. Remarkably, this revolution in circuit integration has continued.

The semiconductor industry, which is primarily the manufacturers of silicon-based integrated circuit (IC) chips, has managed to put exponentially increasing numbers of transistors on silicon chips since 1960. This has been driven by Moore's Law (10), which Gordon Moore defined in 1965 by noting that the number of components on ICs was doubling every 12 months. Since 1970 the number of bits on a DRAM (DRAM = dynamic random access memory) chip has doubled every 18 months, while the number of transistors on a microprocessor chip has doubled every 24 months (11). Moore's law is a self-fulfilling prophecy because it has provided goals for the (silicon-based) semiconductor industry. These goals are now codified in projections by the International Technology Roadmap for Semiconductors (ITRS) (12).

The exponential growth of circuit integration on chips characterized by Moore's law has led to an extraordinary growth in semiconductor device revenue. This has increased by a 13–16% compound annual growth rate (CAGR) from 1960 (13). Allowing for inflation, this corresponds to about a 10% CAGR in constant dollars. Semiconductor device revenue plotted on a semilogarithmic plot as a function of time leads to a curve that is nearly a straight line over the 40-year period from 1960 to 2000. However, this masks the cyclic, precipitous changes in revenue that have characterized the semiconductor industry. In particular, the drop in revenue from $200 billion in 2000 to $140 billion in 2001 was the largest drop in the history of this industry, with corresponding implications for the companies involved.

The semiconductor industry was originally driven by advances in DRAM technology, then by advances in microprocessor technology, as can be seen from the corresponding focus of Intel in the mid-1980s. More recently, there is increasing emphasis on chips for communications and consumer electronics. The present market for the industry is primarily silicon-based ICs containing MOSFETs (MOSFET = metal-oxide-semiconductor field effect transistor). These chips are

the foundation for the increasing convergence of the computer, communications, and consumer electronic industries.

The doubling of transistors on chips predicted by Moore's law has been primarily the result of shrinking transistor sizes. Multiplying the length and width of transistors by a factor of $1/\sqrt{2}$ or 0.7 will double the number of transistors on the same size chip. The consequences of this trend can be seen in Figure 1, which compares the rates of technological change in the microelectronics and industrial revolutions. Technological change is measured in DRAM memory bits or transistors for the microelectronics revolution (11) and in the horsepower of steam engines for the industrial revolution (14). The result is exponential growth, as shown by straight lines on a semilogarithmic plot. The degree of technological change in the microelectronics revolution is orders of magnitude greater, even though it has taken less than half the time taken by the industrial revolution. Considering the social changes brought by the industrial revolution, one may expect even greater changes from the microelectronics revolution.

As the silicon semiconductor field has matured, its focus has shifted from the development of new devices to the manufacture (fabrication) of the same devices, mostly MOSFETs, made from mostly the same materials, but with ever-smaller dimensions. Table 1, taken from ITRS 2004 (12), shows a rate of progress based on the continuation of Moore's law for the next several years. Each technology generation is identified by a dimension that is the minimum half-pitch of the first metal (M1) interconnect level. Historically, this has been the half-pitch for DRAMs. Note that we are in a deep submicron regime, characterized by quoting half-pitch dimensions in nanometers rather than microns.

Nanometer dimensions give a new meaning to questions of doping and contamination. When semiconductor devices were first developed, they were characterized by extraordinary levels of chemical purity, with typical impurity concentrations on the order of $10^{17}\,\mathrm{cm}^{-3}$ or 2 parts per million atoms (ppma) because the number density of silicon is $5 \times 10^{22}\,\mathrm{cm}^{-3}$. Consider a regular prism with dimensions $(100\,\mathrm{nm})^2 \times 10\,\mathrm{nm}$, not unreasonable for deep submicron device features. It would contain 5000 silicon atoms. An impurity concentration of 2 ppma would mean there was only one impurity atom in 100 such prisms.

The starting material for bulk devices has been a wafer cut from a single crystal of silicon, 200 mm (8 in.) in diameter and $\sim 650\,\mu\mathrm{m}$ thick, with a global flatness variation across the wafer of $<3\,\mu\mathrm{m}$ (15). Chemical perfection of these wafers is essential. The largest impurity in these wafers, oxygen, is kept $\sim$20–35 ppma. In this range, oxygen improves the mechanical strength of the wafer and promotes intrinsic gettering (16), in which oxygen is used to draw fast diffusing metal contaminants into the bulk of the wafer and away from the surface on which active devices are formed. Specific metallic contaminants, particularly heavy metals, eg, Fe, Ni, Cu, and Au, need to be kept <0.01 ppba because their presence can degrade device performance by reducing minority carrier lifetimes (15). Perfection is especially necessary on the silicon surface. A {100} of plane silicon contains 6.8×10^{14} atoms/cm^2. Surface defect densities must be <1 part in 10^5–10^6 defects/cm^2 for satisfactory MOSFET operation. Good silicon devices required the discovery (8) that the thermal oxidation of silicon could produce an excellent Si–SiO_2 interface.

The high purity required of silicon and the small size of semiconductor devices place stringent limits on the chemicals used in processing silicon. By far the most important chemical is water, which is used extensively to dilute etchants and clean wafers. Pure water [24] is required to have <0.025 ppba of Na^+ and Cl^-, 0.002 ppba of heavy metals, eg, Fe, and <10 ppb of total oxidizable (organic) carbon (TOC), eg, bacteria. In addition, pure water should have <1 particulate/L, $>0.5\,\mu m$ in size. Water this pure is, in fact, an active chemical that absorbs, dissolves, or reacts with all kinds of materials to create contamination. Similar requirements are placed on other process chemicals.

Defects are equally important. Since experience has shown that the density of particles leading to defects is proportional to the reciprocal of the cube of the particle size (14), smaller device features generally mean a much greater susceptibility of devices to defects. For this defect density, one can show that the yield of a circuit (yield = the fraction of good circuits or chips) remains constant when device features are shrunk to the next technology node. However, this doubles the number of transistors on a chip of the same size. For the high yields (≥ 0.85) acceptable for manufacturing, chip yield will be squared (~ 0.72) when the number of transistors doubles. All of these factors mean that maintaining acceptable yields is a continuing challenge for the semiconductor industry.

At the same time, the semiconductor industry is transitioning to 300 mm (12 in.) wafers that can contain $\sim 2.5\times$ the number of chips as 200-mm wafers. Estimates indicate that the ratio of finished costs for 300–200-mm wafers is only ~ 1.4 (18). This means that each 300-mm fabrication facility (fab) can produce the output of at least 2.5 200-mm fabs at lower cost and can replace them.

2. Semiconductor Materials Theory

Silicon is a Group 14 (IV) element of the Periodic Table. This column includes C, Si, Ge, Sn, and Pb and displays a remarkable transition from insulating to metallic behavior with increasing atomic weight. Carbon, in the form of diamond, is a transparent insulator, whereas tin and lead are metals; in fact, they are superconductors. Silicon and germanium are semiconductors, ie, they look metallic, so that a polished silicon wafer is a reasonably gray-toned mirror, but they conduct poorly. Traditionally, semiconductors have been defined as materials whose resistance rises with decreasing temperature, unlike metals whose resistance falls.

Diamond, silicon, and germanium all crystallize in the diamond crystal structure with atoms tetrahedrally bonded to each other. This crystal structure can be described as two interpenetrating face-centered cubic (fcc) lattices (19). As one moves down the column of Group 14 (IV), the bonds lengthen and weaken, lowering the melting and boiling points on these materials. Table 2 indicates some of the key properties of Group 14 (IV) semiconductors.

The electrical behavior of many materials can be explained elegantly by band theory (21). When atoms are packed together their energy levels split into bands of closely spaced energy levels. Some of these bands overlap and some are separated by energy gaps. Whether a material is an insulator or conductor depends on whether a band is completely or partially filled. In the case of

silicon, the atom contains two electrons that fill the 3*s* level and two electrons in the six states of the 3*p* level. As the atoms are brought together in the diamond configuration, these eight states per atom hybridize and split, forming two bands of energy levels separated by an energy gap. At 0 K, the lower valence band with four bonding states per atom is filled and the upper conduction band, which also has four antibonding states per atom, is empty. Consequently, diamond, silicon, and germanium are essentially insulators with an empty conduction band separated from a filled valence band by an energy gap, E_g, of the size indicated in Table 2. At higher temperatures, some electrons are excited to the conduction band, leaving behind empty states in the valence band. Applying pressure decreases the lattice spacing and increases E_g; thermal expansion decreases E_g. These effects can be exploited in pressure- and temperature-sensitive transducers (22).

Near a conduction band minimum the energy of an electron depends on its momentum in the crystal. Thus, carriers behave like free electrons whose effective mass differs from the free electron mass. Their energy is given by equation 1, where E_c is the energy of the conduction band minimum, p is the effective momentum of the electrons in the crystal, and m^* is the effective mass.

$$E - E_c = p^2/2m^* \tag{1}$$

Because the crystal momentum p is related to the wave vector $k = 2\pi/\lambda$ (by de Broglie's relation, $p = hk/2\pi$, where h is Planck's constant and $h/2\pi = 0.658 \times 10^{-15}$ eVs $= 0.685$ feVs), the energy bands are often drawn as an E versus k diagrams. The nearly filled states at the top of the valence band behave in conduction as positive charge carriers (holes) with a different effective mass. The density of states is the number of distinct states or k-values available for occupation within a given energy interval. There are two states for each unique k-value, corresponding to the two possibilities for electron spin. A low effective mass means that the E versus k diagram has a steep slope and a correspondingly low density of states. Figure 2**a** shows an E versus k diagram for an idealized semiconductor in which the electrons and holes have nearly equal mass. The solid lines in Figure 2**a** correspond to available states for electrons with quantized k-values. If the available states are counted as a function of energy, the density (per unit volume) of states (DOS) for electrons is given, as shown in Figure 2**b**.

This simple picture is complicated because the dependence of energy on momentum may be anisotropic and the energy bands may have several minima or maxima. This is the case for silicon whose minimum conduction band energy surfaces are displaced spheroids along the ⟨100⟩ axes (23). This can be described by introducing longitudinal and traverse electron effective masses, $m_l = 0.92$ m and $m_t = 0.19$ m, respectively. An appropriate average gives the density of states effective mass for electrons at 4 K, $m^* = 1.06$ m (23).

Energy bands help to explain the optical properties of materials. Photons must have energies greater than the energy gap to excite electrons from the valence to the conduction band. Visible photon energies are in the red-to-blue energy range of 1.7–3 eV. Thus, insulators, eg, diamond are frequently

transparent ($E_g > 3\,eV$) and metals are opaque. Silicon with an energy gap of 1.12 eV is opaque to visible light, but transparent in the near-infrared (ir) for wavelengths $\lambda(\mu m) > 1.24/E_g(eV) = 1.1\,\mu m$. This transparency is important for applications, eg, solar cells, that depend on the efficient absorption of light. Normally transparent crystals, eg, sapphire can be colored by the presence of impurities that introduce energy levels into the energy gap and alter the absorption characteristics.

Diamond, silicon, and germanium have an indirect band gap, ie, the conduction band minimum has a different momentum than the valence band maximum. Compound semiconductors, eg, GaAs, generally have a direct band gap in which these momenta match. In direct band gap semiconductors an electron dropping from the conduction band minimum to fill an empty state at the valence band maximum can conserve energy and momentum by emitting a photon, which has very little momentum. The same transition in an indirect band gap semiconductor also requires emitting a quantum of lattice vibration (phonon) to conserve momentum. Because this three-body process is much less likely, optical emission is much less efficient than for direct semiconductors.

Remarkably, although band structure is a quantum mechanical property, once electrons and holes are introduced, their behavior generally can be described classically even for deep submicrometer geometries. Some allowance for band structure may have to be made by choosing different values of effective mass for different applications. For example, different effective masses are used in the density of states and conductivity.

3. Semiconductor Statistics

3.1. Intrinsic Semiconductors. For semiconductors in thermal equilibrium, $N(E)$, the average number of electrons occupying a state with energy E is governed by the Fermi-Dirac distribution. Because, by the Pauli exclusion principle, at most one electron (fermion) can occupy a state, this average number is also the probability, $P(E)$, that this state is occupied (see Fig. 2**c**). In equation 2, K is the Boltzmann constant and T is the absolute temperature in kelvin.

$$\langle N(E)\rangle = P(E) = \frac{1}{1 + e^{(E-E_F)/KT}} \tag{2}$$

E_F, the Fermi energy, normalizes the distribution $P(E)$ at all temperatures. At low temperatures the Fermi-Dirac distribution is rectangular, with $P(E) \approx 0$ for $E > E_F$. However, at normal operating temperatures for a semiconductor device, $|E - E_F| > KT$, so the exponential in the denominator of equation 2 dominates and $P(E)$ is well approximated by a classical, exponential, Maxwell-Boltzmann distribution.

In an undoped, intrinsic semiconductor the equilibrium concentrations of electrons, n, and holes, p, are described by a lever rule derived from the law of mass action (eq. 3):

$$np = n_i^2 = N_c N_v e^{-E_g/KT} \quad (3)$$

In an intrinsic semiconductor, charge conservation gives $n = p = n_i$, where n_i is the intrinsic carrier concentration as shown in Table 2. The parameters N_c and N_v are the effective densities of states per unit volume for the conduction and valence bands. In terms of these densities of states, n and p are given in equations 4 and 5, where E_c and E_v are the energies of the

$$n = N_C e^{-(E_C - E_F)/KT} \quad (4)$$

$$p = N_V e^{-(E_F - E_V)/KT} \quad (5)$$

conduction and valence band edges, $E_g = E_c - E_v$, and E_F is the Fermi level. Electrons and holes obey Fermi-Dirac statistics. The above formulas are the result of integrating, over all energies, the product of the density of states and the probability for finding a carrier at a particular energy. These formulas assume that the Fermi level is located in the energy gap, E_g, away from the conduction and valence bands so that the statistics obey the classical Maxwell-Boltzmann limit. The parameters N_c and N_v are functions of temperature and are proportional to $T^{3/2}$ and $N_{c,v} = 2M_{c,v}(2\pi m_{c,v} * KT/h^2)^{3/2}$, where $M_{c,v}$ = the number of equivalent minima or maxima in the conduction and valence bands, respectively, and $m_{c,v}^*$ = the density of states effective masses of electrons and holes.

3.2. Extrinsic Semiconductors. The most common impurities or dopants introduced into silicon crystal are Periodic Table Group 15 (V) and Group 13 (III) elements. These dopants are substitutional, ie, they replace silicon atoms in the crystal lattice. Group 15 (V) elements contain five valence electrons, four of which bond covalently with the surrounding silicon atoms. These impurities are called donors because their fifth electron is weakly bound and can be donated to the conduction band by thermal excitation. This weak bonding is approximately described by a hydrogen atom model in which the energy levels of the hydrogen atom are adjusted for the effective mass of the electron and the relative dielectric constant of the semiconductor, $\varepsilon_r = 11.9$ for silicon. In silicon, As, P and Sb have shallow donor energy levels, E_d, that are 0.045, 0.044, and 0.038 eV below E_c, respectively. A similar argument can be made for Group 13 (III) impurities. In silicon, B, Al, Ga, and In have shallow acceptor levels, E_a, which are 0.045, 0.057, 0.065, and 0.16 eV above E_v, respectively. In practice, the most common donors are phosphorous and arsenic; the most common acceptor is boron. Most commonly, in the semiconductor industry these impurities are introduced into silicon by ion implantation followed by a thermal treatment to activate the impurity.

The variations in E_d and E_a and the much larger value for In show the limitations of a simple hydrogen atom model. Other elements, particularly transition metals, tend to introduce several deep levels in the energy gap. For example, gold introduces a donor level 0.54 eV below E_c and an acceptor level 0.35 eV above E_v in Si. Because such impurities are effective aids to the recombination of electrons and holes, they limit carrier lifetime.

The carrier concentrations in doped or extrinsic semiconductors to which donor or acceptor atoms have been added can be determined by considering

the chemical kinetics or mass action of reactions between electrons and donor ions or between holes and acceptor ions. The condition for electrical neutrality is given by equation 6.

$$n + N_{a-} = N_{d+} + p \tag{6}$$

When the predominant dopants are donors, the semiconductor is n-type. In n-type semiconductors, the majority carriers are electrons and the minority carriers are holes so that $N_{d+} > N_{a-}$ and $n \sim N_{d+}$. By the lever rule, $p \sim n_1^2/N_{d+} < n$, so the hole concentration is greatly reduced by the introduction of donors. Similarly, for p-type semiconductors, $N_{a-} > N_{d+}$ and $p \sim N_{a-}$.

For lightly doped n-type semiconductors at normal operating temperatures there is complete donor dissociation (donor saturation). The parameter $N_{d+} \sim N_d$, corresponding to E_F well below E_d. In the presence of acceptors, $n = N_{d+} - N_{a-} \sim N_d - N_a$. This frequently occurs because semiconductor devices are customarily made by the diffusion or implantation of excess donors into a p-type substrate or vice versa.

As the temperature increases, the intrinsic carrier concentration rises exponentially so that at some point $n \sim p \sim n_i$ rather than $n \sim N_{d+}$. Likewise, as the temperature falls, $E_a - E_v$ becomes large compared to KT and carrier freeze-out occurs below ~ 100 K, so that $n \sim N_{d+} < N_d$. All of this assumes that $n \leq N_c$ or $p \leq N_v$, corresponding to E_F in the energy gap. When enough donor atoms are added so that $n \geq 10\ N_c$, the Fermi level moves into the conduction band, $E_F > E_c$, and the semiconductor is said to be degenerately doped or n^+. Now Fermi-Dirac statistics apply and the degenerately doped semiconductor behaves like a metal. In particular, carriers do not freeze out at cryogenic temperatures. Similar comments apply to degenerately doped p^+ semiconductors.

The equilibrium lever relation $np = {n_i}^2$ can be regarded from a chemical kinetics perspective as the result of a balance between the generation and recombination of electrons and holes (15). The recombination rate in extrinsic semiconductors is limited by the lifetime of minority carriers which, according to the equilibrium lever relation, have much lower concentrations than majority carriers. Recombination and generation rates generally are greatly accelerated by the presence of deep levels in the band gap because the required energy jumps are smaller. Surface and interface states can also act as recombination centers, drastically reducing minority carrier lifetimes.

3.3. Noise. So far, our discussion has been limited to equilibrium statistics. Actually, there are fluctuations about the equilibrium values, $\Delta N = N - N$. For electrons, the mean-square fluctuation is given by $\langle (\Delta N)^2 \rangle = \langle N \rangle (1 - \langle N \rangle)$, where $N(E)$ is the Fermi-Dirac distribution. This mean-square fluctuation has a maximum of one-fourth when $E = E_F$. These statistical fluctuations act as electrical noise and limit maximum signal levels.

Semiconductor devices are affected by three kinds of noise. Thermal or Johnson noise is a consequence of the equilibrium between a resistance and its surrounding radiation field. It results in a mean-square noise voltage that is proportional to resistance and temperature. Shot noise, which is the principal noise component in most semiconductor devices, is caused by the random passage of

individual electrons through a semiconductor junction. Thermal and shot noise are both called white noise, since their noise power is frequency independent at low and intermediate frequencies. This is unlike flicker or $1/f$ noise that is most troublesome at lower frequencies, because its noise power is approximately proportional to $1/f$. In MOSFETs, there is a strong correlation between $1/f$ noise and the charging and discharging of surface states or traps. Nevertheless, the universal nature of $1/f$ noise in various materials and at phase transitions is not well understood.

4. Semiconductor Transport

Charge carriers in a semiconductor are always in random thermal motion with an average thermal speed, v_{th}, given by the equipartion relation of classical thermodynamics as $m^* v_{\mathrm{th}}^2/2 = 3KT/2$. As a result of this random thermal motion, carriers diffuse from regions of higher concentration. Applying an electric field superposes a drift of carriers on this random thermal motion. Carriers are accelerated by the electric field, but lose momentum to collisions with impurities or phonons, ie, quantized lattice vibrations. This results in a drift speed, v_{d}, which is proportional to the electric field; $v_{\mathrm{d}} = \mu_{\mathrm{e}} E$, where E is the electric field in volts per cm (V/cm) and μ_{e} is the electron's mobility in units of $\mathrm{cm}^2/\mathrm{Vs}$.

The electron current density J_{e} has units of amperes per square centimeter ($\mathrm{A/cm}^2$) and in a semiconductor results from drift and diffusion. In the absence of concentration gradients, equation 7 reduces to Ohm's law, $J_{\mathrm{e}} = nq\mu_{\mathrm{e}} E = \sigma E$, where σ is the conductivity, q is the magnitude of the charge of an electron, and D_{e} is the electron's diffusion constant. The Einstein relation relates μ_{e}, and D_{e} by $D_{\mathrm{e}} = KT\mu_{\mathrm{e}}/q$. A similar equation can be written for hole currents, J_{h}, if $+qD_{\mathrm{e}}$ is replaced by $-qD_{\mathrm{h}}$. High mobilities generally lead to high currents. In practice, $\mu_{\mathrm{h}} < \mu_{\mathrm{e}}$, for most semiconductors.

$$j_{\mathrm{e}} = nq\mu_{\mathrm{e}} E + qD_{\mathrm{e}} \frac{\partial n(x)}{\partial x} = \mu_{\mathrm{e}} n \frac{\partial E_{\mathrm{F}}}{\partial x} \tag{7}$$

Ohm's law assumes that the drift speed of electrons in an electric field, $v_{\mathrm{d}} = \mu_{\mathrm{e}} E$, is small compared to their average speed, v_{th}, in a Maxwell-Boltzmann distribution. At high electric fields, $E \geq 10\,\mathrm{kV/cm}$, v_{d} no longer increases with electric field and approaches a limiting saturation speed, v_{s}, determined primarily by optical phonon emission. Figure 3 shows the variation of drift speed with electric field for electrons and holes in various semiconductors.

At still higher fields, carriers can acquire enough energy from motion in an electric field to create electron–hole pairs by impact ionization. For silicon, the electron ionization rate, which is the number of pairs generated per centimeter of electron travel, depends exponentially on electric field. It is $\sim 2 \times 10^3\,\mathrm{cm}^{-1}$ for a 50-kV/cm field at 300 K. The electric field causes electrons and holes so created to travel in opposite directions. They may create other electron–hole pairs causing positive feedback, which leads to avalanche breakdown at sufficiently high fields.

5. MOS Capacitance

So far, the fundamental factors that affect the current a device can carry have been considered. In practice, the speed of digital (and analog) devices is determined by the time it takes to charge or discharge a capacitor, C, where $\Delta Q = C\ \Delta V = I\,\Delta t$. Thus, from this general relationship, low values of Δt require some combination of low C, low ΔV, or large I. Different device technologies and circuit techniques achieve this in different ways, but it can be seen that an understanding of device and circuit capacitance is essential to an assessment of device performance.

This is particularly true for the MOSFET, where conduction is controlled by a metal/oxide/semiconductor (MOS) capacitor. When a voltage is applied to the metal plate of a MOS capacitor most of the voltage drop is across the oxide, but some drop is across the semiconductor. If the semiconductor is p-type, a positive voltage applied to the metal plate induces negative charges in the semiconductor. Initially, this causes holes to move away from the surface of the semiconductor, leaving behind a carrier depletion region containing immobile, ionized acceptors, n_{a-}. Figure 4 shows the charge distributions and band bending in a MOS capacitor as applied voltage is changed. As the voltage is increased, mobile electrons are drawn to the surface. Strong inversion is said to occur when the density of electrons at the surface equals the original density of holes. Strong inversion is generally used to define the threshold voltage for conduction in MOSFETs.

The magnitude of the capacitance depends on the voltage approaching the oxide capacitance in accumulation and inversion, and dropping to ~ 0.4 of the oxide capacitance in depletion, because the negative charge due to ionized acceptors is farther from the metal plate. The rise of capacitance in inversion assumes that electrons can follow the changes in the electric field. This only happens at low frequencies of 5–100 Hz for the $M/SiO_2/Si$ capacitor. At high frequencies, depletion continues and the capacitance continues to drop with increasing voltage reaching a minimum value corresponding to the maximum width of the depletion layer.

6. The Si–SiO_2 Interface

The simple picture of the MOS capacitor presented in the last section is complicated by two factors, work function differences between the metal and the semiconductor and excess charge in the oxide. The difference in work functions, the energies required to remove an electron from a metal or semiconductor, is $q\phi_{ms} = -25$ meV for an aluminum metal plate over a 50-nm thermally grown oxide on n-type silicon with $n = 10^{16}\,cm^{-3}$. This work function difference leads to a misalignment of energy bands in the metal and semiconductor that has to be compensated by a variation of the energy band with distance. When there is no misalignment the flat-band condition results.

Excess charge in the oxide has the effect of charge on the metal plate, but has a larger effect because it is closer to the silicon surface. There are several

components of excess charge. Interface trapped charge at the $Si-SiO_2$ interface actively exchanges with carriers at the semiconductor surface. Fixed interface charges at or near the interface are immobile in an applied electric field. Oxide trapped charges that are also immobile can be created by irradiation or hot-electron injection. They have an important influence on the wear-out and reliability of MOS devices. Mobile ionic charges, eg, Na^+ ions, are notorious because they shift threshold voltages as they move under bias-temperature aging.

Interface states played a key role in the development of transistors. The initial experiments at Bell Laboratories were on metal/insulator/semiconductor (MIS) structures in which the intent was to modulate the conductance of a germanium layer by applying a voltage to the metal plate. However, only $\sim 10\%$ of the induced charges were effective in charging the conductance (2). It was proposed (1) that the ineffective induced charges were trapped in surface states. Subsequent experiments on surface states led to the discovery of the point contact transistor reported in 1948 (3).

At a semiconductor–vacuum interface, dangling bonds extend from the surface atoms into the vacuum. Surface reconstruction minimizes the electronic energy at a semiconductor surface by forming bonds, but the presence of a large number of surface states can completely overshadow any effects of doping near the surface. Thus, it is important to passivate the surface by removing most of the surface states. Because thermal oxidation could passivate the silicon surface, the first silicon MOSFET could be created in 1960 (8). Electrons are strongly bound in the interfacial Si—O bonds and do not participate in conduction. Thermal oxidation can reduce the number of surface states by four orders of magnitude.

The excellence of a properly formed SiO_2-Si interface and the difficulty of passivating other semiconductors' surfaces has been one of the most important factors in the development of the worldwide market of silicon-based semiconductors. MOSFETs are typically produced on {100} silicon surfaces. Fewer surface states appear at this $Si-SiO_2$ interface, which has the fewest broken bonds. A widely used model for the thermal oxidation of silicon has been developed (24).

7. Device Physics

The ideal rectifier or diode is a two-terminal device that allows current flow in only one direction. The transistor is a three-terminal device in which current flow through two terminals is controlled by the third.

7.1. *p-n* Junctions. In addition to its use as a rectifier, the *p-n* junction (19) is the fundamental building block for bipolar, junction FET (JFET), and MOSFET transistors. A thorough understanding of *p-n* junctions explains much of transistor behavior. The theory (4) of the *p-n* junction and its role in bipolar transistors was presented within a year of the discovery of the point-contact transistor.

At an abrupt interface between *n*- and *p*-type semiconductors, there is an enormous difference between electron and hole concentrations on the two sides of the interface. This corresponds to a difference in Fermi energies on the two

sides. Thermal equilibrium with no applied voltage and no current flow requires a constant Fermi energy throughout the material. This equilibrium is achieved if electrons diffuse into the p-type region and holes diffuse into the n-type region, where they become minority carriers. This leaves behind a depletion region at the interface with ionized acceptors on the p-side and ionized donors on the n-side. This charge separation leads to a built-in or diffusion voltage difference, V_{bi}, across the interface. This built-in voltage corresponds to a difference in band energies as shown in Figure 5**d**.

$$qV_{bi} = KT \ln (n_{no} p_{po}/n_i^2) \sim KT \ln (n_d n_a/n_i^2) \tag{8}$$

Because $n_{no}\ p_n = {n_i}^2 = n_{po}\ p_{po}$, where p_{no} and n_{po} are the equilibrium minority concentrations in the n- and p region, respectively, the equation for V_{bi} allows the determination of the concentrations of minority carriers on either side of the depletion region. Thus, equation 8 can be rewritten as

$$n_{po} = n_{no} \exp(-qV_{bi}/KT) \tag{9}$$

where n_{po} is the concentration of minority electrons in the p region.

In general, in a planar process, p-n junctions are formed just below the surface of the silicon wafer by implantation of donor ions into a p-type region or acceptor ions into an n-type region. Thus, the general concern is with n^+-p or p^+-n junctions. As the initial wafer concentration of acceptors or donors in silicon rises from 10^{14} to $10^{18}\,\mathrm{cm}^{-3}$, V_{bi} increases from ~ 0.81 to $1.04\,\mathrm{V}$ for a p^+-n junction and is $\sim 10\,\mathrm{mV}$ higher for an n^+-p junction.

When a positive voltage from p to n is applied across the p-n junction, the voltage drop is mostly across this depletion region and reduces the built-in voltage. This forward voltage drop decreases the width of the depletion region as it reduces the charge imbalance required to sustain the built-in voltage. In simple models of junction behavior, voltage drops are neglected across the neutral regions, which are the n- and p regions outside the depletion region. Actually, there are resistive voltage drops across these regions that are proportional to the current flow. This provides a parasitic resistance that needs to be included in a more accurate device model.

When electrons are injected as minority carriers into a p-type semiconductor they may diffuse, drift, or disappear. That is, their electrical behavior is determined by diffusion in concentration gradients, drift in electric fields (potential gradients), or disappearance thorough recombination with majority carrier holes. Thus, the transport behavior of minority carriers can be described by a continuity equation. To derive the p-n junction equation, steady state is assumed, so that $\partial p_n/\partial t = 0$, and a neutral region outside the depletion region is assumed, so that the electric field is zero. Under these circumstances, the continuity equation reduces to a diffusion equation (eq. 10),

$$D_e \frac{\partial^2 n_p}{\partial x^2} - \frac{n_p - n_{po}}{\tau_p} = 0 \tag{10}$$

where τ_p is the lifetime of minority carriers in the p-type material. The electron current density is given by equation 11.

$$j_e = qD_e \frac{\partial n_p}{\partial x} = -\frac{qD_e n_{po}(0)(e^{qV/KT} - 1)}{L_p} = 0 \tag{11}$$

$L_p = (D_e \tau_p)^{-1}$ is the minority carrier diffusion length for electrons in the p region. The parameter $n_{po}(0)$ is the minority carrier concentration at the boundary between the depletion layer and the neutral region. The sign of this equation indicates that electron injection into the p region results in a positive current flow from p to n as shown in Figure 6.

Hole injection into the n region similarly results in a positive current flow from p to n^+, leading to the Shockley diode equation (eq. 12),

$$I = I_s(e^{qV/KT} - 1) \tag{12}$$

with a reverse current, I_s, where A is the area of the junction.

$$I_s = qA\left(\frac{D_e n_{po}}{L_p} + \frac{D_h p_n}{L_n}\right) \tag{13}$$

For positive voltages the current, I, can become exponentially large and is limited by junction heating and burnout. A forward-biased silicon p-n junction has the voltage drop shown in equation 14. For negative voltages, $I \sim -I_s$.

$$V = (q/KT)\ln(I/I_s) \approx 0.7\,\mathrm{V} \tag{14}$$

For an n^+-p junction, the lever rule gives $p_{no} \ll n_{po}$, so that $I_s \sim qA(D_e n_{po}/L_p)$. For reasonable values ($\mu \sim 1000\,\mathrm{cm^2/(V{\cdot}s)}$, $n_a = 10^{16}\,\mathrm{cm^{-3}}$, $\tau_p = 1\,\mu\mathrm{s}$, $L_p \sim 50\,\mu\mathrm{m}$), the reverse saturation current would be $17 \times 10^{-12} = 17\,\mathrm{pA}$ for a square centimeter of junction area. Typical reverse saturation currents are about 1000 times greater as a result of generation–recombination currents in the depletion region (7). As the reverse voltage bias increases, the field increases in the depletion region until avalanche breakdown occurs, resulting in the characteristic shown in Figure 6.

The depletion region width and minority carrier distribution near the junction must be altered in order to change the voltage across the p-n diode. This corresponds to two capacitive components, the junction capacitance that dominates in reverse bias and the diffusion or charge-storage capacitance that dominates in forward bias. The junction capacitance corresponds to charge separation caused by the concentration of ionized donors and acceptors in the depletion region. Applied voltage changes, dV, across the depletion region change the width of the depletion region. This changes the charge it contains, dQ, which then defines the junction capacitance, $C = |\,dQ/dV\,|$. Because the depletion layer width depends nonlinearly on the applied voltage, this capacitance also is voltage dependent. In addition to the junction capacitance, the injection and removal of minority carriers from the neutral regions has electrical effects similar to capacitance and defines the diffusion capacitance. This capacitance depends

on the lifetime of minority carriers. Doping the *p-n* junction with gold has been used to reduce the minority carrier lifetime, thereby reducing the diffusion capacitance and decreasing the switching time associated with the *p-n* junction.

Avalanche or tunnel breakdown occurs in the *p-n* junction because of the large electric fields in the depletion region for sufficiently large reverse bias. Increasing the doping on either side of the junction decreases the depletion width and lowers the avalanche breakdown voltage, although the actual breakdown field increases. As the electric field increases the slope of band energy with distance increases. When the distance between conduction and valence bands at constant band energy is <10 nm, electrons can tunnel from the valence band to the conduction band. As shown in Figure 7, the energy barrier through which the electrons tunnel is essentially triangular, corresponding to Fowler-Nordheim tunneling. For doping levels $>10^{18}\,\mathrm{cm}^{-3}$ in silicon the critical field for avalanche breakdown exceeds the threshold for tunneling and tunnel breakdown occurs. Junction avalanche breakdown can be useful since it can be nondestructive. The junction breakdown voltage can be used as a reference voltage as in Zener diodes. Avalanche breakdown is also used for programming nonvolatile memories by hot-carrier injection. On the other hand, hot-carrier injection is a wear-out mechanism for MOSFETs.

7.2. Schottky Diodes and Ohmic Contacts. A metal–semiconductor junction may be rectifying or ohmic. Such a rectifying junction (Schottky diode) has a nonlinear, asymmetric dependence of current on voltage like the *p-n* junction, which could be called a Shockley diode (19). An ohmic contact has a linear, symmetric dependence of current on voltage. Figure 8 shows band diagrams for Schottky barrier junctions. In equilibrium, metal and semiconductor Fermi levels must align, resulting in a Schottky barrier height $q\phi_{\mathrm{BO}} = q\phi_{\mathrm{M}} - q\chi$ at the interfaces. The parameters $q\phi_{\mathrm{M}}$ and $q\chi$ are the metal work function and electron affinity, the energies required to remove an electron from the Fermi energy in the metal and the conduction band edge in the semiconductor, respectively.

Both ohmic and rectifying behaviors are possible, depending on the sign of $q\phi_{\mathrm{BO}}$. Unlike the *p-n* junction, the current in a rectifying Schottky barrier largely is controlled by thermal emission of majority carriers over the barrier. When there is no voltage difference between the metal and the semiconductor there is no current flow, ie, the electron flow from metal to semiconductor balances that from semiconductor to metal. A positive voltage applied to the semiconductor raises the barrier to electron flow from semiconductor to metal, but does not affect that from metal to semiconductor. This leads to the same form of current voltage characteristics as the *p-n* junction (eq. 12), but there are differences. The Schottky diode has a larger reverse saturation current and its switching speed is not limited by the minority carrier lifetime. Consequently, there is a large difference between forward and reverse currents (rectifying behavior) for a positive barrier (Fig. 8**b**), but very little difference (ohmic behavior) for a negative barrier (Fig. 8**a**) that does not constrain electron flow.

Because rectifying barriers are more common than ohmic barriers, dopant concentrations of $\sim 10^{19}\,\mathrm{cm}^{-3}$ are typically used to form ohmic contacts on silicon. The distance over which band bending occurs decreases as the semiconductor doping increases because more dopants allow more rapid depletion. Tunneling

through the depletion region is possible if it is $< \sim 10$ nm, leading to ohmic behavior with a rectifying barrier.

8. Bipolar Transistors

Figure 9 shows schematically, that the *n*-*p*-*n* bipolar junction transistor (BJT) may be regarded as two back-to-back *p*-*n* junctions separated by a thin base region (19,25,26). If external voltages are applied, so that the BE junction is forward biased and the BC junction is reverse biased, electrons injected into the base from the emitter can travel to the base-collector junction within their lifetime. If the time for minority carrier electrons to transit the base (base transit time) is short compared with their lifetime, α, the fraction of electrons that reaches the collector junction is nearly one. When they reach the collector they are accelerated across the depletion region, giving rise to a current, $I_c = \alpha I_E \sim I_E$. Kirchoff's current law requires that the current flowing through the emitter contact must be supplied by currents flowing through the base and collector contacts, $I_E = I_C + I_B$. Substituting this equation for I_E gives a formula for I_C in terms of I_B, $I_C = \alpha I_B/(l - \alpha) = \beta I_B$. If $\alpha = 0.99$, then $\beta = 100$, and there is a substantial current gain from the base to the collector.

This simple picture explains most BJT behavior. So long as the BE and BC junctions are forward and reverse biased, respectively, the bipolar transistor is in the active region with a current gain, β. The BE junction must be forward biased by 0.5–0.6 V for appreciable current to flow and the BE input characteristics are those of a *p*-*n* junction. Thus, emitter current depends exponentially on V_{BE} according to the diode equation (eq. 12). This diode law holds over a very wide range of currents from nanoamperes to milliamperes. Because β may vary by a factor of 3 over this range, the BJT may be best understood as a transconductance amplifier, where I_C is determined by V_{BE}, rather than a current amplifier, where I_C, is determined by I_B. This transconductance amplifier perspective is particularly helpful when considering the design of differential amplifiers.

Figure 10**a** shows the symbol commonly used for an *n*-*p*-*n* BJT in a grounded-emitter circuit. Figure 10**b** shows a large-signal equivalent circuit that captures the bipolar transistor's essential characteristics for normal operation. In normal operation, the BJT operates with direct current (dc) voltages set so that the BE junction is forward biased and the BC junction reverse biased. (This is known as the forward-active region.) The input impedance is low since it corresponds to the forward-biased BE junction. However, between collector and emitter, a much larger current, determined by V_{BE}, flows through the reverse-biased BC junction. Thus, the output characteristics shown in Figure 10**c** display a very high resistance for higher values of V_{CE}. This is represented in the equivalent circuit by a current source, I_F, controlled by V_{BE}, that follows (eq. 12), the diode equation. Note that the BJT is a minority-carrier device, because the minority carriers in the base, electrons for an *n*-*p*-*n* BJT and holes for a *p*-*n*-*p* BJT, determine its performance.

A high gain transistor requires α to be nearly equal to one. The parameter α is the product of the base transport factor (the fraction of the minority current that reaches the collector) and the emitter efficiency (the fraction of emitter

current due to minority carriers injected into the base rather than the emitter). In practical BJTs, the emitter efficiency limits α, so that β is proportional to $n_E/p_B W$, where W is the base thickness, and n_E and p_B are the emitter and base dopings. To maximize β, the Gummel number, $p_B W$, must be minimized.

At high frequencies the capacitances associated with the depletion regions of the BE and BC junctions (C_{dBE} and C_{dCB}) and a diffusion capacitance (C_{DE}), which is caused by minority carriers associated with the forward-biased BE junction (28, Section 6.4.4), will tend to short out the transistor at high frequencies and reduce its performance. The most often used figure of merit for small-signal applications is the cutoff frequency, f_T, the frequency at which the small signal current gain, $\partial I_C/\partial I_B$, drops to unity when $R_C = 0$ (28, Section 8.1).

$$1/(2\pi f_T) \sim \tau_F + (k_B T/qI_C)(C_{dBE} + C_{dCB}) \tag{15}$$

where τ_F, the forward transit time, is the ratio of absolute values of the total minority-carrier charge to I_C.

Generally, an n-p-n BJT is fabricated by successively implanting acceptor dopants for the base and donor dopants for the emitter into an n-type semiconductor. Thus, the dopant concentrations must increase from the collector region to the base and then the emitter region, so that the semiconductor regions can change from n-type in the collector to p-type in the base, and then back to n-type in the emitter. This means that the current in a BJT flows vertically, so that the current flow depends on the area of the BE junction. Note that the minimum base width (thickness) is determined by the depletion layer widths of the BE and BC junctions. Narrow base widths or low base dopings allow the collector and emitter regions to come in contact above some value of V_{BC}. This condition is called punchthrough and leads to large currents uncontrolled by the base. The BJTs must be isolated from one another to prevent unwanted interactions in integrated circuits; this is most commonly done by shallow trench isolation (STI) where trenches etched in the silicon are filled with SiO_2 to prevent conduction between circuit elements.

Performance of an NPN BJT can be increased significantly by incorporating Ge into the base region of a Si bipolar transistor and using a polysilicon emitter. The reduced band gap of the SiGe alloy provides a reduced barrier to electron injection into the base. This reduces hole injection from the base into the emitter, which increases gain by increasing emitter efficiency. A Ge concentration gradient creates an electric field in the base that aids electron transport and reduces τ_F. Consequently, such a base region increases collector current, current gain, and cutoff frequency. For example, SiGe technology can increase f_T from 48 to 70 GHz for NPN BJTs of comparable dimensions (28, Section 8.3.3). This can be particularly important for BiCMOS mixed-signal applications, which combine BJT and CMOS transistors in analog circuits with CMOS in digital circuits. The compatibility of SiGe technology with conventional CMOS processing makes it less expensive than other compound semiconductor approaches. Because only lateral PNP BJTs with significantly worse performance may be available in a SiGe process, improvements in circuit performance may be limited.

9. Field-Effect Transistors

9.1. Long-Channel Behavior. Unlike the bipolar junction transistor, field-effect transistors (FETs) are unipolar devices in which the current flow is determined only by majority carrier transport. The majority carriers flow through a channel at the semiconductor surface (19,26,27). In the MOSFET, the most common type of FET, a conducting channel between two opposing p-n junctions, is formed by the application of a field at the silicon surface by a MOS capacitor. When the field is sufficiently high, an inversion layer is formed at the surface of the silicon as shown in Figure 4**c**. In the junction FET (JFET), a conducting channel between two ohmic contacts is constricted by the depletion regions of two opposing p-n junctions on either side of the channel.

A traditional n-type MOSFET (NFET) is shown in cross section in Figure 11. It is formed by growing a thin thermal oxide on the surface of a p-type silicon wafer and depositing and patterning a polysilicon gate that will serve as the MOSFETs metal plate. This is a self-aligned structure in which the polysilicon gate serves to mask the gate oxide when the n^+ source and drain regions are formed by ion implantation. Since the gate polysilicon is also implanted, its conductivity is increased. For low voltages on the polysilicon gate, the n^+-p-n^+ regions from the drain to the source form back-to-back diodes. When a gate voltage, V_{GS}, greater than some threshold voltage, V_T, is applied, an inversion channel forms at the surface. An important difference from the isolated MOS capacitor is that the n^+ source and drain junctions are a ready supply of electrons for the channel. Thus, the channel can be formed at high speed even at cryogenic temperatures because the source and drain regions are degenerately doped.

The NFET has three regions of operation. The cutoff region occurs for $V_{GS} < V_T$. In this region, conduction between the drain and source will be low, whatever V_{DS} may be, because one of the back-to-back p-n junctions will always be reverse biased. This leakage current is small, but nonzero and allows charge to leak off capacitors that are isolated by cutoff MOSFETs. Because this is how bits are stored in dynamic random access memory (DRAM) cells, DRAMs must be regularly refreshed to retain memory.

When $V_{GS} > V_T$ the NFET conducts. The conduction current is determined by Q/t_t, where Q is the amount of charge in the inversion layer and t_t is the transit time for electrons to travel from source to drain. $Q = C'_o LW(V_{GS} - V_T)$, where $C'_o = \varepsilon_{ox}\varepsilon_o/T_{ox}$ is the gate oxide capacitance per unit area and L and W are the length and width of the channel, respectively. The parameters $t_t = L/v_D = L^2/\mu V_{DS}$, where v_D is the electron's drift speed, μ is the electron's mobility in the channel, and V_{DS} is the drain-to-source voltage. A more accurate expression for conduction in the resistive or linear region for low V_{DS}, which takes into account the reduction in gate voltage near the drain, is equation 16:

$$I_{DS} = \mu C'_o \frac{W}{L}\left[(V_{GS} - V_T) - \frac{V_{DS}}{2}\right]V_{DS} \tag{16}$$

The channel conductance ($\sim I_{DS}/V_{DS}$) is proportional to the ratio of width to length for the channel, W/L. The parameter $\mu C_o'$ is determined by the fabrication process, and L by photolithography leaving the layout (W/L) to be adjusted by the circuit designer to achieve an appropriate trade-off between circuit speed and area. For V_{DS} small compared with $V_{GS} - V_T$, the channel conductance is a constant, increasing with $V_{GS} - V_T$.

As V_{DS} increases, the depletion width at the drain junction grows and can accommodate more charge. Thus, less charge is needed in the inversion layer to balance the gate charge. Because the surface potential at the drain edge of the channel is V_{DS}, when $V_{GS} - V_{DS} < V_T$ inversion can no longer be maintained and the inversion layer is pinched-off at the drain end. For $V_{DS} \geq V_{GS} - V_T$, the MOSFET is in the saturation region and the saturation current I_{DS} is given by equation 17. For $V_{DS} > V_{GS} - V_T$ the pinched-off region can

$$I_{DS} = \frac{1}{2}\mu C_o' \frac{W}{L}(V_{GS} - V_T)^2 \tag{17}$$

accommodate a large potential drop, which fixes I_{DS} at its saturated value. However, if the electric field in the channel, V_{DS}/L, is large enough the drift velocity may reach its saturation value, v_s, as in Figure 3, and equation 17 approaches equation 18.

$$I_{DS} = \frac{1}{2} W C_o' v_s (V_{GS} - V_T) \tag{18}$$

Several features should be noted about MOSFET operation. Unlike BJTs, currents flow horizontally in MOSFETs and the amount of current is limited by the depth and width, W, of the channel, as well as the channel resistance, which is proportional to the length, L, of the channel. As a result, MOSFETs do not have as much current drive as BJTs, where the current is limited by the area of the emitter junction. The discussion so far has assumed that $V_T > 0$, corresponding to an enhancement mode NMOSFET or NFET. Suitable channel doping can cause $V_T < 0$, corresponding to a depletion mode NMOSFET. Depletion mode MOSFETs conduct when $V_{GS} = 0$ and were used to provide pull-up resistors in NMOS technology. The MOSFET is a symmetrical device that can conduct equally well in either direction. This allows it to be used in switch or pass-transistor logic as well as static and dynamic logic families.

The previous discussion about NMOSFETs applies equally well to PMOSFETs or PFETs. The PMOSFET channels have hole conduction between p^+ drain and source regions implanted into an n-type substrate. The parameters $V_T < 0$ and $V_{GS} < V_T$ for conduction. Because the hole mobility is approximately half the electron mobility, PFETs provide about half the gain for a given device width. Designers can compensate for this in the layout of a circuit by making PFET gate widths twice as wide as NFET gate widths.

9.2. CMOS Logic. Both NFETs and PFETs can be combined to form complementary MOS (CMOS) circuits (28). Figure 12 shows the circuit diagram for a CMOS inverter, the basic building block for CMOS logic, and the corresponding device cross section. It is customary to choose $V_{ss} = 0$ (ground) as the

NFET source and V_{DD} (supply line) as the PFET source. For many years, corresponding to gate lengths $\geq 1\,\mu m$, the supply voltage remained at $V_{DD} = 5\,V$. The cross-section is for a *p*-well process in which the NFETs n^+ source and drain regions are implanted into a *p*-type well formed by implantation of acceptors into an *n*-type substrate. The scale of a process is determined by the length of the polysilicon gates, which is 2 nm for a 2-μm process. Note that the thickness of the gate oxide (40 nm for a 2-μm process) is much less than the other thicknesses of material.

When 5 V is applied to the input, the NFET is turned on and the PFET is cut off. The pull-down current flowing through the NFET pulls down the output to zero. Because the output is connected to the gates of subsequent MOSFETs, the output sees a capacitor. Thus, pulling down the output to zero requires discharging this capacitor. If the output was originally 5 V, the NFET is in the saturation region and maximum current flows. As the output drops, the NFET enters the resistive region, less current flows, and this increases the fall time of the output. Similarly, when 0 V is applied to the input, the NFET is cut off and, because $V_{GS} = -5\,V$, the PFET turns on and the pull-up current flowing through the PFET charges the capacitor load to 5 V in some rise time.

Note that there is a four-layer p^+-*n*-*p*-n^+ structure between the p^+ PFET and n^+ NFET drain diffusions. Such a structure behaves like a thyristor and can be triggered into a high current mode, called latchup, which can destroy CMOS circuits (29). Latchup can be prevented if the gains of the parasitic p^+-*n*-*p* and *n*-*p*-n^+ BJTs in the p^+-*n*-*p*-n^+ structure can be kept sufficiently low. Shallow trench isolation (STI) is very effective in reducing latchup.

Current only flows in the CMOS inverter when the load capacitor is being charged or discharged. No current flows to maintain a logic level (0:0 V, 1:5 V). Because power is dissipated only when current flows, the amount of power dissipation in a CMOS circuit is proportional to circuit activity, the charging or discharging of capacitors. In equation 19, N_{SW} is the number of circuits switching with load capacity C, and f is the frequency of operation (30); f would be the clock frequency for synchronous digital circuits.

$$P = N_{SW} C f V_{DD}^2 \tag{19}$$

Traditionally, CMOS has been regarded as a low power circuit technology. However, this is only true for circuits operating at low frequency, with low activity, or at low supply voltages. Modern microprocessor chips, operating at gigahertz frequencies with hundreds of millions of transistors can dissipate considerable power. This leads to major concerns about heat dissipation and, for laptop computers, battery drain.

9.3. Moore's Law and Device Scaling. The remarkable increase in the number of MOSFETs per chip, following Moore's law, has been the direct result of shrinking device dimensions. If the gate length, L, and width, W, are decreased by a factor of 2, four times as many FETs can be placed in the same area. Generalized scaling (31) has allowed the behavior of FETs to remain the same when their size shrinks. In constant-field scaling, the electric fields are kept the same as device dimensions shrink. This allows the circuit

characteristics to keep the same shape, which means that the gate oxide thickness and the depth of the source and drain junctions must shrink by the same factor, $1/\alpha$, as L and W. It also means that V_{DD} and V_T need to shrink by the same factor, $1/\alpha$. Shrinking V_T requires increasing the substrate doping by α to maintain the relative size of the depletion region. Constant field scaling reduces both the MOSFET input capacitance (~load capacitance) and gate delay by $1/\alpha$. Both the dc and dynamic power consumption of MOSFETs are reduced by a factor of $1/\alpha^2$, as equation 19 would predict. On a physical basis, constant-field scaling works with little modification down to gate lengths of ~100 nm. The doubling of functions on a chip predicted by Moore's law can be achieved with $\alpha = \sqrt{2}$.

An alternative is constant-voltage scaling in which the supply voltage is held constant while scaling device dimensions. This reduces gate delay by a factor of $1/\alpha^2$, which is good, but increases MOSFET power consumption by a factor of α, which is bad. This is particularly true for applications such as laptop computers where low power simplifies packaging requirements and increases battery life.

Not only the MOSFETs, but their interconnections need to shrink when a chip is scaled. However, wire resistance scales as α and wire capacitance scales as $1/\alpha$, so that this RC time constant with a significant effect on interconnect delays does not change when CMOS chips are scaled. Thus, interconnect delays play a larger role when CMOS chips are scaled to submicron dimensions.

As device dimensions shrink, subthreshold leakage becomes an increasing problem. The parameter I_{DS} does not cut off sharply at V_T. Below V_T, I_{DS} depends exponentially on $(V_G - V_T)/KT$ (28). This means that when $V_{GS} = 0$, a noticeable leakage current will still flow, leading to significant power consumption when there are millions of transistors on a chip. Note that this is exacerbated by low threshold voltages.

Constant field scaling requires that voltages shrink as device dimensions shrink to maintain a constant electric field in the device. This means that, as device dimensions shrink by a factor of $1/\alpha$, the depletion layer thickness, the supply voltage, and V_T must all shrink by the same factor of $1/\alpha$. For an NFET the depletion layer thickness depends upon n_A, the concentration of acceptors in the silicon substrate. Thus, n_A must be increased to reduce the depletion layer thickness and, consequently, V_T.

Table 3 shows projections for CMOS technology selected from ITRS 1999 (32), beginning with the half-pitch technology node of 180 nm in 1999 and continuing to 100 nm in 2005. Below the technology node we see the gate length in nanometers for microprocessors, which is the most demanding application. Note that these lengths are substantially smaller than the half pitch lengths. Moore's law defines technology generations by the doubling times for functions on a chip. For microprocessors this can be measured by millions of transistors per chip. Table 3 shows that the doubling of transistors on a microprocessor chip is expected at half pitches of 150, 120, and 100 nm, while across-chip clock frequency increases by $\sqrt{2}$ at these half pitches. This is what scaling would one to believe, but a closer look shows that scaling, which had been a good guide up to this point for the semiconductor industry, is beginning to break down. The MPU gate lengths (L) shrink by 0.71 ($\sim 1/\sqrt{2}$), 0.80, and 0.81

for these half pitches. (Note that the defined half-pitch technology nodes of 180, 130, and 100 nm show a shrinkage of 0.72 ($\sim 1/\sqrt{2}$) and 0.81; a 90 nm half pitch would be much closer to a $1/\sqrt{2}$ shrink.) Table 3 shows that doubling the number of transistors on a chip has required increasing chip sizes and increasing clock frequency has required holding V_{DD} nearly constant.

Constant field scaling requires the oxide thickness to shrink along with W and L to maintain the same electric field in the gate oxide as the gate voltage shrinks. Now, a limit is set by electrons or holes tunneling through the gate oxide and Table 3 shows a steady increase in the leakage current, measured in nanoampere per unit width (W) of the gate in microns. Oxide tunneling currents for gate voltages of 2 V are $\sim 1\,A/cm^2$ for oxide thicknesses of 2 nm, rising to $10\,kA/cm^2$ for oxide thicknesses of 1 nm (28). For $W = 1\,\mu m$ and $L = 100\,nm$, $1\,A/cm^2$ would correspond to a tunneling current of 100 nA, which is much greater than the leakage current limits set in Table 3. Consequently, a thermal oxide is no longer a suitable gate dielectric and t_{ox} has been replaced by an equivalent effective oxide thickness (EOT), defined by $EOT = (3.9/\kappa_{eff})\ t_{gate\ dielectric}$, where $\kappa_{eff} = 3.9$ for SiO_2. Silicon nitride (Si_3N_4) has $\kappa_{eff} \sim 7$ so that silicon nitride and oxide stacks are generally used for the gate dielectric, allowing thicker dielectric layers with much lower tunneling currents.

Equation 19 indicates that increasing clock frequency while maintaining supply voltages is a recipe for high chip power; Table 3 confirms this. Removing the heat generated by this power has become an increasing problem, since it will raise chip temperatures and this leads to increased reliability problems. Issues of battery drain for laptop computers compound the problem.

9.4. Scaling to Deep Submicron Dimensions. Table 1 showed device and chip properties from ITRS 2004; essentially the same categories are listed as in Table 3. Note that Moore's law continues to be maintained. For each technology generation (node) the functions per chip double, but the chip size remains fixed at $280\,mm^2$. Note also, that the technology nodes now decrease by $1/\sqrt{2}$ and that hp90, the 90 nm half-pitch for a DRAM, in 2004 replaces the projection of 100 nm in 2005 from Table 3. However, these results have not been achieved by simple constant-field device scaling.

In Table 1, the physical gate length for a high performance microprocessor is the final, as-etched length of the bottom of the gate electrode. This is $\sim 0.4\times$ the technology node and is distinguished from the printed gate length, which is $\sim 1.4\times$ greater. The on-chip local clock frequency is for high performance, lower volume microprocessors in localized portions of the chip, rather than a global across-chip clock frequency. This primarily reflects the increasing influence of interconnect delays, particularly for global interconnects crossing a substantial fraction of the chip. The supply voltage has remained nearly constant and the source-drain leakage current, $I_{sd\ leak}$, continues to rise. The result is a continued rise in maximum allowed power, which needs to rise by only 20% from hp90 to hp65 and by only 5% from hp65 to hp45. It can be argued that classical scaling of silicon technology died somewhere around the 130–90-nm node, and power consumption has hit a wall (34).

To quote ITRS 2004 (12), "the scaling of the numbers in the tables reflects a particular scaling scenario in which we have attempted to optimally scale to meet the key goal for high-performance logic, 17% per year average improvement

in the NMOS intrinsic switching speed, while delaying as long as feasible the projected need for major innovations. These include innovations such as metal gate electrode, high-κ gate dielectric, and novel doping and annealing techniques to reduce the parasitic source-drain resistance." Three key concerns for improving transistor performance are fast switching speed, which is closely related to high drive currents, low leakage currents for low standby power, and controlling the short channel effect (SCE) that causes V_T to decrease when channel lengths are reduced.

The approaches for achieving higher switching speeds are defined by the need for drive MOSFETs to charge or discharge MOFSET input capacitances. The general relation $\Delta Q = I\ \Delta t = C\ \Delta V$ can be rewritten as $\Delta t = C\ \Delta V/I$. Shorter switching times require some combination of lower C, lower ΔV, or higher I. A convenient way of expressing this is in terms of the intrinsic MOSFET delay, $\tau = (C_{\text{gate}} \times V_{\text{DD}})/I_{\text{d, sat}}$ (12), which shows the close relationship between switching speed and drive current.

Strained silicon is a nonscaling approach to achieving high drive currents that was introduced by Intel in the 90-nm node. In general, tensile strain increases electron mobility, benefiting NFETs, while compressive strain increases hole mobility, benefiting PFETs. Strain can be created locally by film stresses due to device structures (process-induced strain) or globally by straining silicon wafers (whole-wafer strain) (35). Interestingly, (uniaxial) process-induced strains give the biggest boost to PFET performance, while (biaxial) whole-wafer strains give the biggest boost to NFET performance. Both tensile and compressively stressed nitride contact liners can be incorporated in a CMOS process flow to improve both NFETs and PFETs (36).

A strained silicon wafer can be created by growing a pseudomorphic top silicon layer on top of a buffer layer ~ 3–4-nm thick that forms a crystal lattice larger than silicon. Typically this buffer layer is silicon with $\geq 20\%$ germanium (35). However, thermally induced out-diffusion of Ge into the strained Si layer can decrease its strain and lead to a significant increase of interface trap densities when it reaches the gate oxide interface (37). Alternatively, strained silicon can be combined with silicon-on-insulator (SOI) technology, in which MOSFETs are fabricated on a thin layer of crystalline silicon that is separated from the substrate by a thick ($\sim$100 nm) layer of SiO_2. The SOI almost eliminates junction capacitances, eliminates the dependence of V_T on bulk silicon voltages, and reduces soft error effects due to radiation. (The reduction of junction capacitances for the drive transistors will increase device speeds.) This has led to a complex engineered wafer fabrication process in which the strained silicon layer produced on a donor wafer is transferred to an SOI wafer. Thus, deep submicron devices require a wafer technology that is well beyond the bulk silicon wafers used in earlier CMOS generations.

Moving to deeper submicron CMOS transistors is likely to require some combination of new materials and new transistor structures (29). Gate stacks that include high κ gate dielectrics, likely HfO_2 with $\kappa = 26$–30, reduce tunneling currents because they can be thicker for the EOT needed to improve MOSFET performance. Unfortunately, implementation has been delayed for two reasons. First, high κ dielectrics are incompatible with polysilicon gates because the Fermi level is pinned at the poly/high κ interface, causing high V_T. Thus, high

κ dielectrics may require metal gates with a p-type metal for PFETs and an n-type metal for NFETs. Ironically, metal gates may return to MOSFETs when oxides are no longer used. Second, poly/high κ transistors have severely degraded channel mobility. Good results have been reported using HfSiON gate dielectrics, delivering 90% of the mobility and good V_T stability.

Alternatively, one can move from present two-dimensional (2D) transistor structures to new three-dimensional (3D) structures that wrap the gate electrode around the gate dielectric. Three-dimensional structures provide better control over current leakage and other short channel effects as well as improving performance. Of these structures the finFET, shown schematically in Figure 13, is promising because of its manufacturability and scalability (38). Because the subthreshold swing is almost ideal, one gets lower leakage current and higher drive current. This relaxes the requirements on the gate dielectric, so that the introduction of high-κ gate dielectrics can be delayed by at least a generation.

As scaling becomes more difficult, radically new approaches to nanoscale devices are being researched. One alternative being actively researched is the use of carbon nanotubes, which can be metallic or semiconducting depending on how they are fabricated, that can be used to form transistors. Fabricating large-scale circuits of appropriate nanotubes remains challenging (39). However, this topic is beyond the scope of this article.

10. Other Applications

As microprocessors and memories move beyond computers to become embedded in communications and consumer products, eg, cell phones and other wireless devices, other applications are emerging for silicon-based semiconductor technology. Predominant among them are flash memories, flat-panel displays, microelectromechanical systems (MEMS), and power electronics. New materials are beginning to be incorporated in silicon-based IC devices. Some of these, SiGe and SiC, are silicon-based semiconductors that offer advantages for higher performance or higher temperature operation (see SEMICONDUCTORS, COMPOUND SEMICONDUCTORS). It is also the case that, as the minimum feature sizes of devices shrink to nanometer dimensions, back-end processing of multilayer interconnects is becoming more important in determining IC performance. Lower temperature operation can improve both CMOS and interconnect performance while reducing the impact of subthreshold leakage currents and exponentially reducing Arrhenius-based degradation mechanisms. Nevertheless, the prospects for cryogenic operation are limited by refrigeration technology, even though IC cooling has become an increasing concern for high performance ICs (40).

Interestingly, even though silicon is an indirect gap material, the value of integrating light emission with silicon chips has led to several approaches for efficient light emission from silicon. One approach that has achieved an external quantum efficiency of 10% (comparable to standard III–V LEDs) involves implanting rare earths, eg, erbium or cerium, in a layer of silicon-rich oxide, which is silicon dioxide enriched with silicon nanocrystals 1–2 nm in diameter (41). Another approach uses the Raman effect to produce a continuous laser beam at a new wavelength, when driven by an external laser. The limits imposed

by two-photon-generated free carrier absorption on stimulated Raman scattering in silicon waveguides were avoided by designing a PIN diode structure, which has a layer of intrinsic silicon in the middle of a PN junction, along a low loss silicon-on-insulator (SOI) rib waveguide (42).

11. Nonvolatile and Flash Memories

Volatility, the loss of memory when power is removed has been a concern since MOS SRAM and DRAM memory chips were introduced in the mid-1960s. Two concepts for nonvolatile memory (NVM) were introduced in 1967. Charge can be retained for over ten years when stored at 125°C on a floating gate of polysilicon surrounded by oxide. Alternatively, charge can be stored in the gate insulator of a metal/nitride/oxide/semiconductor (MNOS) memory transistor. A 1 kb uv-erasable programmable read-only memory (EPROM) chip was introduced in 1971 shortly after 1-kb random access memory (RAM) chips were marketed.

Electrically erasable NVMs (EEPROMs) remained ~10 % of the total semiconductor market in the 1990s, reaching a market value of ~$4 out of a $65 billion memory market by 1999 (43). By that time flash EEPROMs featuring bulk erasure of memory blocks had grown to ~75 % of the NVM market. Flash memory transistors (cells) have a floating gate between the control gate and channel of an NFET. An excess of electrons on the floating gate shields the channel when a positive voltage is applied to the control gate, corresponding to raising the threshold voltage at which the channel conducts.

Flash memories are generally fabricated as NOR or NAND arrays of memory cells. In NOR arrays, the drain electrodes of memory transistors are connected in parallel to bit lines, while their source electrodes are connected to a common source line. In NAND arrays, eight or more memory transistors are arranged in series sandwiched between ordinary NFETs that allow selective connections to a bit line or source line. In both types of array, word lines connect a row of gate electrodes for the memory transistors.

The NOR cells are erased by Fowler-Nordheim (FN) tunneling of electrons from the floating gate to the control gate when a negative voltage is applied to the control gate and a positive voltage is applied to the source electrode. These cells are generally programmed by grounding their source and placing high voltages on the control gate and drain electrodes so that hot electrons are created in the channel (CHE), allowing some with energy greater than the 3 eV Si/SiO_2 energy barrier to travel to the floating gate (CHE injection). The NAND cells are erased by FN tunneling from the floating gate to the channel and programmed by FN tunneling from the channel to the floating gate. The NAND flash is usually used for disk replacement because their cell density is higher but their reading speed is lower.

Since the threshold voltage of an NVM transistor depends directly on the number of electrons stored on the floating gate, n bits can be stored in a single cell provided 2^n distinct voltage levels can be programmed and detected. The parameter n is generally two and two-bit per cell NOR and NAND memories have become fairly common. Samsung researchers have reported a 4 Gb four-level NAND flash memory fabricated in a 3.3-V, 90-nm CMOS process (44).

Recently, flash memory has become the Moore's law technology driver (45), replacing DRAMs and microprocessors. With transistor densities doubling about every year, flash became a $4.8 billion market in 2004. More flash memory, in terms of the number of transistors, is expected to ship in 2005 than has been produced in the entire history of flash production to date. Flash chips are leading in the introduction of new technology. They are already using hafnium-based high κ gate dielectrics and will likely introduce new transistor structures beyond the 45-nm node. Infineon has built the world's smallest nonvolatile flash memory cell using finFETs with 20-nm gate dimensions. These could be used in 32-Gb flash memory chips.

12. Displays

Full-color flat-panel displays (FPDs) became a significant market as a necessary component of a growing mobile laptop or notebook computer market. The FPDs consist of an array of closely spaced, separately contacted picture elements or pixels. Thin-film transistor active matrix liquid-crystal displays (AMLCDs) have become the dominant technology for laptop computer displays. The AMLCDs sandwich liquid-crystal material between two plates of glass. One plate contains a matrix of thin-film transistors (TFTs); the other plate contains an array of color filters and polarizers. Assembly involves placing spacers and the injection of LCD material between the plates, followed by sealing the two plates together (46). In operation, the TFT connects the LCD capacitor, which is formed by the LCD material between electrodes on each plate, to a source line when the TFT is activated by a gate line. Each pixel is driven over a gray-scale range as the TFT MOSFET is driven from threshold to saturation. Amorphous-silicon TFTs have been used, despite their poor mobilities ($<1\,\mathrm{cm}^2/\mathrm{Vs}$), because the glass plates require low temperature processing ($<500^\circ\mathrm{C}$). Polysilicon TFTs, with mobilities $50-100\times$ higher, would be preferable, allowing driver inverters to be processed on the same plate, provided a suitable process can be found. Organic TFTs, using materials such as pentacene ($\leq 1.7\,\mathrm{cm}^2/\mathrm{Vs}$), are being considered as well.

At this time, small AMLCD displays are ubiquitous on cell phones and cameras. Larger AMLCD displays are now common as flat panel computer monitors and televisions. Indeed, the hope of replacing cathode-ray tube (CRT) television displays with FPD displays was an original incentive for developing this technology. There are several competing large-screen FPD technologies (47). These include plasma FPD, projection FPD, and field emission FPD (FED). Plasma FPDs can be quite large (60 in. diagonal screen size) and emit light by forming a plasma discharge between two glass plates with phosphors deposited on one plate. They are $\sim$50% brighter than AMLCD FPDs. FEDs use an array of field emission microtips as cathodes that emit electrons onto phosphor stripes on a glass plate serving as an anode. The distance between the cathodes and anode is a few millimeters. Projection FPDs may use an array of rotatable micromirrors and are fabricated from silicon wafers using the techniques described in the next section.

13. Microelectromechanical Systems (MEMS)

Micromechanics uses silicon as a structural material, using microelectronic fabrication techniques to micromachine micromechanical structures and machines. Although single-crystal silicon is brittle and fractures when its elastic limit is exceeded, it has a higher elastic limit than steel and remains strong under repeated cycles of tension and compression. Chemical etching techniques have been developed that allow formation of basic 3D elements, eg, pits, trenches, holes, diaphragms, and cantilevers. Anisotropic etchants, eg, potassium hydroxide (KOH), are particularly useful for *bulk micromachining* of silicon because the etch rate depends dramatically on crystal plane. For silicon, the KOH etch rate is $\sim$400:1 for { 100 }/{ 111 } crystal planes (48). More recently, surface micromachining has become popular. *Surface micromachining* is based on the patterning of a thin sensor film, typically polysilicon, that has been deposited on a sacrificial spacer material, typically SiO_2, that is subsequently removed. Since deposited films for surface micromachining exhibit built-in stress, sensor and spacer thicknesses were limited to $\sim 1\mu$m to facilitate integration. Micromechanical applications include the fabrication of ink-jet nozzle arrays, gas chromatographs, and microgrooved heat exchangers for cooling semiconductor chips. Techniques have been developed for fabricating rotating gears and miniature motors.

A simple example of a MEMS element is a cantilever beam formed from silicon. This was the basis for an integrated surface-micromachined accelerometer used by automobiles to determine when safety requires air bag inflation. Analog Devices developed an accelerometer based on a multiple cantilever structure that occupied $\sim$5% of a BiCMOS analog IC containing the sensor electronics for this MEMS chip (49). Similar process technology has been used to fabricate an array of micromirrors, whose rotation can be controlled electrically, on an IC chip. This is the basis of a light modulator for a high definition projection system developed by Texas Instruments (50) with 442,368 mirrors for a 768×576 pixel array on a silicon chip. Silicon cantilevers containing probe tips 1.7 μm high with apex radii $<$20 nm are the basis for atomic force microscopy (AFM). This technology has been adapted for ultradense data storage in IBMs Millipede project (51) in which such a cantilever tip writes a 1 on a polymer medium by thermomechanically forming a pit in the polymer surface. The areal density of information that can be stored is determined by the probe tip. Areal densities approaching in terabit per square inch (Tb/in.^{2}) can be achieved, which exceeds the theoretical limit for magnetic storage. High data rates are achieved by using multiple cantilevers in parallel. Arrays of $64 \times 64 = 4096$ cantilevers have been fabricated.

14. Power Semiconductors

Power semiconductor devices cover a diverse range of applications from 100 W at microwave frequencies (microwave ovens) to 100 MW at low frequencies (motor drives or power supplies) (52). Automotive electronics and switching

power supplies require relatively low voltages ($<100\,V$), but high current construction.

Avalanche breakdown sets a fundamental limit on the maximum operating voltage for power devices. Because depletion layer curvature increases the crowding of electric field lines and reduces the breakdown voltage, multiple-junction field rings or other termination techniques are used to reduce the depletion layer curvature and increase the breakdown voltage. Large device areas are required to handle large current flows. For example, a single power thyristor may require $>400\,cm^2$. Large chip areas for a single device require careful attention to chip yields, which tend to decrease exponentially with chip area. More vertical channel structures, eg, VMOS, DMOS, or UMOS (52), offer higher performance than the surface channels of conventional MOSFETs. The surface of the chip is covered with an array of parallel MOSFET cell structures.

Silicon carbide, SiC, is a compound semiconductor composed entirely of Group 14 (IV) elements and is expected to have properties intermediate between those of silicon and diamond. Several polytypes of SiC exist, but homoepitaxial growth on 6H−SiC substrates has been preferred. As Table 1 indicates, its high band gap leads to spectacular decreases in intrinsic carrier concentration which means that SiC devices should have excellent performance at high temperatures. Early application for SiC devices were as blue light-emitting diodes (LEDs) and uv photodiode detectors. This means that SiC is a desirable material for power devices, since higher current and voltage ratings can be achieved with higher temperature operation. However, much lower defect densities will be required to achieve the large device areas that power devices need.

BIBLIOGRAPHY

"Semiconductors (Theory)" in *ECT* 2nd ed., Vol. 17, pp. 834–861, by P. J. Dean, Bell Telephone Labs, Inc.; "Semiconductors (Theory and Application)" in*ECT* 3rd ed., Vol. 20, pp. 601–633, by S. A. Schwartz, Bell Laboratories, Inc.; "Semiconductors (Silicon-Based)" in *ECT* 4th ed., Vol. 21, pp. 720–750, by K. Rose, Rensselaer Polytechnic Institute; "Semiconductors, Silicon-Based Semiconductors" in *ECT* (online), posting date: December 4, 2000, by K. Rose, Rensselaer Polytechnic Institute.

CITED PUBLICATIONS

1. J. Bardeen, *Phys. Rev.* **71**, 717 (1947).
2. W. Shockley and G. L. Pearson, *Phys. Rev.* **74**, 232 (1948).
3. J. Bardeen and W. H. Brattain, *Phys. Rev.* **75**, 1208 (1949).
4. W. Shockley, *Bell Syst. Tech. J.* **28**, 435 (1949).
5. W. Shockley, *Electrons and Holes in Semiconductors with Applications to Transistor Electronics*, D. Van Nostrand Co., Inc., Princeton, N.J., 1950.
6. G. L. Pearson and W. H. Brattain, *Proc. IRE* 1794 (1955).
7. C. T. Sah, R. N. Noyce, and W. Shockley, *Proc. IRE* **45**, 1228 (1957).
8. D. Kahng, *IEEE Trans. Electron Devices* **ED-23**, 655 (1976).

9. A. L. Robinson, *Science* **195**, 1179 (1977), special issue devoted to the electronics revolution.
10. G. Moore, *IEEE Spectrum* **30** (Apr. 1979).
11. M. L. Hammond, *Semiconductor Int.* **51** (Jan. 2004).
12. Available at http://www.itrs.net/. ITRS, 2004.
13. M. L. Hammond, *Semiconductor Int.* **102** (Jul. 2004).
14. K. Rose, *IEEE Circuits Devices* **26** (Nov. 1991).
15. H. R. Huff and F. Shimura, *Solid State Technol.* **103** (Mar. 1985).
16. T. M. Brown, *Semiconductor Int.* **223** (May 1987).
17. A. E. Braun, *Semiconductor Int.* **44** (Apr. 2002).
18. S. K. Ghandhi, *VLSI Fabrication Principles*, 2nd ed., John Wiley & Sons, Inc., New York, 1994.
19. S. M. Sze, *Physics of Semiconductor Devices*, 2nd ed., John Wiley & Sons, Inc., New York, 1981.
20. M. Ruff and co-workers, *IEEE Trans. Electron Dev.* **41**, 1040 (1994).
21. C. Kittel, *Introduction to Solid-State Physics*, 6th ed., John Wiley & Sons, Inc., New York, 1989.
22. H. V. Malmstadt, C. G. Enke, and S. R. Crouch, *Electronic Analog Measurements and Transducers*, W. A. Benjamin, Inc., Menlo Park, Calif., 1973.
23. R. F. Pierret, *Advanced Semiconductor Fundamentals*, Addison-Wesley Publishing Co., Reading, Mass., 1987.
24. B. E. Deal and A. S. Grove, *J. Appl. Phys.* **36**, 3770 (1965).
25. B. G. Streetman, *Solid State Electronic Devices*, 4th ed., Prentice-Hall, Inc., Englewood Cliffs, N.J., 1988.
26. S. K. Ghandhi, *The Theory and Practice of Microelectronics*, John Wiley & Sons, Inc., New York, 1968.
27. A. S. Grove, *Physics and Technology of Semiconductor Devices*, John Wiley & Sons, Inc., New York, 1967.
28. Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*, Cambridge University Press, Cambridge, U.K., 1998.
29. P. Singer, *Semiconductor Int*. **46** (Apr. 2004).
30. K. Bernstein and co-workers, *IBM J. Res. Devel.* **39**, 33 (1995).
31. R. H. Dennard and co-workers, *IEEE J. Solid State Circuits.* **SC-9**, 256 (1974).
32. *International Technology Roadmap for Semiconductors*, 1999 ed., 1999.
33. N. H. E. Weste and K. Eshraghian, *Principles of CMOS VLSI Design*, 2nd ed., Addison-Wesley Publishing Co., Reading, Mass., 1994.
34. A. E. Braun, *Semiconductor Int.* **19** (June 2005).
35. P. Singer, *Semiconductor Int.* **28** (Dec. 2004).
36. P. Singer, *Semiconductor Int.* **28** (Jan. 2005).
37. C. Arena and co-workers, *Semiconductor Int.* **40** (Mar. 2005).
38. L. Peters, *Semiconductor Int.* **47** (Mar. 2005).
39. P. Singer, *Semiconductor Int.* **26** (Oct. 2004).
40. K. Rose and co-workers, *Crit. Rev. Solid State Mater. Sci.* **24**, 63 (1999).
41. P. Singer, *Semiconductor Int.* **26** (Dec. 2002).
42. P. Singer, *Semiconductor Int.* **26** (Apr. 2005).
43. W. D. Brown and J. E. Brewer, eds., *Nonvolatile Semiconductor Memory Technology*, IEEE Press, New York, 1998.
44. S. Lee and co-workers, *Proc. IEEE Int. Solid-State Circuits Conf.*, Paper 2.7, 2004.
45. J. Chappell, *Electronic News* (July 17, 2005).
46. P. Singer, *Semiconductor Int.* **99** (Nov. 1996).
47. R. DeJoule, *Semiconductor Int.* **59** (Jan. 1997).
48. K. E. Peterson, *Proc. IEEE* **70**, 420 (1982).

49. T. A. Core and co-workers, *Solid State Technol.* **39** (Oct. 1993).
50. M. A. Mignardi, *Solid State Technol.* **63** (July 1994).
51. P. Vettiger and co-workers, *Proc. IEEE International Electron Devices Meeting*, p. 763, 2003.
52. B. J. Baliga, *Physics of Power Semiconductor Devices*, PWS Publishing Co., Boston, Mass., 1995.

GENERAL REFERENCE

M. K. Balazs, *Solid State Technol.* **75** (Oct. 1993).

KENNETH ROSE
SIDDHARTH DEVARAJAN
Rensselaer Polytechnic Institute

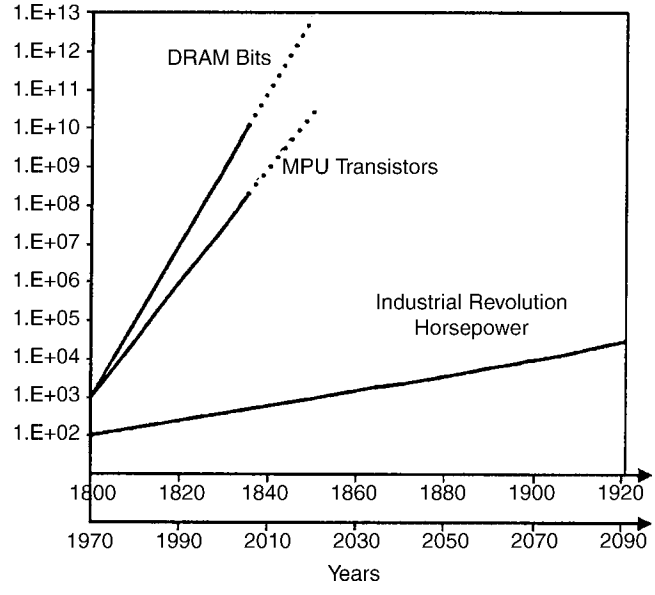

Fig. 1. This figure compares the rates of technological change in the microelectronics and industrial revolutions.

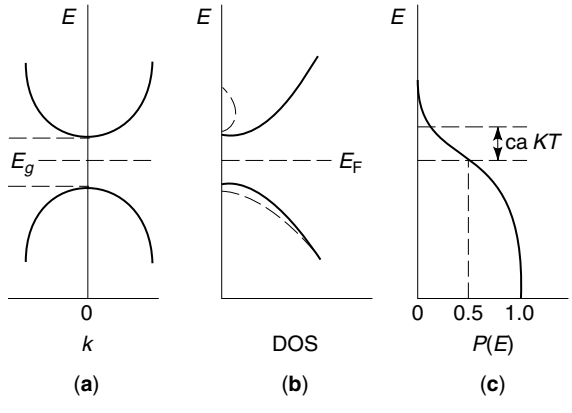

Fig. 2. (**a**) Energy, E, versus wave vector, k, for free particle-like conduction band and valance band electrons; (**b**) the corresponding density of available electron states, DOS, where E_F is Fermi energy; (**c**) the Fermi-Dirac distribution, ie, the probability $P(E)$ that a state is occupied, where K is the Boltzmann constant and T is the absolute temperature in kelvin. The tails of this distribution are exponential. The product of $P(E)$ and DOS yields the energy distribution of electrons shown by (– – –) in (**b**).

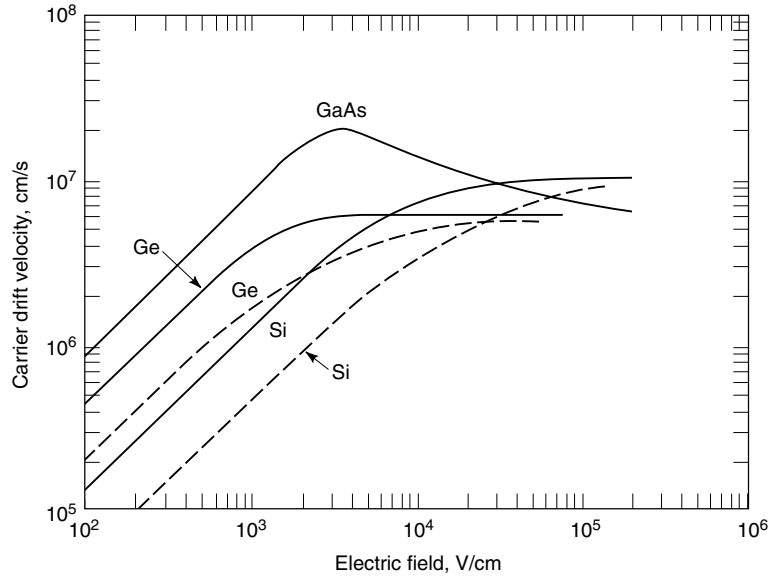

Fig. 3. Electron (——) and hole (– – –) velocities versus electric field for high purity silicon, Si; germanium, Ge; and gallium arsenide, GaAs, at 300 K (19).

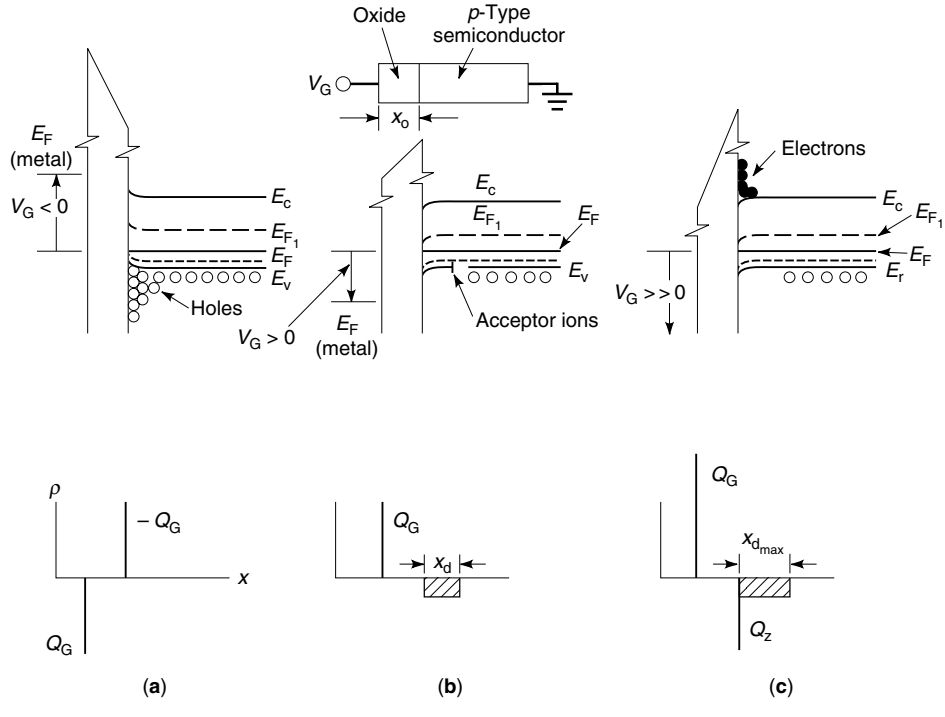

Fig. 4. Charge distributions in the voltage-biased MOS capacitor: (**a**) accumulation of majority carriers near surface; (**b**) depletion of majority carriers from surface; (**c**) inversion, accumulation of minority carriers near surface (20). The parameter V_G = gate voltage; Q_G = gate charge; x_d and $X_{d_{max}}$ = depletion widths; and ρ = charge density.

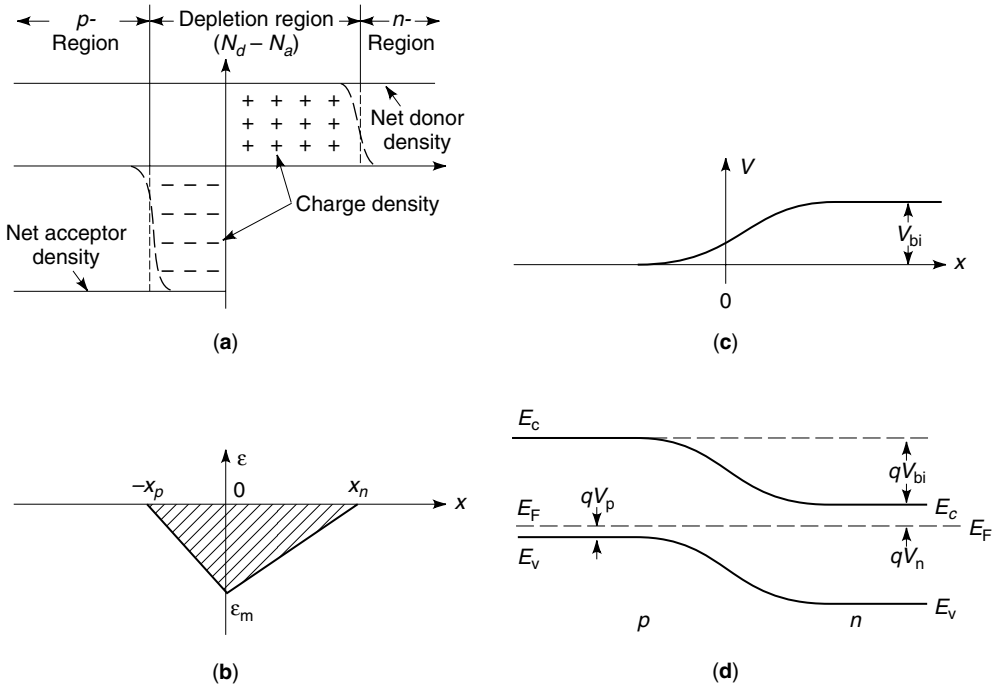

Fig. 5. An abrupt *p-n* junction in thermal equilibrium: (**a**) space-charge distribution where (– – –) indicate majority carrier distribution tails and the charge density is the result of unneutralized impurity ions; (**b**), where ε_m is the maximum electric field in the depletion region, and (**c**) spatial variations of the electric field and corresponding potential where V_{bi} is built-in potential or voltage and ⊿ is the area of diffusional potential; (**d**) energy band diagram where qV_p and qV_n are energies of ionized acceptors, a, and donors, d, relative to the band edges E_v and E_c, respectively. Where the band edges are flat, there is no electric field and the potential does not change (neutral regions) (19).

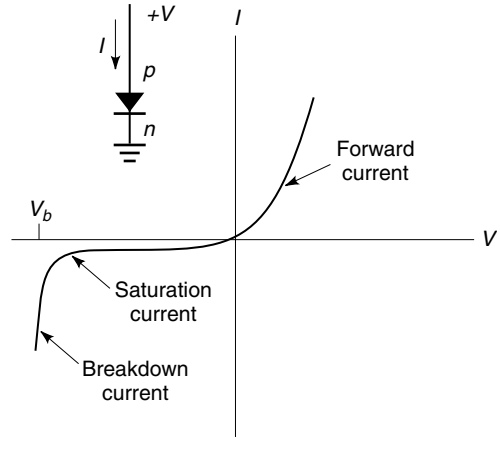

Fig. 6. Circuit symbol and current, I, voltage, V, characteristic of a *p-n* junction diode, where V_b = breakdown voltage (25).

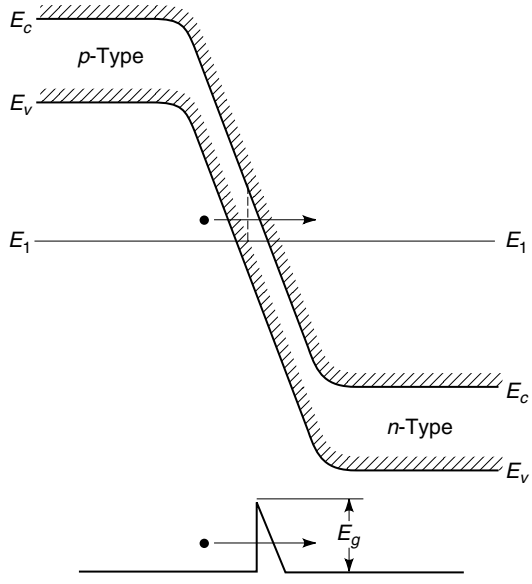

Fig. 7. Band-to-band tunneling in a semiconductor where the triangular barrier leads to Fowler-Nordheim tunneling (26).

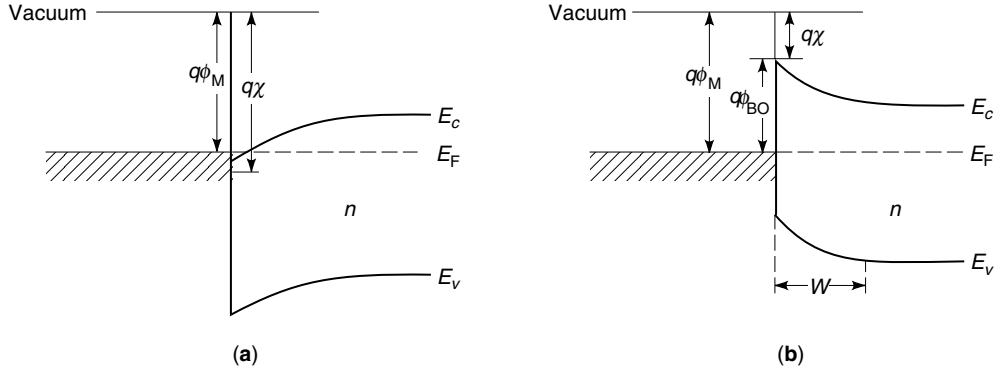

Fig. 8. Schottky barrier band diagrams: (**a**) a rare situation where the metal work function $q\phi_M$ is less than the semiconductor electron work affinity $q\chi$, resulting in an ohmic contact; (**b**) normal Schottky barrier with barrier height $q\phi_{BO}$. When the depletion width W is <10 nm, an ohmic contact forms.

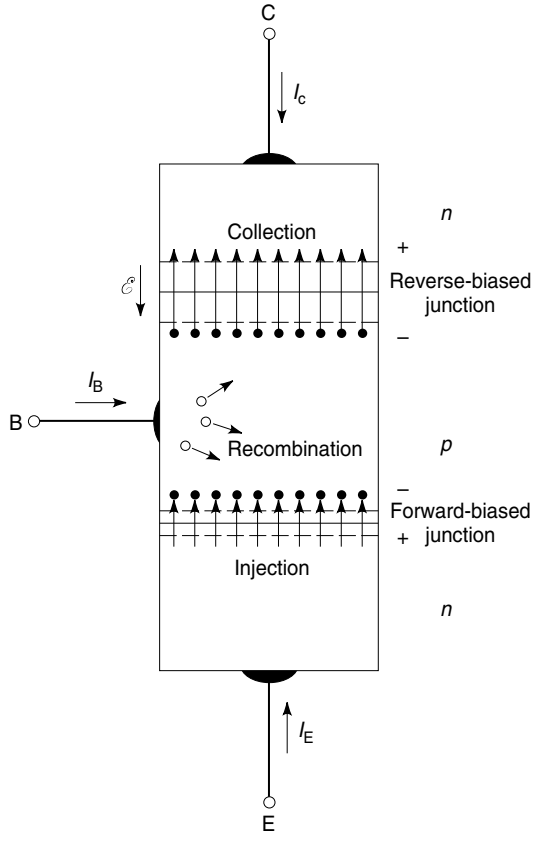

Fig. 9. The *n*-*p*-*n* transistor biased in its active region, where $I =$ current, (– – –) indicate depletion regions at the *p*-*n* junctions, and $\mathcal{E}$ is the electric field; their width is reduced for the forward biased base-emitter (BE) junction and increased for the reverse biased base-collector (BC) junction. Note the directions of electron (●) and hole (○) flow (26).

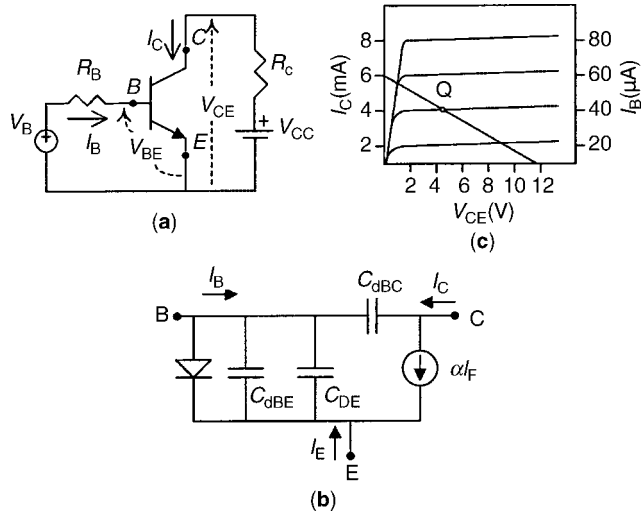

Fig. 10. (**a**) Common symbols for an *n-p-n* BJT in a grounded-emitter circuit. (**b**) Large-scale equivalent circuit that captures the transistor's essential characteristics for normal operation. (**c**) Output characteristics for the *n-p-n* BJT.

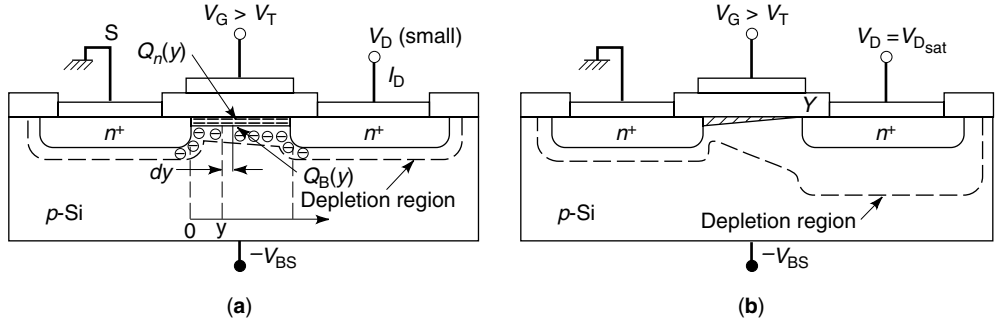

Fig. 11. NMOSFET operated (**a**) in the linear (low drain voltage) region, where S is the grounded source contact, y is the distance from source to drain, and d is differential increase; and (**b**) at the onset of the saturation region, $V_D = V_{D_{sat}}$. The point Y indicates the channel pinch-off point. For $V_D > V_{D_{sat}}$ point Y moves to the left, reducing the effective channel length, but the drain current remains nearly constant. See text (20).

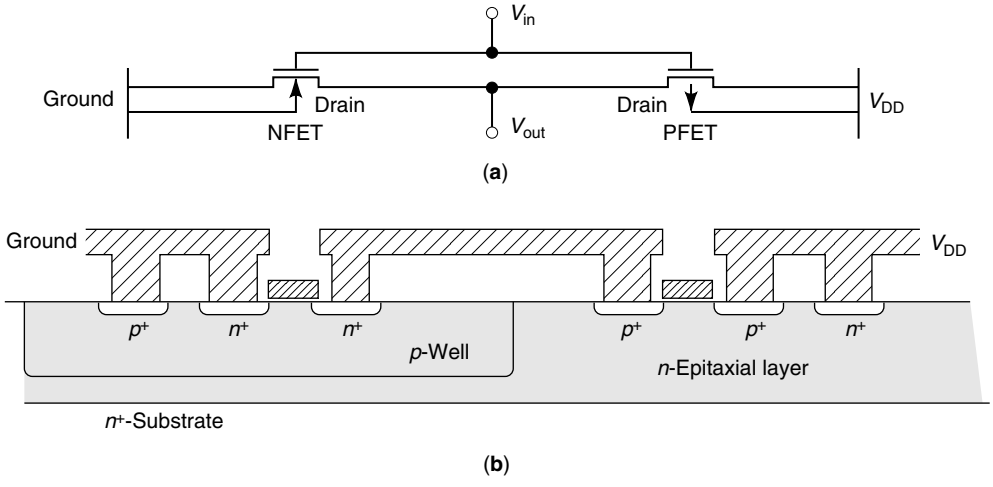

Fig. 12. (**a**) The CMOS inverter circuit. The FET circuit symbols emphasize that MOS–FETs are actually four-terminal devices in which the n substrate is connected to V_{DD} for the PFET and the p substrate is connected to the ground for the NFET. Note the conventions on drain location for the PFET and NFET. (**b**) Corresponding cross-sectional view roughly to scale for a 2-μm CMOS process, where □ represents silicon, □ SiO_2, ▨ polysilicon, and ▨ aluminum.

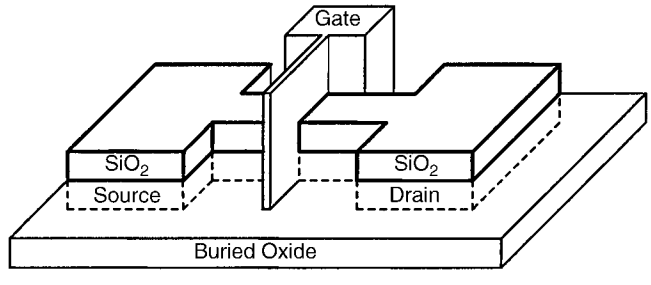

Fig. 13. The schematic representation of the structure of finFET.

Table 1. **Projection of Silicon Technology**[a]

Year	2004	2007	2010
technology node, nm	hp90	hp65	hp45
MPU physical gate length, nm	37	25	18
functions per chip (M transistors)	386	773	1546
chip size, mm^2	280	280	280
V_{DD}, V	1.2	1.1	1.0
on-chip local clock frequency, GHz	4.2	9.3	15.1
EOT, mm	1.2	0.9	0.7
$I_{sd\ leak}$ (at 25°C), nA/μm	50	70	100
maximum power, W	158	189	198
MPU printed gate length, nm	53	35	25

[a]See Ref. 12.

Table 2. **Important Properties[a] of Group 14 (IV) Semiconductors**

Material	Cubic lattice constant, pm	Band gap, eV	Intrinsic carrier concentration, cm^{-3}	Relative dielectric constant, ε_r	Mobility, $cm^2/(V \cdot s)$	
					Electrons	Holes
C (diamond)	356.683	5.47	7.6×10^{-26}	5.7	1800	1200
GH-SiC	[b]	3.0	1.6×10^{-6}	9.66		
SiC					380	75
Si	543.095	1.12	1.45×10^{10}	11.9	1500	450
Ge	564.613	0.66	2.4×10^{13}	16.0	3900	1900

[a] At 300 K.

[b] a-SiC crystallizes in the wurtzite lattice with $a = 308.6$ pm and $c = 1511.7$ pm (values from Refs. 19 and 20).

Table 3. **Moore's Law Scaling**[a]

Year	1999	2001	2002	2003	2005
DRAM ½ pitch, mm	180[b]	150	130[b]	120	100[b]
MPU gate length, nm	170	100	85	80	65
functions/chip (M transistors)[c]	23.8	47.6		95.2	190
chip size, mm^{2}[c]	340	340		372	408
V_{DD}, V[b]	1.8	1.5	1.5	1.5	1.2
clock frequency, GHz	1.25	1.77	2.10	2.49	3.5
EOT, mm[d]	1.9–2.5	1.5–1.9	1.5–1.9	1.5–1.9	1.0–1.5
I_{off} at 25°C, nA/μm[e]	5	8	10	13	20
maximum power, W[c]	90	115	130	140	160

[a]See Ref. 12.
[b]Identified technology node.
[c]For a high performance microprocessor chip.
[d]Effective oxide thickness = EOT.
[e]For a minimum L high performance device.