IOPscience

HOME | SEARCH | PACS & MSC | JOURNALS | ABOUT | CONTACT US

Physical limits of silicon transistors and circuits

This article has been downloaded from IOPscience. Please scroll down to see the full text article.

2005 Rep. Prog. Phys. 68 2701

(http://iopscience.iop.org/0034-4885/68/12/R01)

The Table of Contents and more related content is available

Download details:
IP Address: 193.205.222.52
The article was downloaded on 30/03/2009 at 15:01

Please note that terms and conditions apply.

INSTITUTE OF PHYSICS PUBLISHING REPORTS ON PROGRESS IN PHYSICS

Rep. Prog. Phys. **68** (2005) 2701–2746 doi:10.1088/0034-4885/68/12/R01

Physical limits of silicon transistors and circuits

Robert W Keyes

IBM Research Division, Yorktown, NY 10598, USA

Received 27 April 2005, in final form 11 August 2005
Published 19 September 2005
Online at stacks.iop.org/RoPP/68/2701

Abstract

A discussion on transistors and electronic computing including some history introduces semiconductor devices and the motivation for miniaturization of transistors. The changing physics of field-effect transistors and ways to mitigate the deterioration in performance caused by the changes follows. The limits of transistors are tied to the requirements of the chips that carry them and the difficulties of fabricating very small structures. Some concluding remarks about transistors and limits are presented.

(Some figures in this article are in colour only in the electronic version)

0034-4885/05/122701+46$90.00 © 2005 IOP Publishing Ltd Printed in the UK

Contents

	Page
1. Introduction	2703
1.1. Advent of the transistor	2704
1.2. Electronic computation	2705
1.3. Integrated electronics	2705
2. Semiconductor devices	2708
2.1. The pn junction	2708
2.2. The transistor	2710
2.3. Computing with transistors	2712
3. Field-effect transistors	2713
3.1. Models of the FET	2713
3.2. Additional FET considerations	2715
3.3. The inverted layer	2715
4. Miniaturization of field-effect transistors	2716
4.1. Scaling	2717
4.2. Subthreshold current	2719
4.3. Miniaturized MOSFETs	2720
4.3.1. Short channel effects	2720
4.3.2. Drain-induced barrier lowering	2721
4.3.3. Gate tunnelling	2721
4.3.4. Drain currents	2723
4.3.5. Short channel experiments	2723
4.3.6. Mobility	2724
4.4. Novel FET designs	2725
5. The chip	2726
5.1. Wires	2726
5.2. Propagation of pulses on wires	2727
5.2.1. Space for wires	2728
5.3. Wire lengths	2729
5.4. Nonlocality	2731
5.5. Long wire delays	2731
5.6. Power on a chip	2733
5.7. The billion transistor chip	2735
6. Fabrication	2735
6.1. The lithographic process	2736
6.2. Processing limitations	2737
6.3. New transistor designs	2738
6.4. Defects	2739
6.5. Additional factors	2741
7. The transistor?	2741
8. Limits	2743
References	2744

1. Introduction

Readers may be surprised to find a paper on transistors in a physics journal. However, the transistor was invented by physicists and the subject of transistors is permeated with physics. The early studies of transistors were recognized as physics and published in physics journals, but a half-century of use of transistors in electronics has passed the subject into technology and engineering. Physics, however, continues to be an essential part of the continuing development of electronics by providing the basic understanding needed to deal with phenomena that gain importance and must be incorporated in models as the dimensions of transistors are decreased.

In return, by becoming the foundation on which major industries are based, transistors have given a powerful stimulus to the growth of solid state physics. The computational capabilities provided by transistorized computers have greatly expanded the range of techniques available to experimental science and have made whole new fields of physics possible. Many aspects of the manufacturing technology developed by the transistor industry have been adapted to experimental physics. The symbiosis of physics and technology, of basic science and semiconductor engineering, has enabled both to thrive.

The capacity of transistor electronics to reshape our world continues to astonish both observers of and participants in its development. The power to process information provided by electronics has given new abilities to and caused great changes in commerce, entertainment, warfare and communication, in addition to all aspects of engineering and science. Although information processing with machines had been practiced since the late nineteenth century, first with punch card machines and relays and later with vacuum tube computers, the invention of the transistor was the trigger that released the latent power of automated information processing and led to the vital role of electronic computing facilities in modern society. The demand for expansion of that role through lowering the cost and increasing the speed of information processing systems continues.

The rapid advancement of transistor electronics has relied on miniaturization, making transistors and their adjuncts smaller. Decade after decade of reducing the size of transistors naturally leads one to wonder how long that can go on. Thus the search for limits is a search for an answer to the question: 'how small can useful devices be made?' Seeking a limit to miniaturization is not a physics experiment, however, the quest must keep in mind 'useful' or 'why smaller and smaller?' Pushing transistor technology to its limits serves an economic purpose: it lowers the cost of transistors and of the computing power of the systems that use transistors.

The limits of silicon technology have long provided fuel for speculation among those with some acquaintance with the subject, both casual and serious. Regarding the former, the late Rolf Landauer is reported to have said 'Anyone can write $E = h/t$ on a napkin and persuade you over lunch that it imposes a fundamental limit on the power-delay product. The trouble begins when people publish their napkins' (Sandberg 1999). The more thoughtful assessments have been the subject of a good number of excellent reviews (Thompson *et al* 1998, Frank *et al* 2001, Meindl *et al* 2001, Plummer and Griffin 2001, Solomon *et al* 2002). These pertinent reviews are written by experts intimately involved with continued development of the technology, published in engineering journals, and necessarily lead into details that are most conveniently discussed with the specialized vocabulary that has grown up in the industry. The author does not here attempt to compete with these comprehensive and authoritative reviews, our aim will be to convey the essential physics of the topic for scientists not closely aligned with the development of the technology.

The early sections here are devoted to the history and development of transistor technology and the computer industry that supports it in order to clarify the environment that constrains and motivates the quest for miniaturization. Following that, semiconductor devices and transistors

are treated in a little detail with the aim of furnishing a background that will enable one with some acquaintance with solid state physics to follow the subsequent discussion of the changes in transistor characteristics as sizes are reduced. Miniaturization of transistors stresses the environment in which they are used and drives modifications of that. Thus, further sections discuss limitations imposed by the hardware that hosts transistors. Finally, the ability of the industry to manufacture ever-smaller devices in a way that meets the goal of reducing the cost of information processing must be considered.

1.1. Advent of the transistor

The telephone industry was a major user of relays and vacuum tubes in the first half of the twentieth century. Both were prone to failure and used large amounts of power. In 1946 the need for a better switch led the director of the Bell Telephone Laboratories, Mervin Kelly, to assemble a group of physicists charged with finding a solid state replacement for the vacuum tube and relay switches used in the Bell system (Brinkman *et al* 1997, Riordan and Hoddeson 1997). Germanium and silicon were chosen as working materials because they were readily available through their use as detectors of high frequency electromagnetic radiation during the war. The group was successful and transistor action was soon found in germanium (Bardeen and Brattain 1948, 1949). The novel phenomena that were discovered by the group opened an expanded window on to the solid state and the work was honoured by the award of the 1956 Nobel prize in physics. The promise of small size and high reliability from the newly discovered device immediately attracted widespread attention as an obvious candidate to replace the vacuum tube in many applications. Commercial adoption of the new invention was fast: transistor radios were brought into the market in 1954 and 100 000 were sold. Both vendors of vacuum tube computers and academic engineers were stimulated to explore computing with transistors, and Bell Labs built a 700 transistor computer for the Air Force in 1954 (Riordan and Hoddeson 1997).

The discovery of transistor action promoted the study of the solid state from a poor relation to a major part of physics. Reproducible transistor action depended on the use of good quality materials, leading to the development of methods of producing single crystals of high purity and major advances in material science. Research soon led to the discovery and exploitation of many other semiconductors, including the group III–group V compounds and additional novel effects. Efficient light emission and lasing action found in semiconductors added another dimension to semiconductor physics and spawned new industries. The photovoltaic effect was put to use as a source of electrical energy in solar cells.

Although transistor action was first discovered in germanium, transistors have since been demonstrated in many other semiconductors. Germanium persisted as the favoured transistor material until about 1960, when it became apparent that for a number of reasons silicon was preferable as a vehicle for manufacturing transistors. The oxide SiO_2 has such a prominent place in the list that it has been called 'nature's gift to the integrated circuit industry'. The oxide is easily formed by oxidation of silicon and is an excellent electrical insulator. The SiO_2 has a benign interface with silicon with a low concentration of interfacial electron traps and forms a tenacious chemically resistant coating on silicon that has been used to great advantage in the fabrication of devices. Silicon is a plentiful element at the surface of the earth.

Further, a number of other silicon compounds (metal silicides, silicon nitride) have favourable properties that aid the continued advance of silicon electronics. Silicon also has a desirable set of physical attributes for a room temperature transistor. Its energy gap of about 1.1 eV is large enough to allow a silicon chip to be heated to 100 ˚C without carriers thermally excited across the gap affecting transistor action. Yet the gap is not too large; in higher band

gap semiconductors acceptor and donor dopants are frequently so far from the band edges that they trap holes and electrons in immobile states at room temperature. (The effective masses of semiconductors tend in a rough way to increase with increasing energy gap and large masses lead to large binding energies of charge carriers to donors and acceptors (Kohn 1957).) The very high solubilities of both acceptor and donor atoms in silicon are advantageous in allowing low resistivity regions to be formed.

The dependence of computer technology on unique properties of silicon have earned that element the appellation 'new steel'. The global semiconductor industry is now an annual quarter-trillion euro behemoth that thrives on the continuing advancement of transistor technology.

1.2. Electronic computation

Computers use binary digital representation of information. Physical entities that represent binary information ideally must be in one of the two possible states, described as appropriate by 1 and 0, ON and OFF, or as a switch being CLOSED or OPEN. The earliest electrical computers used relays as switches. Computers only attracted commercial interest after vacuum tubes, much faster components than relays, were introduced to computing in the ENIAC, demonstrated in 1946. The ENIAC was the first successful general purpose computer, and was a large machine with 18 000 tubes that occupied a $10 \times 15\,\mathrm{m}^2$ room. The ENIAC spawned a small industry that manufactured vacuum tube computers. A decade after the transistor became known transistorized computers began to replace vacuum tube machines. The IBM 7000 series computers, announced in 1958, with up to 30 000 transistors were the first large machines to be based on transistors.

Switches are readily adapted to the electrical implementation of binary digital logic operations with zeros and ones. Switches are nonlinear, they are in either one position or the other. The early electrical computers were built with relays, electrically controlled switches and the electronic logic devices that succeeded them, vacuum tubes and transistors, emulate relays, turning the flow of electrons that represent information ON and OFF with electrical signals. The signals that exercise control in electronic computation do so by erection and removal of potential barriers. The voltages that create the barriers must be much larger than the thermal spread of electron energies to effectively simulate an on–off switch: $qV \gg kT$ or $V \gg kT/q$, 30 mV at 350 K. Small signals evoke only a linear response and do not suffice for digital operations.

The kT/q voltage scale can be used to find the order of magnitude of other electrical quantities involved in electrical logic. The impedance of simple structures is determined by geometrical factors times ϵ_0 and μ_0, the electrical and magnetic permittivities of space and is not too far removed from the so-called 'impedance of free space', $Z_0 = (\mu_0/\epsilon_0)^{1/2} = 377\,\Omega$. A basic measure of current in logic is therefore $kT/qZ_0 = 70\,\mu\mathrm{A}$. Similarly, the scale of power is $((kT/q)^2/Z_0) = 2\,\mu\mathrm{W}$ (Keyes 1969, 1975b). While such a semiquantitative argument leaves quite a bit of room to play in, electronics at 300 K cannot do fast logic with sub-microwatt power and does not need to command watts or kilowatts to power a logic gate. In fact voltage V of 10–100 kT/q is needed for an adequate approximation to on–off switching, while dielectric and geometric factors usually cause impedances to be only 0.1–0.3 Z_0 so the power is expected to be a few orders of magnitude larger than the basic value.

1.3. Integrated electronics

It soon became obvious that more devices permitted more computing power, but transistors in their first decade were limited by being manufactured and handled and wired one by one.

The integrated circuit, invented in 1960, by making it possible to make and handle many devices in a single operation opened a door to cost reduction in electronics and increasingly larger systems. Manufacturing integrated circuits was a planar process; devices and their interconnections could be made by performing operations on a single surface of silicon. The modern integrated circuit is a small (several mm^2 to several cm^2 in area) thin slice of silicon that contains many transistors and other components and is known as a chip (Reid 1984). The introduction of the 4004 microprocessor with 2300 transistors on a single chip in 1970 was a milestone in the story of the transistor in information processing; it represented more computing power than the ENIAC on a single chip.

The efforts of the information processing industry to satisfy an insatiable demand for more and more computing power per dollar continues to motivate increasing integration. Applications that would have once been impractical become reasonable. Scientific and engineering users seize greater power to increase the depth of detail by decreasing the mesh size in simulations and to attack larger problems and employ more physically realistic models. Sophisticated computer games provide complex actions with increasingly realistic figures. Consumers find convenience in cash registers that can read a bar code and look up a price and ATMs that check a customer's balance and count out money.

Striving to meet the demand for more computing power drives silicon technology towards ever-greater levels of integration, the number of devices that can be placed on a single chip of silicon. The cost of manufacturing a chip is only weakly dependent on the content of the chip, and more devices on a chip means less cost per device and it is devices that provide the computing power.

Chips are actually made by handling wafers, thin slices of a single crystal cylinder of silicon that are large enough to contain a few hundred chips into which they are eventually separated. The size of chips has grown during the era of integrated electronics and the size of wafers has grown to accommodate hundreds of chips. Using numbers typical of the present time, a 200 mm diameter wafer might contain 200 chips each of which holds 25 million transistors; the cost of making a wafer full of devices is shared by 5×10^9 transistors.

The need to produce devices at ever-lower cost also leads memory towards increased integration, multigigabit chips are entering the market. Faster, more powerful computers handle larger problems and need rapid access to increasingly larger collections of data. The measure of memory is simply cost per bit rather than the criteria applied to logic chips that include a measure of performance. The well-established magnetic core memories continued to be used in computers for a decade after transistors were introduced for logic. The demonstration of a 1000 bit memory chip in 1970 marked the beginning of cost dominance by semiconductor memory. Since then supplying the demand for more data storage with semiconductor memory through higher and higher levels of integration has been another major thrust of the integrated circuit industry.

The phrase 'Moore's Law' has been rendered threadbare through incessant repetition in the popular and the technical press. In 1965, Moore (1965), a pioneer in the silicon industry as one of the founders of Intel observed that for some time the silicon electronics industry had doubled the number of devices on a chip each year. Moore (1978) went on to attribute the steady growth in the number of devices on a chip to three factors: smaller devices that allow more devices per unit area, larger chips and 'cleverness' or 'compaction'. The last factor refers to making better use of available space on a chip by reducing the area per device by innovations in processing and device design rather than miniaturization. For example, a capacitor might be made by depositing a conductive layer, an insulating film and another conductor on the chip surface. Space can be saved by doing this with several layers, one atop another. Still less space is used by etching a deep hole in the substrate and depositing layers on the surface of

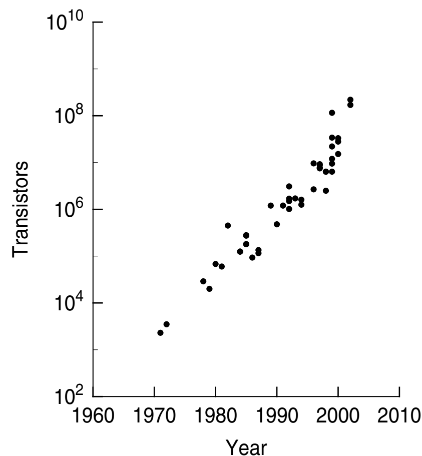

Figure 1. The increasing number of transistors on microprocessor chips.

the hole. Figure 1 shows the results of these inputs as the growing number of transistors on a microprocessor chip since the Intel 4004 in 1970.

The transistor must be considered in the context of the chip. A variety of aspects of a many-transistor chip must be adapted to an increasing chip content. Devices must be connected into functioning circuits by wires; the larger the number of devices, the more the amount of wire needed. Electrical power must be supplied to the circuits and means for removing the heat produced by the power are needed. The chip must be connected to other components which supply the power and which can exchange information-bearing signals with the chip. Smaller, more densely packed devices rather obviously stress the technology that provides these other functions.

Integrated electronics today fabricates the transistors, wires and other components that appear on a chip by subjecting wafers to a long series of process steps in large factories. Patterns of materials are deposited and/or removed in specific areas. Optical (broadly interpreted) radiation passing through masks selects the areas to be affected at each step. The series of operations and the tools that control them must be frequently modified to accommodate smaller features and greater component densities.

By the end of the 1980s the industry view of miniaturization of electronics matured and changed. More than a decade of constant progress towards smaller structures and more devices per chip created the confidence that the development of even more powerful, lower cost electronics could be continued into an indefinite future. Circuits that used field-effect transistors (FETs) were replacing the high power bipolar circuits that had dominated high speed computing. Microprocessors were becoming a major product of the semiconductor industry and playing an increasing part in computing hardware. The evolving device requirements and continued progress towards reduced dimensions and higher levels of integration required new generations of equipment to deal with obstacles that were appearing in the road to continued miniaturization. The dimensions of structures on a chip were falling below the wavelength of visible light—new light sources with shorter wave length were needed. Monochromatic light sources with lenses designed to minimize aberrations and maximize resolution for lithography became essential (Singh *et al* 1997). A new light source meant other new optical equipment with new lenses, and new photoresist materials sensitive to the shorter wavelength had to be found. Shorter times and lower temperatures were required to reduce the distances that

dopants diffused during processing. Smaller defects became important, demanding more rigorous standards of cleanliness. At the same time technology allowed increased chip sizes, helping to increase the number of devices per chip, but calling for larger wafers. Larger wafers meant new apparatus for the growth of larger crystals and for new processing tools that could handle the larger wafers. The transitions to smaller device dimensions and larger substrates pervaded all aspects of a manufacturing facility.

The new tooling requirements had to be communicated to the equipment manufacturers and material suppliers at an early date. The Semiconductor Industry Association was formed to make periodic assessments of the state of the semiconductor art and present quantitative visions of future semiconductor technology. The studies were made publicly available as 'Roadmaps' for the guidance of the suppliers of materials, supplies and semiconductor manufacturing equipment, of the semiconductor industry and of pre-competitive research efforts in university and government laboratories. The forecasts continue today as the International Technology Roadmaps for Semiconductors (ITRS)—international efforts based on inputs from hundreds of volunteers drawn from several national semiconductor associations. The Roadmaps transcend simple extrapolation to examine possible technological directions through the coming fifteen years in detail and identify barriers to progress and potential avenues of research for avoiding them as well as problems for which no solution can be seen at the present time (ITRS 2001, 2003).

However, the history of the integrated circuit industry is one of the discovery of unexpected difficulties and solving them by research and development efforts. A pessimistic view would be that the 'no known solution' barriers seen in the Roadmaps announce a set of limits with dates at which the limits would be felt. However, the view of the Roadmaps and of the institutions that support them is that these are challenges that will be mastered with new ideas and inventions. The Roadmaps have been published for over a decade, and usually technology has advanced more rapidly than anticipated by the Roadmaps.

2. Semiconductor devices

The action of transistors and diodes depends on the existence in semiconductors of mobile charges, both electrons in the conduction band and holes, unfilled states in the valence band (Sze 1981, Taur and Ning 1998, Hess 2000, Seeger 2002). The electrons are introduced by the addition of donor elements, those with one more valence electron than the host atom that it replaces, as phosphorus in silicon, and the holes by acceptor elements, usually boron, from Group III of the periodic table, with one less electron replacing a host atom. Transistors have p-type regions containing acceptors and holes and n-type regions doped with donors and containing electrons, both also containing small numbers of minority carriers of the opposite sign.

2.1. The pn junction

The contact between n and p regions is a pn junction and its properties are important in transistor action. Figure 2 shows how the Fermi level and the band edges are related in a pn junction. The Fermi level is constant throughout and the conduction band edge is near it in the electron-containing region and near the valence band in the p region. The Fermi level in the zone between the n and p regions is far from either band edge, meaning that there are few electrons and few holes present. Dopant concentrations are conveniently referred to the intrinsic concentration, n_i, the concentration of holes and electrons that would be present by being thermally excited from the valence band to the conduction band in the absence of

impurities:

$$n_i^2 = C\exp(-E_g/kT). \tag{2.1}$$

The constant C depends on the densities of states near the band edges and E_g is the gap in energy between the valence band and the conduction band. The intrinsic concentration of silicon at 300 K is $n_i = 1.45 \times 10^{10}$ cm^{-3}. The Fermi level corresponding to the intrinsic case is designated by F_i.

The variation of the potential between the n and p semiconductors is treated in an excellent approximation (based on the rapid decrease of carrier concentrations as a band edge moves away from the Fermi level) by assuming that the region between the two is entirely depleted of mobile charge carriers and thus is determined by the charge of the donor and acceptor atoms (Shockley 1949). The potential, ϕ, in the depleted region varies in accordance with the Poisson equation

$$\frac{d^2\phi}{dx^2} = \frac{-N_D q}{\epsilon} \tag{2.2}$$

in the region dominated by donors. Here N_D is the net concentration of positively charged donor atoms and a similar equation holds when acceptors, concentration N_A, dominate. Also ϵ is the electric permittivity of the silicon.

Also, although the donor and acceptor doping atoms bind the added carriers at low temperatures, essentially all are ionized at 300 K in silicon and the electron and hole concentrations are equal to the numbers of donor and acceptor atoms in, respectively, n-type and p-type material. Then the number of donor atoms, N_D, equal to the number of electrons, n, in equilibrium may be expressed as $n = N_D = n_i\exp((F_c - F_i)/kT)$, where F_c is the Fermi level in n-type material. A similar equation applies in p-type semiconductor: $p = N_A = n_i\exp((F_i - F_v)/kT)$, where N_A is the concentration of acceptor atoms and F_v is the Fermi level. Holes and electrons can recombine to reach an equilibrium state in which the Fermi level controls the concentrations of both carriers and the minority carrier concentrations can be determined from the reaction np $= n_i^2$. In a p-type semiconductor $n = n_i^2/p = n_i/N_A$ and similarly in n-type material $p = n_i^2/N_D$. The magnitude of the change in ϕ from the conduction band to the valence band is known as the built-in voltage of a junction, shown in figure 2, and is determined by the energy gap of the semiconductor and the doping by donors and acceptors. The built-in voltage V_{bi} is $F_c - F_v$, the displacement of the energy bands that aligns the Fermi levels across the junction (figure 2) and can also be written as

$$V_{bi} = kT\log\left(\frac{N_D N_A}{n_i^2}\right), \tag{2.3}$$

by using the preceding expressions for N_A and N_D. In reasonably conductive materials where the Fermi level in each type is close to the respective band edge V_{bi} approaches the energy gap.

The application of a voltage to the junction shown in figure 2 creates a difference in the position of the Fermi level between the n and p regions. A positive voltage V applied to the p region, forward bias, lowers the barrier between the two sides of the junction by an amount qV and attracts electrons into that region and forces holes in the opposite direction. The number of carriers that can pass over the barrier is thermally activated, the current is proportional to ($\exp^{qV/kT} - 1$), vanishing when $V = 0$. A voltage in the reverse direction, of opposite sign, raises the barrier to a current of majority carriers in each direction. A small current which is still governed by ($\exp^{qV/kT} - 1) < 0$ flows in the opposite direction and saturates rapidly with increasing reverse voltage, $V < 0$.

There are frequent interesting cases in which the doping on one side of the junction is much larger than on the other side, the one-sided junction case. Let the p side of the junction,

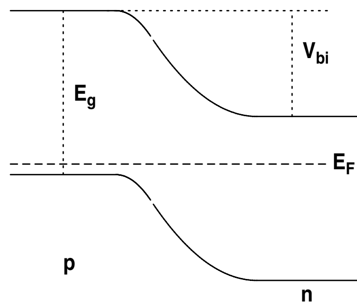

Figure 2. The variation of the band energies through a pn junction in the absence of an applied voltage. V_{bi} is known as the 'built-in voltage'. The p side of the junction is more heavily doped than the n side.

doping N_A, be the lightly doped side. Figure 2 is drawn with asymmetric doping to suggest the nature of this case. The more rapid variation of the potential on the heavily doped side means that almost all of the depleted layer lies on the lightly doped side of the junction where the potential varies as

$$\phi = \left(\frac{N_A q}{\epsilon}\right)(x^2 - 2xw_D). \tag{2.4}$$

The extent of the depleted space is w_D and since most of the potential drop is on the lightly doped side the built-in voltage becomes nearly $V_{bi} = \phi$ (equation (2.4)) and

$$w_D = \left(\frac{2\epsilon V_{bi}}{N_A q}\right)^{1/2}. \tag{2.5}$$

The widening of the depleted layer under reverse bias is another aspect of the pn junction. The extent of the layer is found by substituting $(V + V_{bi})$ for V_{bi} in equation (2.5). As the reverse voltage applied increases the width of the layer increases, but only as the square root of the voltage (equation (2.4)) and the electric field in the layer increases. At sufficiently high reverse bias the electric field in the depleted region can accelerate charge carriers to the point at which they can lose energy by exciting an electron from the valence band into the conduction band, creating a free electron and a hole. These again gain energy from the field and may excite more charge carriers and create an avalanche of current, dielectric breakdown. The field needed to cause dielectric breakdown in silicon is approximately 4×10^5 V cm^{-1} and limits the voltage that can be used in semiconductor devices.

Heavy doping thins the depleted layer of the junction. Both sides of the junction can be doped so heavily that the depleted layer is reduced to the point at which holes and electrons can pass through it by tunnelling. When biased in the forward direction the tunnelling currents can produce a negative resistance, a device known as the Esaki diode or tunnel diode (Esaki 1958), which has found applications, though not in logic circuits (section 7). Tunnelling currents play no role in transistor action and are not welcome in logic with transistors.

2.2. *The transistor*

The principle of the transistor is simple (Sze 1981, Taur and Ning 1998, Hess 2000, Seeger 2002). Potential barriers hold charge carriers of one sign in place while charge neutrality in the

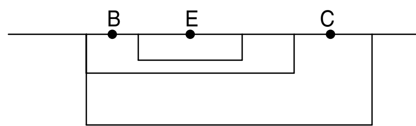

Figure 3. A planar npn bipolar transistor showing contacts to Emitter, Base and Collector.

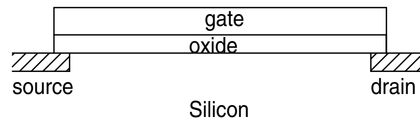

Figure 4. Schematic drawing of a FET.

device is maintained by mobile carriers of the opposite sign. The mobile charge can then carry currents under the influence of electric fields and concentration gradients. Large quantities of electric charge can be carried through the transistor by the mobile charge with only a minuscule current needed to replace any leakage of the fixed charge.

The first transistors were of the bipolar type. The two signs of charge in semiconductors permit the construction of two kinds of transistors, npn, where electrons are the mobile carriers, and pnp. The earliest transistors used point contacts to inject and to collect charges but modern bipolar transistors, like the planar transistor illustrated in figure 3, use pn junctions for the same purpose. Potential barriers formed by pn junctions confine holes in an npn transistor and mobile electrons occupy the same space as the confined holes. Biasing the base of an npn transistor positively with respect to the emitter draws electrons into the base where their charge is compensated by holes supplied through the base contact. The electrons diffuse through a concentration gradient from the emitter across the base to the collector where they are captured by a positively (reverse) biased collector junction, while escape of the holes is prevented by the junction barriers. The concentration of donors in the emitter is made much larger than that of acceptors in the base to make the emitter current mostly an electron current. Electron current can continue to flow from emitter to collector while only a small base current is needed to replace holes lost to the base through the emitter current and by recombination of holes and electrons in the base. Bipolar transistor action depends on the hole–electron recombination lifetime being much longer than the time needed for the electrons to cross the base and enter the collector.

A very thin insulating layer separates the fixed and mobile charges in the FET, illustrated in figure 4. Again there are two kinds of FET, designated as n-type and p-type depending on whether electrons or holes are the mobile charge carriers. A layer of charge attracted to the surface by a positive voltage applied to the gate in an n-type FET forms a connection between the source and drain regions, mimicking the action of a mechanical relay. The advent of commercially useful FETs followed that of bipolar transistors by more than a decade because of the formation of a large concentration of charge trapping states at the interface between the semiconductor and the insulator. Charge on the gate would induce charge in the traps rather than mobile carriers in the semiconductor. Reproducible production of high-gain FETs had to await the development of a refined silicon-SiO_2 technology which could produce interfaces between the silicon and the oxide insulator with a low trap density. The difficulty persists in field-effect devices in material systems other than silicon-SiO_2. The silicon FET dominates contemporary electronics and is the focus of current concern with limits.

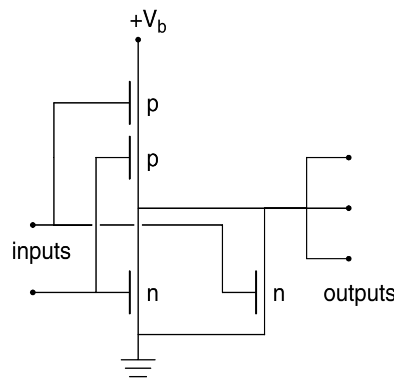

Figure 5. A CMOS NOR circuit. If either input is positive the outputs are connected to ground through the n transistors. When both inputs are low the p transistors are turned on and connect the output to V_D.

The widespread use of silicon-SiO_2 field-effect transistors has earned them their own name: MOSFET. The acronym MOS means metal–oxide–semiconductor with 'metal' referring to the gate. The gate in early FETs was made of the metal aluminium and the name persists even though the gate material today is more commonly heavily doped polycrystalline silicon.

The development of CMOS (complementary MOS) circuits greatly reduced the use of power in electronic switching circuits. CMOS combines n-type and p-type FETs in a circuit in which there are two stable states. Figure 5 is an example of a CMOS circuit. Current flows only when the circuit is switching from one of the states to the other. The low power demand makes CMOS the circuit of choice for miniaturized integrated electronics.

Making both npn and pnp transistors on the same silicon surface poses a problem as the bulk semiconductor should be of opposite type in the two cases. This is managed in a p-type substrate by fabricating the p-FETs in wells doped with donor atoms.

2.3. Computing with transistors

The bipolar transistors in early (before about 1985) large computers controlled large currents to achieve high speed. The density of transistors on a chip increased inversely as the square of the length variable with miniaturization while the current in a transistor remained nearly constant as dimensions were reduced (Keyes 1988). Voltages decreased only slowly if at all, so the density of power dissipation increased rapidly. Removing the heat that was produced became a serious problem as devices were made smaller and levels of integration increased. The high power and large currents limited the levels of integration of bipolar chips to values much less than those that the manufacturing technology allowed, and some large machines adopted water cooling to remove the high power densities of arrays of bipolar circuits (Blodgett 1983).

The capability of FET-based microprocessors after their introduction with the Intel 4004 increased rapidly. The improving performance of microprocessor chips that used FETs forecast the end of the dominance of bipolar transistors in large computer systems. FET circuits became competitive with bipolar logic and, aided by the invention of low power CMOS, FETs dominated large processors by 1990 and systems engineers learned how to build large supercomputers from many microprocessors.

3. Field-effect transistors

3.1. Models of the FET

First an oversimplified model is used to give a qualitative idea of the current–voltage characteristics of an FET. Referring to figure 4, an n-type FET, let $c_i = \epsilon_i/t$ be the capacitance per unit area of the insulator that separates the gate from the semiconductor, where ϵ_i is the dielectric constant of the insulator and t is its thickness. A positive voltage applied to the gate attracts electrons to the interface between semiconductor and insulator. A minimum gate voltage V_T, the threshold voltage, is required to cause electrons to appear at the interface. The charge per unit area in the electron channel at the interface is then

$$\rho_s = c_i((V_G - V_T) - V). \tag{3.1}$$

Here V is the voltage in the channel, which is connected at one end to the source contact and at the other end to the drain. Considering the source as the zero of potential, a current flows when there are electrons in the channel and a positive voltage is applied to the gate. The current is driven by the electric field parallel to the surface, $\mathrm{d}V/\mathrm{d}x$, and for a transistor of width w in the direction normal to the current is

$$i = \rho_s v w = c_i((V_G - V_T) - V)\mu\left(\frac{\mathrm{d}V}{\mathrm{d}x}\right)w, \tag{3.2}$$

where v is electron velocity, μ is electron mobility and x is distance from the source. Equation (3.2) is readily integrated to give

$$ix = \mu w c_i\left[(V_G - V_T)V - \left(\frac{V^2}{2}\right)\right]. \tag{3.3}$$

The current through the transistor is found when $x = L$, the distance from source to drain and $V = V_D$ is the voltage applied to the drain.

$$i = \left(\frac{\mu w c_i}{L}\right)\left[(V_G - V_T)V_D - \frac{V_D^2}{2}\right]. \tag{3.4}$$

Equation (3.4) as a function of V_D has a maximum when $V_D = V_G - V_T \equiv V_S$. When V_D is increased beyond the saturation voltage V_S the current remains constant, and it is said to be saturated.

When the drain voltage is raised above V_S no charge remains at the drain to carry the current (equation (3.1)). The channel is said to be 'pinched off'. In fact, this does not happen as the carrier velocity saturates at high fields and a few carriers remain to carry the current under the influence of a very high electric field between the drain and the point of pinch-off. The field near the drain can become comparable to the fields in the depleted layers and the gradual channel approximation fails.

The threshold plays an important part in a discussion of limits. Figure 6 shows the variation of the energy bands of a semiconductor along a vertical line from the gate electrode through the insulator into the body of the p-type semiconductor. In these plots E_c and E_v are energies of the conduction and valence bands and E_F is the Fermi energy. The Fermi levels of the gate and the semiconductor are aligned in figure 6(*a*), no voltage is applied to the device. Figure 6(*b*) shows the 'flat-band' condition, i.e. a voltage that produces a vanishing electric field in the bulk semiconductor has been applied to the gate. The flat-band condition is a convenient starting point for a discussion of thresholds. In practice, the voltage necessary to achieve the flat-band condition must be added to the threshold. In figure 6(*c*) a positive voltage that lowers the energy of the conduction band to the point at which the distance of the conduction band above the bulk Fermi level is equal to the energy distance between the Fermi level and the

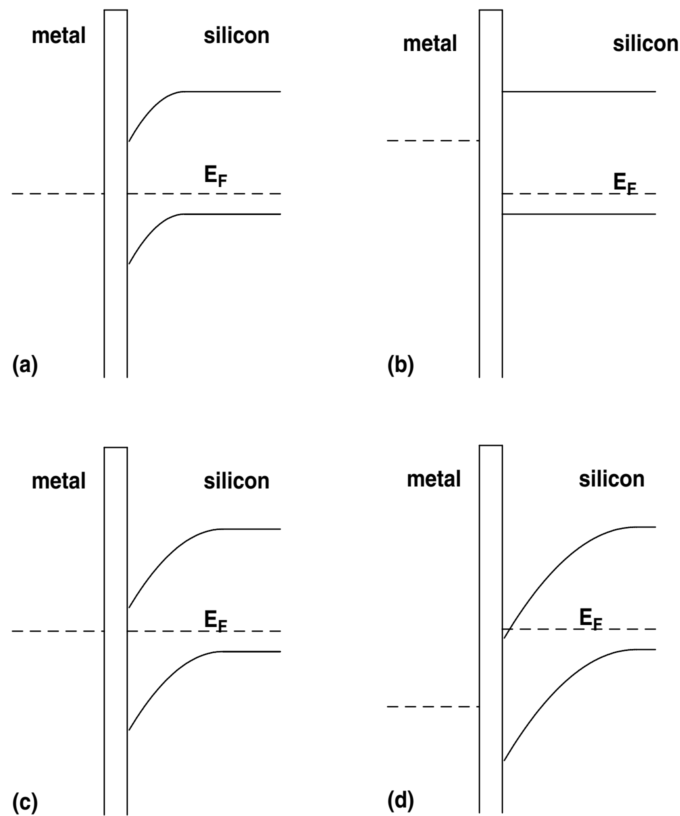

Figure 6. Energy bands at the surface of an n-type FET. (*a*) No voltage is applied to the metal gate. (*b*) Flat-band. A gate voltage has removed the variability of potential within the silicon. (*c*) Threshold, by convention defined as the point at which the potential difference between the bottom of the conduction band at the surface and the Fermi level is equal to the difference between the valence band in the bulk silicon. (*d*) Deep inversion, the Fermi level has entered the conduction band creating a thin layer of a degenerate electron gas at the surface.

valence band in the p-type substrate has been applied to the gate. The gate voltage at this point is by convention called the threshold voltage. At this threshold the concentration of electrons at the oxide–silicon interface is approximately equal to the concentration of holes in the substrate. Figure 6(*d*) shows the beginning of deep inversion, when the Fermi energy enters the conduction band.

The solution of Poisson's equation in the depletion approximation in the substrate where the charge is dominated by acceptor atoms is

$$\phi = -\psi + \left(\frac{N_A q}{\epsilon}\right)\left(\frac{wx - x^2}{2}\right). \tag{3.5}$$

Here x is the distance into the substrate from its contact with the insulator, N_A is the concentration of acceptor atoms in the depleted region, ϕ is the change in potential measured from the p-type interior of the substrate, ψ is the change in potential in the semiconductor from its contact with the insulator to the bulk and w is the width of the depleted layer. The boundary conditions are $\phi = -\psi$ at $x = 0$, $\phi = 0$ at $x = w$ and $w = \sqrt{2\psi\epsilon/N_A q}$ is the width of the depleted region.

At the threshold as defined above $q\psi_t = E_G - 2(E_F - E_v)$, where $E_G = E_c - E_v$ is the energy gap. The electric field in the insulator at this point is determined by $\epsilon_{ox} F_{ox} = \epsilon w(N_A q/\epsilon)$ when the field in the semiconductor is found by differentiating equation (3.5). Substituting the value of w gives $F_{ox} = (2N_A q\psi_t/\epsilon)^{1/2}/\epsilon_{ox}$. Adding the potential drop in the insulator to that in the semiconductor gives

$$V_T = \psi_t + \left(\frac{t_{ox}}{\epsilon_{ox}}\right)\left(\frac{2N_A q\psi_t}{\epsilon}\right)^{1/2}, \tag{3.6}$$

as the voltage on the gate electrode at threshold.

Increasing the gate voltage above threshold attracts more electrons to the channel. The added charge on the gate is compensated by the charge in the channel and there is little change in the depleted layer. The charge in the depleted region is $Q_d = N_A w$ and V_T can also be written as $V_T = \psi_t + Q_d/c_{ox}$, where $c_{ox} = \epsilon_{ox}/t_{ox}$ is the capacitance of the insulator per unit area.

3.2. Additional FET considerations

This simple description hides various aspects of the FET which are discussed in detail in (Sze 1981, Taur and Ning 1998), which deserve brief mention.

The mobility in equations (3.2)–(3.4) is not a simple constant as assumed above but is influenced by several aspects of the conditions of operation. The mobility is less than the mobility in the bulk semiconductor because mobile carriers are scattered by the semiconductor–insulator interface. The closer the carriers are to the interface, the stronger the scattering, so that the mobility is less, the greater the transverse electric field that holds the electrons near the interface.

At high electric fields electrons and holes gain kinetic energy from the electric field so rapidly that their effective temperature rises above that of the solid silicon. The charged particle's interactions with the lattice change, and the electrons approach a limiting velocity as the field increases. At high enough fields a few of the 'hot' electrons gain enough energy to escape over the barrier that holds them in the semiconductor and escape into the oxide, where they become trapped and where their charge changes the threshold voltage of the transistor. The escape of hot electrons from the silicon to the traps in the oxide is one of the limits of the design and use of field-effect transistors.

If the source-to-drain distance becomes less than the mean free path of the electrons, charge carriers may traverse the distance without being scattered, a regime known as 'ballistic' transport.

3.3. The inverted layer

An effect on the threshold of an FET discovered by Fowler *et al* (1966) in MOSFETs in deep inversion (figure 6(*d*)) can be attributed to 'particle in a box' quantization in the narrow channel (Stern and Howard 1967). FETs in a state intended to carry current are operated in the deep inversion regime in which the Fermi level in the layer is above the bottom of the conduction band. The electrons are a statistically degenerate gas and a rough idea of the nature of the inversion layer can be obtained from a drastically simplified statistical view of the layer (a two-dimensional Thomas–Fermi model). Consider a layer of electrons on a p-type surface. In deep inversion the concentration of electrons is much larger than the concentration of acceptor atoms and the latter will be neglected. The electron gas can be considered degenerate. Then

the concentration of electrons at any point is

$$n = \left(\frac{8\pi}{3h^3}\right)(2me(E_F - \phi))^{3/2}. \tag{3.7}$$

Poisson's equation must also be satisfied:

$$\frac{\epsilon d^2\phi}{dx^2} = \left(\frac{8\pi e}{3h^3}\right) = n\epsilon = (2me(E_F - \phi))^{3/2}. \tag{3.8}$$

Equation (3.8) has an analytic solution:

$$(E_F - \phi) = \frac{B}{(x+a)^4}. \tag{3.9}$$

Here B is

$$B = \frac{400(3h^3\epsilon/8\pi e)^2}{(2me)^3}. \tag{3.10}$$

The parameter a can be determined from F_S, the electric field in the semiconductor at the surface or from N_S, the total number of electrons in the surface layer.

$$N_S = \frac{\epsilon F_S}{e} = \frac{4B\epsilon}{ea^5}. \tag{3.11}$$

The electron density in the layer is given by

$$n = \frac{5N_S a^5}{(x+a)^6}. \tag{3.12}$$

For a field of 3×10^5 V cm^{-1} in silicon at the interface, for example, a is about 4 nm and the electron concentration in the channel is 2×10^{12} cm^{-2}; the electrons are confined within a few nanometres of the surface. The lowest energy state of an electron in a 4 nm box in silicon is several tens of millivolts above the bottom of its well and the quantization has a significant effect on the threshold voltage of an FET (Stern and Howard 1967, Stern 1972, Tang *et al* 2004).

Much has been left out of the above cursory treatment. An effect of importance in miniaturized transistor arises from the multivalley nature of the silicon band structure (Kohn 1957). Different contributions to the conduction band and to the valence band are affected differently by the narrow channel confinement and participate differently in the channel wave functions. The splitting is dependent on crystalline orientation and is different for different crystallographic orientations of the silicon surface (Fischetti *et al* 2003). The lowest energy band and the mobility of a charge carrier therefore varies with the orientation of the surface.

4. Miniaturization of field-effect transistors

Two forces propel the miniaturization of MOSFETs. Besides the obvious increase in the number of devices per chip, miniaturization allows devices to switch faster. Decreasing distances allows electrons and holes to pass through devices in shorter times and devices are closer together, reducing device–device travel time. Capacitance is dimensionally equivalent to length; smaller devices have less capacitance to charge. Performance, measured as speed of operation, is improved.

The state of miniaturization is described by a measure of the minimum dimension which the fabrication technology can define on a chip. The dimensional feature that is used as the measure of miniaturization is somewhat arbitrary and may be called such names as 'rule' and 'technology'. The Semiconductor Industry Association for purposes of its periodic Roadmaps (ITRS 2001, 2003) uses the measure half of the pitch (the line-to-line distance) of the finest conductors on a chip and calls it 'technology node'.

Table 1. Constant field scaling.

Quantity	Symbol	Scaling factor
Dimension	L	s
Voltage	V	s
Concentration	n	$1/s$
Electric field $= V/L$	E	1
Current density $ne\mu E$	j	$1/s$
Current $= jL^2$	i	s
Capacitance $\sim L$	C	s
Resistance $\sim 1/L$	R	$1/s$
Voltage in resistance	iR	1
Power $= iV$	P	s^2
Power/area $= iV/L^2$	Q	1
Time $= CV/i$	t	s
Energy $= Pt$	W	s^3
Number of dopant atoms $= nL^3$	m	s^2

4.1. Scaling

In a first approximation miniaturization reduces all dimensions of a device by the same factor in order to retain the model of the physics of the device. For that reason miniaturization is also called 'scaling' and rules have been promulgated to guide scaling. An early influential set of rules for reducing dimensions by a factor s is given in table 1 (Dennard *et al* 1974). Voltages are reduced in proportion to the dimensions in this scenario in order to keep electric fields constant, giving it the name 'constant field scaling'. Limiting the electric field prevents electrons from gaining enough energy to exhibit the hot electron effects mentioned above. Hot electrons may acquire enough energy to surmount barriers, even enough to escape from the channel of an FET into the gate insulator where they can become trapped and where their charge will alter the transistor characteristics or cause dielectric breakdown as described in section 2.1.

Scaling of dimensions requires that the thickness of depleted layers between n- and p-type material be reduced in the same way as other dimensions. Table 1 addresses the thinning of depleted layers: since the width of a depleted layer is proportional to the square root of the voltage supported by the layer divided by the doping concentration (equation (2.4)), increasing the doping level by a factor $1/s$ in addition to reducing the voltage reduces the depleted layer thickness by the factor s. The physics then controls the scaling of many other electrical characteristics of the scaled device, as shown in the table.

Table 1 illustrates how constant field scaling advances the goals of miniaturization: increased speed, the time scale is reduced by factor s, and lower power, the power per device is reduced by s^2. Of course, the density of devices on a chip is also increased by factor s^2, but the power per unit area is constant and the energy dissipated per switching event decreases by a factor s^3.

Scaling forms a convenient starting point for a discussion of miniaturization, but technology, for a variety of reasons, has found it difficult to strictly follow the scaling rules of table 1. There are certain fixed voltages that are impossible or extremely difficult to change. First is the built-in voltage which is important in determining the thickness of depleted layers, but which cannot be decreased much below the semiconductor energy gap, as seen in the p–n junction (figure 2). Thermal energy, kT, that prevents an abrupt effect of voltage barriers on electron currents is fixed by the environment and is not scaled. To change the number

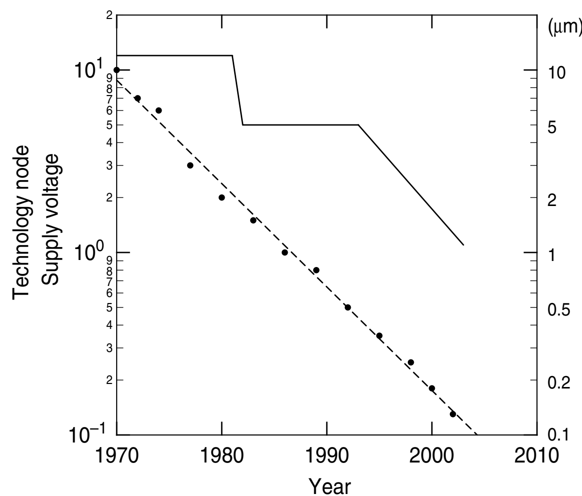

Figure 7. The changes in widely used power supply voltages and the minimum dimension on chips since the introduction of the microprocessor in 1970. The dots representing dimension refer to the right-hand scale.

of electrons that pass a potential barrier by orders of magnitude, as required of a switch, the height of the barrier must be changed by a voltage large compared with kT/q. Also, fabrication processes leave appreciable uncertainty in the parameters of devices that must be accounted for in circuits (Orshansky 2000).

The randomness in the position of dopant atoms in semiconductors is an inherent source of uncertainty in small devices. Table 1 shows that the number of atoms in any particular region of a device, m in the table, scales as s^2, decreasing rapidly as dimensions are reduced. The random fluctuation in m is of order $m^{1/2}$, becoming a larger fraction of m in smaller devices. This effect in the depleted layer under the channel causes fluctuations from device to device in the threshold voltage of FETs (Hoeneisen and Mead 1972, Keyes 1975a, 1994, De *et al* 1996). The smaller number of dopants exposes transistors to the effects of fluctuations in their number which may also have to be taken into account in signalling voltages. Voltages must overwhelm both the uncertainties of fabrication and the thermal distribution of energies and cannot in practice always be reduced as desired.

The voltage drop in the ohmic resistance inherent in devices, for example in the source and drain in figure 4, scales as a constant. Since the voltages involved in device action are reduced with dimensions the resistive drops loom larger.

For these and other reasons the full reduction in voltage suggested by table 1 has not occurred. Reducing voltage was retarded by being tied to industry-wide standards for many years. Scaling according to the rules of table 1 decreases the device current, an eventuality to be avoided to the extent that it decreases speed of response or if current also flows to entities that are not scaled, such as long wires on a chip. Circuit and device designs must respond to a business quest for greater speed, an objective that is advanced by setting voltage and the resulting currents higher than prescribed by scaling. Figure 7 shows how the most common voltage used in computer circuitry has varied with time and compares it with the progress of miniaturization as measured by the minimum dimension of features on chips.

Certain length parameters characterize silicon and are not controlled by the manufacturing process and cannot be scaled. Selected length parameters of n-type silicon at 300 K are plotted

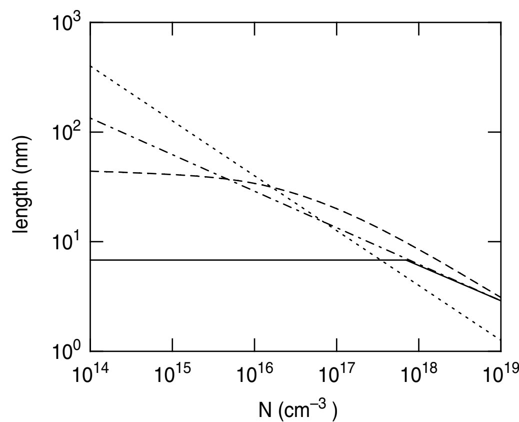

Figure 8. Some material parameters of n-silicon as a function of the donor concentration at 300 K. The dashed line is the electron mean free path, the dotted line is the screening radius, the dot–dash line is the average distance between donor atoms, and the solid line is the electron wavelength, determined by the temperature at low concentrations but decreasing as the electron gas becomes degenerate.

as a function of donor concentration in figure 8. The approach of the dimensions of transistors to inherent material length parameters does not mean that the devices will not work, but it does mean that new physical effects may become important and new device models may be needed. For example, device dimensions near or below the electron mean free path mean that an electron might traverse that dimension without being scattered, its motion is not controlled by a mobility parameter, and when a dimension approaches the electron wavelength quantum effects must be considered. Although hard to depict in figure 8, tunnelling probabilities increase with decreasing dimension, as will be seen later.

Therefore, as devices are miniaturized new physical effects that must be taken into account appear and models must be enlarged to include them. The availability of powerful computers permits the inclusion of complex effects in detailed numerical simulations of device operation to replace analytic models in exploring device designs. A simulation program follows the progress of an electron through a numerical model of a device structure with many small time steps. The probability of each possible kind of interaction (for example, acceleration by a field, scattering by an impurity, interaction with a phonon, recombination with a hole, loss of energy to the creation of an electron–hole pair) that may effect an electron trajectory are obtained from the basic physics of the phenomena involved. At each time step a random number generator selects an event (or lack thereof) from the known probabilities. Following many electrons (typically thousands) through a structure gives a picture of device action (Price 1978, Jacoboni and Reggiani 1983, Laux *et al* 1990). The probabilistic content of such simulations has earned them the name 'Monte Carlo' programs.

4.2. Subthreshold current

Transistors in digital computers are used as switches in computing and face a limit when they are unable to control current. No current should flow to the drain when a transistor is turned off by a voltage applied to the gate. However, an FET is never completely 'off'. As noted earlier, hole and electron concentrations reach an equilibrium described by $np = n_i^2$. As the gate voltage moves the conduction band upwards in energy and farther from the Fermi level to turn

the transistor off the electron concentration decreases exponentially but does not vanish. The few remaining electrons under the gate carry a small 'subthreshold current'. The subthreshold current is usually characterized by a subthreshold slope, which measures the increase in gate voltage needed to reduce the subthreshold current by a factor of 10. An expression for the subthreshold slope is (Taur and Ning 1998)

$$S = \left[\frac{\mathrm{d}(\log_{10} I_{ds})}{\mathrm{d}V_g}\right] = \left(\frac{kT \log_{10}}{q}\right)\left(1 + \left(\frac{C_\mathrm{d}}{C_\mathrm{ox}}\right)\right). \tag{4.1}$$

Here $C_\mathrm{ox} = \epsilon_\mathrm{ox}/t_\mathrm{ox}$ is the capacitance per unit area of the insulator and $C_\mathrm{d} = \sqrt{Nq\epsilon/(2V_\mathrm{d})}$ is the corresponding capacitance of the depleted layer in the semiconductor. V_d is the voltage drop in the depleted layer and is close to the built-in voltage of the source–substrate junction. The first term in the parentheses in equation (2.1) represents the decrease in the number of thermally excited carriers when the carrier concentration decreases in the way described and has the value $2.3(kT/q) = 60\,\mathrm{mV}$ at 300 K. The second term accounts for the division of the gate voltage between the series-connected gate capacitance and the capacitance of the depleted layer and vanishes if $C_\mathrm{ox} \gg C_\mathrm{d}$. As other voltages are reduced in accordance with table 1 the subthreshold slope is nearly constant and becomes more significant.

The capacitance C_ox really represents the capacitance between the gate electrode and the electrons in the channel and is diminished by two additional factors that gain in importance as dimensions are reduced. The quantization of the electronic states in the channel removes the electrons from the interface with the oxide (Stern 1972), and the applied voltage causes a small depletion of the heavily doped silicon which forms the gate at the oxide interface. Each of these contributes roughly 0.5 nm to the electrical distance between the gate and the channel, thereby reducing C_ox and increasing the subthreshold slope S.

For any subthreshold slope the leakage current when the transistor is 'off' decreases with increase in the voltage. The steady reduction of the supply voltage that accompanies miniaturization reduces the voltage available to turn the transistor off, allowing leakage to become relatively more important.

4.3. Miniaturized MOSFETs

The simple model of an FET presented in connection with figure 4 becomes increasingly inaccurate as channel lengths are reduced. In a long channel MOSFET the channel length is much greater than the depletion widths. The field along the channel is much weaker than the vertical fields and may be treated as a small effect superimposed on the vertical field. The channel ceases to be long in this sense at some stage of miniaturization and the approximation fails because depletion layer widths are related to the energy gap and cannot be scaled in proportion to dimensions as called for in table 1. While in figure 4 the width of the depleted regions associated with the source and drain junctions has been neglected, when the channel length is decreased to the point at which the widths of the depleted regions approach the channel length they begin to affect the device characteristics. The depleted width is in the range from 100 to 1000 nm and when the channel length decreases below about 1 μm, depending on substrate doping, there is a significant effect on transistor characteristics. The simple planar picture of figure 4 must be replaced by the fully two-dimensional view of figure 9.

4.3.1. Short channel effects. When the length of a gate in a miniaturized FET is decreased while leaving all other parameters the same the threshold voltage is found to decrease with gate length, the 'short channel effect'. The reason for the effect is that because of the part played

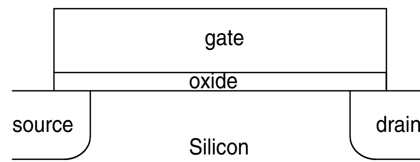

Figure 9. A miniaturized FET, showing how the vertical dimensions become comparable to the lateral dimensions.

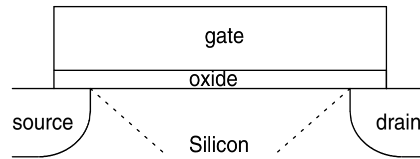

Figure 10. The dotted lines indicate that part of the depleted layer under the gate is controlled by the source and drain, leaving less to be depleted by the gate as the channel length decrease, producing the short channel effect.

by the energy gap in the built-in voltage the extent of the depleted regions associated with the source and the drain is relatively larger in the FET of figure 9 than in figure 4. The charge in the regions depleted by the source and the drain is a significant part of the total charge under the gate. As the gate is shortened the regions depleted by source and drain remain the same, leaving less charge to be depleted by the gate and lowering the threshold voltage.

The electric field normal to the surface and the field perpendicular to the surface are considered separately in the treatment of section 3.1, an approximation that is appropriate when the gate and the channel are large compared with the thickness of the depleted layer beneath the channel. This is known as the gradual channel approximation and fails in the miniaturized devices of contemporary practice when the electric fields from source and drain intrude on an appreciable fraction of the volume beneath the gate. A 'charge sharing' model, illustrated in figure 10 (Nguyen and Plummer 1981), provides a picture of the loss of control by the gate. The point is illustrated by Poisson's equation: $\partial F_x/\partial x + \partial F_y/\partial y = -qN_\mathrm{A}/\epsilon$. Increasing the horizontal field by shortening the channel steals charge from the vertical field. Overall charge neutrality reigns in the device and the charge that compensates the charge in the depleted region under the gate is shared by the gate, source and drain.

The short channel effect translates variability in channel length after fabrication into variability of threshold voltage, it adds to uncertainty in the value of the threshold and to a signal amplitude large enough to turn all transistors on and off.

4.3.2. Drain-induced barrier lowering. A voltage applied to the drain in the operation of the MOSFET enlarges the depleted layer at the drain and, as described above, reduces the amount of charge controlled by the gate. When a positive voltage is applied to the gate the threshold voltage is lowered. This effect is known as drain-induced barrier lowering (DIBL).

4.3.3. Gate tunnelling. Miniaturization aids the flow of certain currents that play no role in the intended function of a transistor. The steady reduction in voltages shown in figure 7 means that the size of the potential barriers that turn off currents is decreased. Decreasing dimensions and lower, thinner potential barriers eventually permit currents that are not an intended part of transistor action.

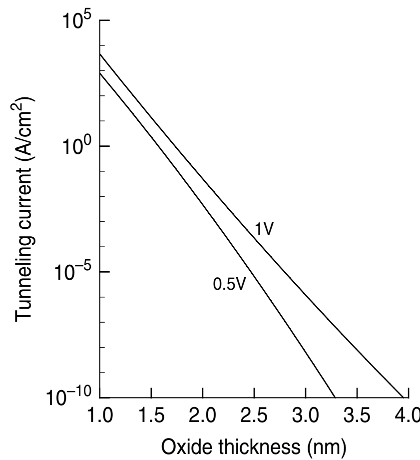

Figure 11. Tunnelling current through an SiO_2 as a function of the oxide thickness and the voltage across the oxide, after Lo (1997).

The gate controls the inversion channel through the capacitance between channel and gate. When the transistor is scaled and the gate length is decreased the insulator is thinned along with the other dimensions as dictated by the rules of table 1 and the capacitance of the gate per unit area increases. This desirable effect is limited, however, by a current of electrons tunnelling through the insulator, which increases rapidly with decreasing thickness and eventually prevents scaling of the gate thickness. A simple geometrical model of oxide tunnelling is susceptible to quantitative analysis (Lo *et al* 1997) and provides the plot of current through the gate oxide to an n-type inversion layer as a function of oxide thickness presented in figure 11 for one and one-half volt potentials across the oxide. The tunnelling current does not prevent the transistor from working, the gate can still control the current between the source and the drain, but the power consumed by the gate leakage current must be supplied and then must be removed from the chip as heat. In CMOS circuits the tunnelling current can flow and consume electrical power as long as there is a voltage difference between gate and channel, even when the drain is not drawing current. Oxide tunnelling is more important in n-type MOSFETs because the larger effective mass of holes in silicon reduces hole tunnelling. Extrapolation suggests that the continuation of miniaturization will lead to more power being dissipated in gate tunnelling than in the useful action of the chip.

The Roadmap regards the acceptable amount of gate leakage as $1\,\mathrm{A\,cm^{-2}}$ growing to $100\,\mathrm{A\,cm^{-2}}$ in 2016 (ITRS 2003). Figure 11 shows that this amount of leakage will be attained with oxide thickness in the range 1.6–1.8 nm for voltage in the range 0.5–1 V, expected to be used with present technologies. However, other authors relax this strict limit (Frank *et al* 2001), and the leakage that can be tolerated depends on the application. For example, the burden of excess power would be a larger disadvantage in a portable battery-powered system than in a fixed system.

In any case, there is a point at which action to control the gate leakage current becomes necessary. Simply not scaling the gate to smaller thicknesses is one route to managing gate tunnelling and is another reason for the failure of strict adherence to table 1 in device design. However, a larger thickness decreases C_{ox} and the smaller gate capacitance limits the control of the channel by the gate, illustrated by an increase in the subthreshold slope due to a lower C_{ox} in equation (2.1) and reducing the current in the on state.

Replacement of the SiO_2 with a material of higher dielectric constant (a concept abbreviated as high-K) is being pursued as the solution to the gate leakage problem. The high dielectric constant means that the insulator can be thicker for a given gate capacitance, maintaining the threshold slope (equation (2.1)) while presumably allowing exponentially less tunnelling current than the oxide. Resort to the high-K route with new materials is being pursued but presents new problems (Kingon *et al* 2000, Foerst *et al* 2004). A material with much higher dielectric constant than SiO_2 and a high bandgap that does not react with silicon or SiO_2 must be found. Retention of the high dielectric constant at the high frequencies involved in the switching of contemporary logic circuits is necessary. The additional requirements are compatibility with the gate electrode material, a high-quality interface with silicon and an ability to survive the high temperatures used in silicon processing. Oxides and silicates are chief candidates (Wong 2002).

The simple one-dimensional picture of high-K insulators must be replaced with a two-dimensional view when the thickness of the insulator approaches other physical lengths in a transistor. A detailed analysis of high-K insulators suggests some constraints on their application (Frank *et al* 1998, 2001) and shows that the physical thickness of the high-K material must be less than the thickness of the depleted layer in the substrate and concludes that dielectric constants larger than 20 are not likely to be useful.

4.3.4. Drain currents. Tunnelling currents may also be found in the substrate at small device dimensions. A positively biased drain with a p-type substrate forms a reverse-biased p–n junction that can carry a small reverse current in an n-type transistor. The more heavily doped the substrate, the narrower the depleted region between drain and substrate and the larger the drain–substrate current. A significant tunnelling current will flow if the depleted region is thin enough, and sufficiently high drain voltages can lead to reverse breakdown of the drain–substrate junction with large currents from drain to substrate. Miniaturization reduces the thickness of the potential barrier in the substrate between source and drain and there is a point at which it becomes thin enough to allow appreciable tunnelling through the barrier from source to drain.

Note also that the source–substrate–drain combination in figure 9 forms an npn configuration as in a bipolar transistor in which the reverse-biased drain collects electrons from the source. Bipolar transistor action with large source–drain currents independent of the gate may occur and combinations of biases that can cause bipolar action must be avoided.

4.3.5. Short channel experiments. The compendium of effects that may limit the progress of electronics to smaller dimensions is large enough to support the conviction that there is a limit. Yet it is complex enough to leave much doubt as to where to draw a line that represents *the limit*. Experience with some of the problems found by actually fabricating transistors with channel lengths far less than current practice, less than 30 nm, at present sheds additional light on the expected limitations.

Transistors with very short channel lengths, to below 10 nm, have been made and studied (Deleonibus *et al* 2000, Doris *et al* 2002, Doyle *et al* 2002, Bertrand *et al* 2004). Experimental fabrication methods, not necessarily adaptable to low-cost manufacturing, are needed for the very small dimensions involved. The channel lengths are smaller than the gate lengths, but at the smallest dimensions the determination of channel length is difficult and the definition somewhat arbitrary. The smallest, shortest gate length transistors worked, having true transistor characteristics. The ITRS Roadmap anticipates that 10 nm gate lengths will be produced starting in the year 2015 (ITRS 2003).

As expected, short channel effects are considerable in sub-30 nm transistors (Doyle *et al* 2002, Bertrand *et al* 2004). Limiting high electric field effects demands continual reduction of power supply voltages with these transistors, as suggested in table 1. DIBL of 100–200 mV per volt of drain voltage were found. Drain conductance increased. Subthreshold slopes were not extreme, 90 mV per decade. Oxide thicknesses are reduced to 1 nm or less, only six times the length of the Si–O bond in SiO_2, to scale with other dimensions. Oxide tunnelling currents of 10^3 A cm^{-2} are to be expected at these dimensions, implying that development of high-K dielectrics will be essential to progress into the sub-30 nm scale of gate lengths (Doyle *et al* 2002). Heavy substrate doping thins depleted layers but allows increased currents between drain and substrate.

Additional processing which tailors the doping of the substrate under the transistor to minimize the above undesirable effects associated with short gates was introduced long ago during the production of much larger devices. ‘Channel engineering’ refers to adding dopants to the substrate under the gate and the contacts to improve transistor characteristics. Short-channel effects, series resistance and threshold voltage are all affected (Thompson *et al* 1998, Gwoziecki *et al* 1999). Many compromises are involved in the design of the contacts to the source and the drain and their extensions. Implantation of ions at an angle with the gate as a mask creates ‘halos’ of heavier doping at the ends of the gate where the contacts meet the channel, extending the contacts under the gate. Normal to the surface the impurity concentration is made low near the channel in order to keep threshold voltages low and the concentration increases to a peak deeper into the substrate to maintain a barrier between the source and the drain. The effect of randomness on thresholds (section 4.1) can be alleviated as part of the adjustment of the depth dependence of the impurity concentration.

If the gate tunnelling can be controlled with new gate insulation, then leakage current between the source and the drain appears to be the most serious obstacle to extreme miniaturization. The experiments with very small transistors showed source–drain leakage currents in the off state increasing at something like the inverse sixth power of the gate length (Doyle *et al* 2002). Off state leakage compromises the operation of the transistor as a switch that is intended to turn the source–drain current on and off.

4.3.6. Mobility. Maintaining high mobility of the holes and electrons is another concern in the miniaturization of transistors. Reduction of mobility would lower the speed of device response and could negate the increase in performance that is one of the incentives for miniaturization. The confinement of the charge carriers in a thin channel near the surface renders them sensitive to scattering by the surface. Surface irregularities, charges trapped at the surface and in the insulator and vibrations of ions in the gate insulator can reduce mobility.

The role of mobility in performance makes increasing mobility an end in itself. Strain, by reshaping the complex energy bands in silicon to reduce the effective mass of the current carriers, advances this objective (Keyes 1986, 2002b, Tiwari *et al* 1997, Mizuno 2000), since mobilities tend to be inversely correlated with effective mass (Keyes 1959). The pursuit of high mobility has led to the introduction of germanium into the silicon material suite. Silicon and germanium are miscible, forming alloys with a range lattice parameter that increases from silicon to germanium. Silicon grown epitaxially on a Ge–Si alloy with a larger lattice constant and strained in such a way that current in the plane of the channel is carried by high mobility low mass electrons (Keyes 1986). Such strained silicon devices are appearing in devices in the market. Charge carriers in germanium have a smaller effective mass and higher mobilities than those in silicon. The reasons that silicon became the material of choice for electronics were detailed in section 1.1, but silicon technology has returned to a thin layer of germanium as a means to obtain higher mobility (IBM 2004b).

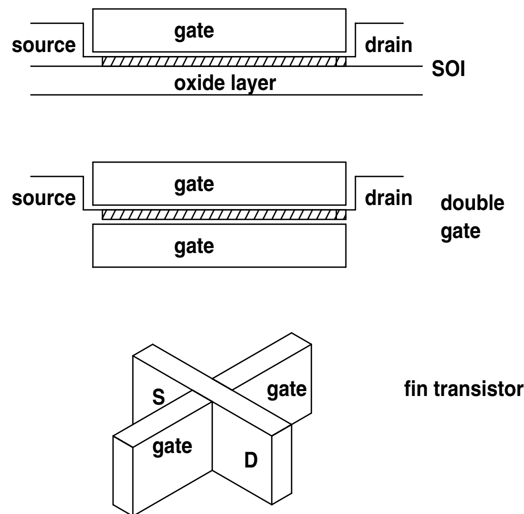

Figure 12. Novel transistors. SOI is the silicon on insulator, built on an oxide layer on a silicon wafer. The shaded area is the active silicon layer. The double gate transistor imbeds a second gate beneath the active silicon layer. The fin transistor is another way of making a double gate transistor; current flows from source S to drain D.

4.4. Novel FET designs

Avoiding the limitations that are appearing in the scaling of MOSFETs to smaller dimensions has encouraged departures from the simple MOSFET which is basically electrodes formed on a silicon surface towards the development of major transistor redesigns. A promising alternative approach makes transistors in a thin layer of silicon on a silicon substrate. SiO_2 is an insulator, giving the technique the acronym SOI for 'silicon on insulator', and it is illustrated in figure 12. This modification leads to substantial changes in the physics of the transistor (Taur and Ning 1998). SOI offers reduced power, faster operation and more efficient use of chip area as compared with devices in bulk silicon (Shahidi 2002). The replacement of part of the source and drain interfaces with bulk silicon by contact with the insulator reduces capacitances, and therefore reduces power dissipation and allows faster operation. Less leakage current from source to drain through the bulk semiconductor adds to the power saved. Since a well is not needed in SOI CMOS circuits some of the steps in wafer processing are eliminated and the elimination of the extra space which was used for the well makes higher component densities possible. The SiO_2 layer also prevents the collection of charge from particle tracks, an advantage in memory chips that may experience 'soft errors' (Ziegler and Srinavasan 1996). The potential barrier between source and drain arising from a heavily doped substrate is not present and increased short channel effects may be found.

The adoption of SOI devices is a major change in device technology. Efforts to make silicon devices in a layer of silicon deposited on sapphire have a long history as soft-error resistant devices for use in space and as high frequency power amplifiers but have never been developed as integrated circuits for computer use because of imperfections in the silicon layer. Fabricating devices in a layer of silicon on amorphous SiO_2 has recently met with greater success and has recently found its way into mainstream products (IBM 2004a).

The ability of technology to precisely control the thickness of the thin silicon layer is a limitation on SOI structures. The silicon layers are so thin that the electron states in them are

quantized; the energy of the states therefore depends on thickness, and variability in thickness causes variability in transistor thresholds. Surface scattering of mobile charge carriers is significant; the thinner the silicon layer, the lower the carrier mobilities, and the thickness also determines whether or not the silicon is fully or only partially depleted. The layer may be made thicker than necessary in order to obtain greater reproducibility, when the layer may be only partially depleted by the gate, allowing some mobile charge carriers to remain in the silicon layer near the insulator at threshold.

A modification of SOI places the mobile carriers in a thin layer of silicon with a gate beneath the layer in addition to the gate above. This ‘double gate FET’ also receives much attention (Doyle *et al* 2002, Shahidi 2002, Wong 2002, ITRS 2003, Lundstrom 2003) and is shown in figure 12. The confinement of the silicon between two parallel gates allows a larger source–drain current to be controlled. The two gates allow thicker layers of silicon to be completely depleted, while retaining the SOI lower subthreshold slope. The structure is difficult to make, the two gates should be accurately aligned, and the thickness of the layer is still a sensitive feature needing tight control. The two gates are usually considered to be biased by the same source, but their existence provides the option of biasing them separately. Difficulties in fabricating a double gate in the plane of the substrate surface have led to interest in ‘fin transistors’ in which the gates are placed on opposite sides of a narrow strip of silicon (ITRS 2003) projecting above the principal silicon surface as depicted in figure 12.

Still another version of a MOSFET is the ‘vertical transistor’, an idea spawned early in the transistor era. The concept uses the thickness of a layer to define a gate length rather than depending on lithography. Ideally it would be possible to build transistors with much smaller gate lengths than those possible with conventional lithography. Vertical transistors have been demonstrated (Hergenrother *et al* 1999).

5. The chip

More and more ever-smaller transistors in a fixed space create problems and meet with limits beyond those found in connection with the transistors themselves. The properties of the chip in which transistors are made and must operate limits their use. The chip provides the communication between devices that is needed to carry out the complex sequences of logic operations that are its purpose. Power dissipated in the switching of transistors and in the wires flows into the body of the chip, which must be cooled to a temperature at which the transistors can operate. Limits to cooling limits what can be done with transistors. The chip must carry enough connectors to its host to transmit and receive the information that integrates the chip and its logic into a larger system. The connectors also carry electrical power to the chip, power that is converted to heat when charge is drawn from the power supply and passed through circuits to ground, the heat that must be removed from the chip. Meeting the requirements of this collection of functions becomes increasingly difficult as larger numbers of tinier transistors are placed on the chips and the chips themselves are constrained by the physical characteristics of large computers. The provision of the hardware that incorporates the chip into a larger system is known as packaging and is dominated by the requirements for communication among diverse segments of a system and for removing heat.

5.1. Wires

In the earliest use of transistors each one was made on a separate chip of silicon and the chips were connected into circuits by discrete wires put in place one by one, an expensive process that limits their use when transistors are numbered in the thousands. Robert N Noyce, a co-inventor

of the integrated circuit, has said 'The integrated circuit is the component industry's solution to the interconnection problem' (Noyce 1977). Today the provision of communication between devices on a chip is part of the chip manufacturing process. The conductors that carry electrical signals from one place to another are called wires, even though they are made by patterned material deposition and removal processes. The largest portion of the wires consists of those that carry signals between the elementary logic circuits on a chip. Other wires carry electrical power to devices and distribute a clock signal which synchronizes the operation of the various parts of the chip.

5.2. Propagation of pulses on wires

The time that signals spend travelling on wires is an important limitation on the rate of operation of a chip. The propagation of signals on resistive transmission lines is governed by the equations (Carslaw and Jaeger 1947, Magnusson 1965, Ho 1982)

$$\frac{\partial V}{\partial x} = rI - l'\frac{\partial I}{\partial t}, \tag{5.1a}$$

$$\frac{\partial I}{\partial x} = -c\frac{\partial V}{\partial t}. \tag{5.1b}$$

In these equations V is the voltage on the line, I is the current and r, l' and c are, respectively, the resistance, inductance and capacitance per unit length. The narrow lines needed to obtain the high wire densities on chips with miniaturized devices have such a high resistance per unit length that neglecting the last term in equation (5.1a) is a useful approximation. Then V satisfies

$$\frac{\partial^2 V}{\partial x^2} = rc\frac{\partial V}{\partial t}, \tag{5.2}$$

which is well known in physics as the equation of heat conduction and has the solution

$$V = V_0 \operatorname{erfc}\left[\frac{x}{2(t/rc)^{1/2}}\right]. \tag{5.3}$$

Here erfc is the complementary error function. For large values of the argument V approaches

$$V = V_0\left(\frac{2}{x}\right)\left(\frac{t}{\pi rc}\right)^{1/2}\exp\left(\frac{-x^2 rc}{4t}\right). \tag{5.4}$$

The response controlled by the exponential is very small until t becomes larger than $rc(x^2)/4$. The time of arrival of a pulse at x is effectively proportional to x^2 and is much longer than the time of flight of an electromagnetic wave.

An elementary scaling or dimensional argument suggests that RC delays might be unaffected by miniaturization. Let Z be the measure of lithographic dimension that decreases with miniaturization. Then the resistance per unit length is inversely proportional to Z^2, $r \sim \rho/Z^2$, capacitance per unit length, $c \sim \epsilon$, is independent of dimension and wire length is proportional to Z. Here ρ is the resistivity of the wire. Capacitance per unit length c is equal to some geometrical parameter times ϵ, the dielectric permittivity of the insulator, and does not depend on dimension Z. Then rcZ^2, the wire delay, is constant. However, rates of operation tend to increase as dimensions are reduced and it is desired that transit times on wires keep up, fuelling a search for faster transit times.

A conflict arises because the increasing density of transistors on chips requires a high density of wires connecting them and the quest for shorter wire delays which can be satisfied with wide, thick wires which lower the resistance r that determines the transit time. Experience

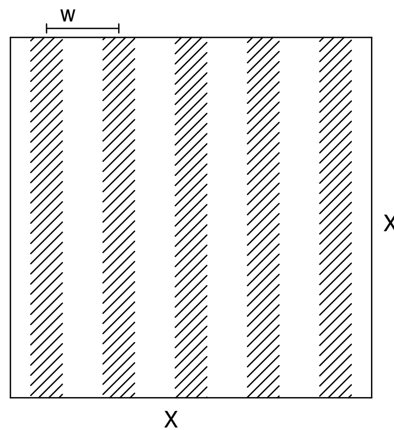

Figure 13. Illustration of the concept of wire pitch, w and used to describe wire density D_w.

has shown that both long and short wires are needed to create the kinds of functioning logic chips that are wanted for computation. Chip wiring attempts a compromise between the use of thin wires for short distances and wide wires for long distances to avoid the longest delays.

5.2.1. Space for wires. Space for wiring on chips is provided as channels in which wire segments can be placed as needed to implement logic functions. Wires are made in several layers to maintain a high density of interconnections, and vias, places where connections between layers may be made, are included in wiring. Alternate layers in which a wire may be placed on chips usually carry channels in mutually perpendicular directions. Distances are measured along these perpendicular axes and are often called ‘Manhattan’ distances. For example, the distance between the opposite corners of a square of side L is $2L$.

Wire channels occupy a substantial area on chips and on packages and the need to devote area to wiring limits the packing density of components. The complexity of the wiring needed to implement logic functions and communicate the results to diverse destinations lends it a pseudo-random character and prevents full use of the channels by physical wires; only a fraction half or less of the available wire channel is actually occupied by a wire. In the subsequent text ‘wire’ usually means the wire channel that needs physical space.

The concept of the ‘wire-limited chip’ (Keyes 1982) gives an overview of the nature of a wire limit. If A is the area of a chip containing M transistors, then $(A/M)^{1/2}$ is the transistor–transistor distance or transistor pitch. Let w be the width of a wire channel (figure 13) and call the length of wire channel measured in transistor pitches required per transistor to implement the logic h. Setting the area available for wire in K layers, KA, equal to the area needed for Mh transistor pitches of wire, $Mhw(A/M)^{1/2}$ gives

$$A = M\left(\frac{hw}{K}\right)^2. \tag{5.5}$$

Here w is the width of the wire channels. The chip area is fixed by the wire requirements, and equation (5.5) shows the effectiveness of many layers in reducing chip area. The point is that as devices are made smaller and their density on a chip increases, adding more layers of wire accommodates the required increase in wire density without too much sacrifice of wire width. This argument is an oversimplification in that the different layers may have different widths

w_j, but wire area is the dominant determinant of chip area today (Davis and Meindl 1998, Davis *et al* 1996, 1998a, 1998b).

The length of wire per unit area, D_{w}, is a useful characterization of the wire density. Referring to figure 13, X/w wires channels can be placed side by side across a layer on a square chip of side X carrying wire channels of width w. The length of these wires is X, the total length in the layer is X^2/w and the length per unit area of the layer is $D_{\mathrm{w}} = 1/w$. The channel density on the chip is found by summing the densities of the wire layers, which may differ from layer to layer as the widths are varied:

$$D_{\mathrm{w}} = \sum_{j=1}^{K} 1/w_j. \tag{5.6}$$

Long wires with long delays limit the performance of logic and their effect is restrained by the use of wide, thick wires on the longer lines as described above. The constraints of fabrication technology mean that wider, thicker wires are placed on different layers than finer lines. The density D_{w} of wire is then the sum (equation (5.6)).

The density of wiring increases along with the increase in transistor density. Wiring at a specific density becomes increasingly difficult as the area to be wired grows. A dimensionless measure of the state of the wiring art that captures the density/area limitation is $D_{\mathrm{w}}A^{1/2}$ or $AD_{\mathrm{w}}{}^2$ (Keyes 1991).

Layers with differing wire widths are used for differing lengths, and longer delays on longer wires must be taken into account in the timing of logical operations. The operation of a chip is governed by a clock: after a short sequence of logical steps information is held in place until released by a clock pulse that is distributed to all parts of the chip. The various parts are kept in step by the clock and the period of the clock is called the *cycle*. The cycle must be long enough to permit all sequences to be completed.

5.3. Wire lengths

The importance of wire lengths in chip design has led to considerable interest in the distribution of wire lengths. In the absence of fundamental principles of computing circuitry that offer guidance, studies of wire lengths rely on empirical observations that have accumulated through years of experience in making useful chips and computing systems. An early investigation of wire limits simulated connections between elements of a square grid of points or ‘nodes’ (Heller *et al* 1977). The wires ran in tracks through the array of nodes and the question studied was as follows: given certain statistical properties of the wires (average length, number of wires per node), how many wire tracks would be needed to allow a randomly selected set of connections between the nodes that satisfied the given statistical constraints to be completely wired? A large number of cases was selected to find the probability of wireability as a function of the number of tracks and a successful wiring of 90% of the statistically identical trials was regarded as a lower limit on the number of tracks or channels needed. Experience with wiring made the authors aware of a tendency for wire lengths to increase as the number of nodes grew and the cases selected for simulation reflected that knowledge. As a result, the required number of wiring tracks grew with the node count.

The most well-known empirical relation relating to interconnection is ‘Rent’s rule’ and concerns the number of connections that must be made to a part to incorporate it into a large system (Landman and Russo 1971). The 1960s packaging hierarchy mounted chips on modules, modules on cards, cards on boards, etc and the names of the parts varied from one manufacturer to another. At that time E Rent observed in an internal IBM memorandum that the number of connections made from one of these partitions of a system to the next higher package

level obeyed a power law: $C = aN^b$, where C is the number of signal connectors, N is the number of logic circuits on the partition and a and b are constants. Rent found that $a \approx 4$ and $b \approx 2/3$. Since then the power law relationship between the number of connections between package levels and component count has been found in many other cases with varying values of constant a and exponent b (Landman and Russo 1971). The rule received considerable attention before around 1980 in an era when the finite space available for external connections limited the amount of logic that could be placed on a chip or other partition of a system, but it receives less attention now that high levels of integration make it possible to integrate on a single chip multimillion transistor systems which need less recourse per device to external entities.

A number of authors have assumed that Rent's rule is also applicable to connections to any arbitrary portion of a chip and used it to derive wire length distributions on a chip (Donath 1979, 1981, Davis *et al* 1998a, 1998b, 2001). While this application is not unreasonable and may be correct (Christie and Stroobandt 2000), it hardly follows from the rather different original significance of Rent's rule, which was an experimental result in cases where the limited availability of connections between package levels and the delays and distortion of signals in the connectors provided a strong incentive for system designers to minimize the number of transitions to another level of the package. The uses of Rent's rule within a chip find that the frequency of wires of a given length decreases with increasing length approximately as a power -0.6 ± 0.2 of the wire length with a rather abrupt termination at a maximum length. They are successful in being able to fit the length distributions of manufactured working chips (Donath 1979, 1981, Davis *et al* 1996, 1998a, 2001).

Christie and collaborators have explored several models that lead to wire length distributions in logic circuits, including the self-similar ideas of fractal analysis (Christie 1993), a neural growth analogy (Christie *et al* 1992) and a probabilistic model that derives Rent's rule (Christie 2001) which yields results similar to those above. The physics suggests still another approach to the problem (Donath 1968, Keyes 1987). The large number of wires on a part and the lack of precise knowledge of where each wire will be located is reminiscent of the problems of statistical mechanics. The lack of *a priori* detailed knowledge of wiring implies that the flexibility, the number of options available to the wiring designer, should be maximized. If the probability that a particular possible connection j with length $L_{j,l}$ will be used is $q_{j,l}$, then the distribution should be the one that maximizes

$$H = \sum_{j=1,l=1}^{\infty} -q_{j,l} \log q_{j,l}. \tag{5.7}$$

The constraints on this maximization are that the probabilities must sum to one and that the total connection length L_{total} is limited by manufacturing technology.

$$\sum_{j=1,l=1}^{\infty} q_{j,l} = 1, \qquad \sum_{j=1,l=1}^{\infty} -q_{j,l} L_{j,l} \leqslant L_{\text{total}}. \tag{5.8}$$

The statistical mechanics view yields an exponentially declining number of wires analogous to a Boltzmann distribution as length increases and does not fit the cases examined by Donath (1981) and Davis *et al* (2001) well. However, examples of exponential distributions can be found (Gardner 1987, Koetzle 1987, Nishi 1988).

Models based on simple rules, though useful guides, do not capture the full limits on wiring. In particular, congestion, an unusual concentration of wire channels wanted in some small region, is not captured in averages, nor is the blocking of wire channels by vias, the connections between wires and the substrate which pass through intervening wire layers, taken into account.

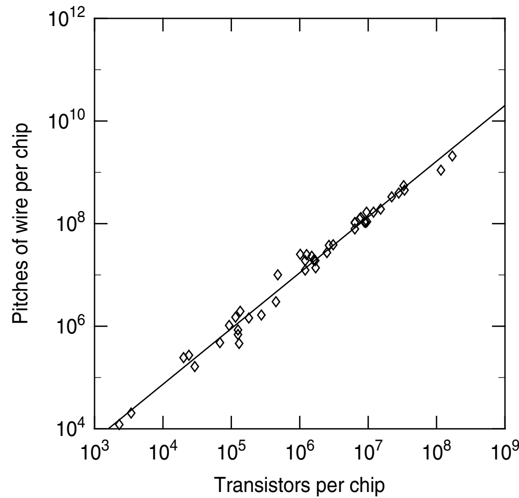

Figure 14. The total length of wire on a microprocessor chip measured in units of the transistor pitch as a function of the number of transistors on the chip, taken from published data. The slope of the line fitted to the points is 1.07.

5.4. Nonlocality

Rent's rule is an example of a property that may be called 'nonlocality', meaning that the larger a part of the system that one examines, the more contact it needs with other parts. Nonlocality in an information processing system is also evidenced by the finding that the larger the number of components on a part, the greater the length of the average wire in it, measured in units of the component pitch (Donath 1979, 1981). Stated another way, the number of components that can be reached by a wire of average length increases as the number of components in the system or in that part of it examined increases. The greater the number of transistors or other functional nodes that a chip contains, the more pitches of wire per transistor are needed (Donath 1979, 1981, Heller *et al* 1984). This requirement is further illustrated by the survey of microprocessor chips reported at the International Solid State Circuit Conferences presented in figure 14 (Keyes 2002a). Here the length of wire channel on a chip measured in units of the transistor pitch is plotted as a function of the number of transistors on the chip. The slope of the line fitted to the data is 1.07, the length of wire increasing faster than the number of transistors.

5.5. Long wire delays

The lengths of many wires on a chip do not scale with lithographic dimension; a small number of wires have lengths of the order of the chip edge (Davidson 1997, Davis *et al* 1998b) and chip sizes and the physical lengths of the longest wires have actually increased over time. As dimensions are reduced and wires become thinner the resistance of the long wires increases rapidly. The problem posed by the increasing resistance of long wires has been contained by a combination of means: placing a layer of wide wire on top of the scaled layers, increasing the aspect ratio (height/width) of the wires and replacing aluminium wire with copper. Nevertheless, continuing miniaturization steadily exacerbates the problem.

The state of affairs is shown in figure 15. The RC delay for the coarsest reported wire layer and a length equal to the chip edge is estimated with equation (5.6) for microprocessors whose

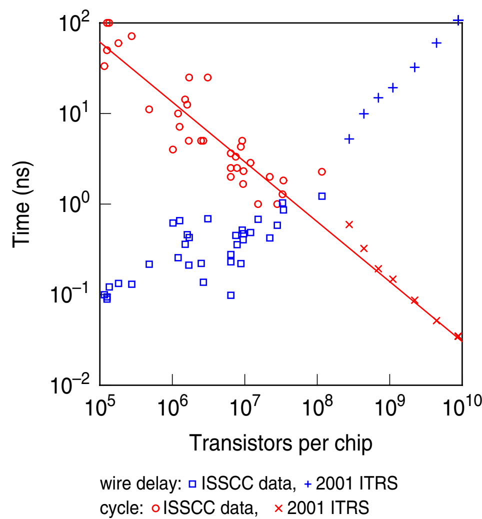

Figure 15. Examples of the cycle time and the wire delay on a global wire with the length of the chip edge for microprocessor chips taken from the published literature and the Roadmap (ITRS 2001). The line is a least squares fit to the circles.

Table 2. The coarsest wire layer of the 2001 SIA Roadmap Chip area $= 310\,\text{mm}^2$.

Year	Mtrans per chip	Clock (ps)	Wire layers	Dielectric constant	Global wire	
					Pitch (μm)	Aspect r
2001	276	540	7	3.3	0.670	2.0
2003	440	320	8	3.3	0.475	2.1
2005	698	193	8	2.85	0.360	2.1
2007	1107	148	9	2.5	0.290	2.2
2010	2213	87	10	2.1	0.205	2.2
2013	4424	52	10	1.9	0.140	2.3
2016	8848	35	10	1.8	0.100	2.4

metallization was described in the proceedings of ISSCC meetings (Keyes 2002a) and in the International Technology Roadmap for Semiconductors (ITRS 2001). The delay is plotted as a function of the number of transistors per chip, used as a measure of the state of the art of integration. The cycle time of the microprocessors is also plotted. The delay is less than the cycle time although approaching it closely in the most recent cases.

The line in figure 15 is a fit of delay versus clock for the ISSCC data and is extended to years included in the Roadmap. The extrapolation shows that the Roadmap forecasts a continuation of the trend in the cycle. The Roadmap expects that wire resistance will be reduced by increasing wire aspect ratios (thickness/width) greater than one, even though this increases the capacitance and there is a point of diminishing returns. Assuming a reduced dielectric constant of the insulator in later years, giving a smaller c in equation (5.6), further reduces the expected delays. The accompanying table 2 presents the relevant data from the Roadmap (ITRS 2001).

Figure 15 also plots the estimated wire delays for a wire the length of the chip edge for the coarsest wire layers (global wires) with the data of table 2. Aggressive miniaturization of the wiring causes long wire delays to increase very rapidly. In spite of the significant modifications

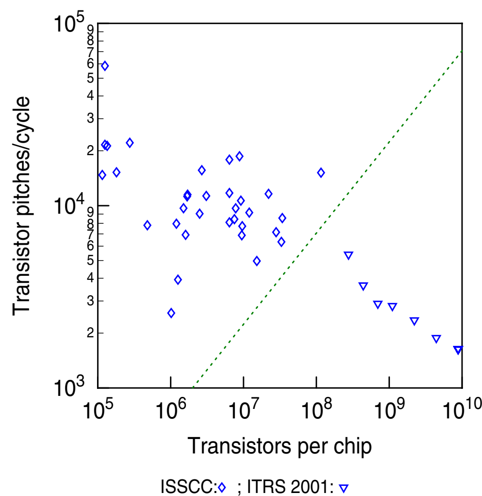

Figure 16. The length of wire (in transistor pitches) accessible in one cycle by a wire in the global layer with the data of figure 15. The dotted line is the locus of points for which the wire length is equal to the chip edge.

of the dielectric constant and aspect ratio the projected delays for the long wires increase to over one thousand times the cycle!

Another view of the same information asks: How far can a signal travel in one cycle time? Figure 16 shows the distance measured in transistor pitches traversed by a signal on the coarsest wire layer in one cycle for the ISSCC data and for table 2. Figure 16 shows that there has been a marked decline in this measure as technology advances. The trend is expected to continue into the Roadmap era. The dashed line that neatly separates the ISSCC data and the Roadmap projections shows the point at which the Manhattan distance of a line with RC delay equal to the cycle time can reach a number of transistors equal to the total number of transistors on the chip. The Roadmap projections allow an increasingly small fraction of the chip to be reached in a cycle. It appears that long wire delays will be a severe limit on the functioning of future chips.

Repeaters, circuits that restore the shape and amplitude of signals and retransmit them periodically along a long path, have been considered to be a way to mitigate the effects of delays on long wires. If there are k repeaters the rcl^2 delay is replaced by $k(rc(l/k)^2)$ (Davidson *et al* 1997). Ten repeaters could reduce wire delays by a factor of 10, although some time would be added by delays in the repeater circuits.

5.6. Power on a chip

Circuits on a chip draw charge from a source that supplies voltage V when operating. The charge enters and leaves a capacitor of some kind (a wire, a pn junction, the gate of an FET) and is subsequently discharged to ground, leaving behind heat in amount CV^2. Capacitances are dimensionally a length and decrease in proportion to the length dimensions as devices are miniaturized. As the scaled density of devices on the chip grows as fast as the inverse square of the dimension, however, the capacitance per unit area and the density at which power is converted to heat increases linearly with miniaturization at constant V.

Several factors affect this simple-minded argument. Compaction may cause the density of devices to increase faster than the inverse square of dimension. Device design seeks to

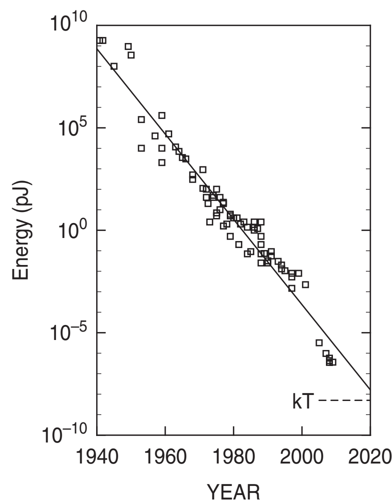

Figure 17. The energy dissipated in a single logic operation.

minimize the use of power, as in the SOI transistor (figure 12). The voltage used has declined steadily, while CMOS circuits draw current only intermittently and the rate at which the devices operate has grown. Figure 17 shows the history of the energy used in an electronic switching event. The comparison of the energy with the thermal energy kT is shown since that, as the energy dissipated in an irreversible binary decision operation, has been regarded as a lower limit to dissipation in a logical decision (Landauer 1961, Keyes and Landauer 1970, Stein 1976). The interest in the switching energy itself has declined, however, because other sources of dissipation in devices now play an important role (sections 4.3.4 and 4.3.5) and the increasing amount of wire on chips of a given size has led to the power used in charging wires to dominate that used in device action (Meindl *et al* 2002).

Miniaturization in the past decade has met another element of the power/cooling problem: power produced by leakage currents. Simple extrapolation suggests that power from leakage currents will exceed that from the device operation in a few years. A high-K gate dielectric and revised transistor designs are expected to postpone the leakage limit (section 4.3.3) (Wong 2002).

Heat is removed from a computer by transferring it to a fluid that can be removed from the system. The fluid is commonly air, although water-cooled systems have been used (Blodgett 1983). Heat transfer at the solid–fluid interface forms a bottleneck in the removal of heat by a fluid, and heat transfer coefficients in forced air cooling are less than 0.3 W cm^{-2} K^{-1}. If the temperature of the chip rises 80 K above the coolant, then up to 25 W might be removed from a 1 cm^2 chip, but the heat sinks which increase the area exposed to the coolant, such as metallic fins, can relieve this limit. In practice the use of forced air and other measures for cooling is severely limited by the space requirements of other packaging functions: signal and power connections, mechanical support and protection and provision for assembling the parts. The heat produced on the chips must be removed not only from the chip but also from the computer that uses it and from the building in which the computer is housed. Each stage contributes to the cost of computing, meaning that limiting the power consumed by a chip is an important objective.

There is a tradeoff between rate of heat transfer and the effort expended; more pressure forces more fluid to flow. In a *tour de force*, Tuckerman and Pease (1981) showed that a kilowatt

can be removed from a one centimetre square chip by forcing water through thin channels etched into the back of the chip. The experiment is not easily translated to a manufacturable process with a packaging technology to support it, however. The use of water cooling also adds significantly to the packaging complexity and the cost of a computer and makes additional demands on the installation.

Marked advantages in performance accrue to a system that would be operated at the temperature of liquid nitrogen (Gaensslen *et al* 1977). A commercial enterprise did develop and market a liquid nitrogen cooled computer, the ETA-10, in the late 1980s (Carlson *et al* 1989). Several systems were sold, but there was no financial success and the business had a brief life.

5.7. The billion transistor chip

The thought of a one billion transistor chip excites much interest in our decimal-oriented world and should become a reality in about two years. According to the extrapolation in figure 16 a signal from a point on the chip will only be able to reach around 1/100 of the transistors on the chip in one microprocessor cycle. The significance of the long delays illustrated in figures 15 and 16 as a limit depends on what one expects of the chips that incur them. A part of a chip might be devoted to a memory bank that need not be accessed in a single cycle. An increasingly attractive use of chips with a very high transistor count is to organize the transistors into a number of relatively independent microprocessors which do not need to communicate with one another every cycle. Such multiprocessor chips are already entering the marketplace (EE Times 2001), primarily as 'dual core' chips containing two microprocessors. A second advantage of multiprocessor products is simplification of the formidable task of designing hundred million and more transistor chips. The projected thousand cycle delay in figure 15 refers to a chip with 10^{10} transistors. Such a chip might contain 1000 ten million transistor microprocessors between which many-cycle delays might be acceptable, for example.

6. Fabrication

The smallest possible physical size of FETs is far beyond that of the contemporary integrated circuits. Working FETs with gate lengths down to 5 nm have been demonstrated (Deleonibus *et al* 2000, Doris *et al* 2002, Doyle *et al* 2002, Bertrand *et al* 2004). However, the ability of fabrication technology to economically build generation after generation of smaller and smaller devices on silicon wafers is by no means assured. The state of the art today is impressive by many measures. For example, a modern silicon chip can contain a kilometre of wire channel, and a 300 mm wafer with 1/4 μm minimum dimension features contains 10^{12} (1/4 μm)^2 pixels (a good camera lens resolves 10^7 pixels). The adoption of the technology by applications that have little to do with silicon electronics, such as magnetic recording, thin film displays, quantum computing, MEMs and experimental science attest to its success.

But a long record of high achievement does not ease the path to the more drastic miniaturization expected to characterize the coming years. Manufacturing technology faces problems in the future for which no solution is now in view. The art of integrated circuit fabrication has been advanced in a series of steps or generations of tools characterized by dimensional parameters shown in figure 7. Roughly, each generation has reduced the minimum dimensions by a factor of 0.7, leading to an increase of a factor of 2 in component density on a chip. Chip sizes have been increased with some help from 'cleverness' enough to increase the number of devices per chip by a factor of 4. The ITRS Roadmap presents a picture of what may be expected of fabrication technology through the next decade (ITRS 2003).

6.1. The lithographic process

Lithography has long been a fertile field for seekers of limits, beginning with a proposal decades ago that features could be no smaller than 1 μm because that was the wavelength of light. Today's manufacturing methods use a series of pattern-defining steps, known as 'mask levels', to create a wafer of working chips. The principal operations in a 'mask level' are: (1) deposit a layer of light-sensitive material (photoresist) on a surface in a way that produces a uniform adherent layer, (2) expose the photoresist to a demagnified image of a mask that contains the pattern to be produced. The exposing optics cannot cover an entire wafer at once, so the wafer is moved at each masking step from one small area to another until the whole wafer is exposed. The light changes the properties of the photoresist making it either less soluble or more soluble (two kinds of resist) and (3) develop, that is, remove the resist from the more soluble areas. The substrate under the resist-free areas is then subjected to further processing. Each of these steps consists of many sub-steps, such as repeated cleansings and heat treatments, and twenty or thirty such mask levels are used in the manufacture of today's chips.

The resist is frequently placed on a layer of SiO_2 that, after exposure and development of the resist, is removed by an acid etch from the resist-free areas. Then the remaining resist can be removed leaving the patterned oxide as a mask for processing of the underlying surface. The parts of the surface that are not masked by photoresist or by SiO_2 may be treated by etching or material deposition or exposure to some reactant or implanted with ions.

Several aspects of optical exposure limit the dimensions of structures. The requirement that each masking and exposing step be precisely placed with respect to all of the preceding steps is one. The steady reduction of dimensions increases the mechanical accuracy required in this alignment, 'overlay' error is expected to be a small fraction of the minimum feature size. High spatial resolution in the image on the wafer is needed and requires large numerical aperture (na) lenses. The diffraction limit of resolution on the wafer is $x_d = k_1\lambda/\text{na}$, where λ is the wavelength of the exposing light and k_1 is a constant, and should be small compared with the minimum pattern dimension. During the last two decades computerized lens designs have increased na to values near one and substantially reduced k_1, making it possible to fabricate features with size considerably less than the exposing wavelength. However, using a high na for high resolution leads to a small depth of field (DoF). At a distance d from the focal plane a point becomes an area of diameter $r = \text{na} \times d$. 'Depth of field' (DoF) is used in lithography to mean the vertical distance from the focal plane at which r is equal to the diffraction limit, $r = x_d$. Thus $\text{DoF} = k_1\lambda/\text{na}^2$ and the lithographic depth of focus decreases as na^2. Another way to look at DoF is to eliminate na from r and x_d, obtaining $\text{DoF} = {x_d}^2/k_1\lambda$. DoF is increased by using the smallest possible wavelength for a given diffraction resolution x_d.

Lithography uses monochromatic light sources with lenses designed for best performance at the source wavelength and a particular transparent material. The sources have progressed through a succession of decreasing wavelengths in order to achieve the higher resolution needed for smaller features. A new lithographic generation means that a different shorter wavelength light source intense enough to permit high throughput must be found and that new resists optimized for the new wavelength will be needed. Lenses must be redesigned and sometimes new transparent materials for lenses and mask supports may be required. The history and possible evolution of light sources for lithography are summarized in table 3.

A change in the wavelength usually also involves the development of a new photoresist material sensitive to the changed radiation. The changes in k_1 that have marked the technology for reducing feature sizes faster than wavelengths has been made at the expense of lowered contrast at the silicon surface which makes additional demands on the resists.

Table 3. Light sources for future lithography.

Node	Wavelength (nm)	Light source	Lens
250-180	365 i-line	Mercury arc	Glass
180	248 Deep UV	Excimer laser (KrF)	Fused quartz
130-65	193 Far UV	Excimer laser (ArF)	Fused quartz + CaF_2
65-45	157 Vacuum UV	Excimer laser (F_2)	Reflective optics
65-32	13.4 Extreme UV	Plasma	Reflective optics

Decreasing wavelengths means higher energy photons, and the availability of transparent materials that can resist damage from the energetic photons is another problem for the extension of lithography to shorter wavelengths. Reflective optics, mirrors made of layers of thin films (Spiller 1976, Spiller *et al* 1992), would relieve the limits on lenses when and if the production of mirrors of the requisite quality at an acceptable cost can be assured.

Exposure with electron beams rather than electromagnetic radiation has often been demonstrated and could eliminate the light source problem and, in addition, relieve the limits associated with resolution and depth of focus. The writing of patterns with electronically controlled scanned beams is not bound by the diffraction effects of electromagnetic radiation and is used to prepare structures beyond the limits of optical lithography for research purposes; for example, the transistors with gate length far beyond the manufacturing practice mentioned above, and for lithographic masks. However, high throughput demands parallelism by simultaneous exposure of a considerable area rather than by a scanning beam. Projection exposures with masks and an electron source are not impossible and have been demonstrated (Berger 1991, Pfeiffer and Stickel 1995) but they have not yet met the requirements of high throughput processing.

When lithographic exposures are made with light with a wavelength much smaller than the features on a chip contact masking may be used: a mask bearing an image with the size of the finished pattern is placed in contact with or very close to ('proximity' masking) the substrate and exposed to the light. Contact masking was used in semiconductor processing in the early development of semiconductor manufacturing when the dimensions on a chip were 10 μm and larger and light sources in the visible range were used. The use of x-rays for lithography today would enable proximity masking to avoid the use of lenses but is limited by the lack of x-ray sources of sufficiently high intensity. A further limitation might arise from exposure of resists by secondary electrons produced by the high energy photons.

6.2. Processing limitations

Maintaining the low cost of electronic components depends on handling a large area of silicon in each operation. Current practice deals with 300 mm diameter wafers containing hundreds of chips. However, perfect uniformity all over a 300 mm wafer of the many parameters that affect a process cannot be achieved. Chemical reactions are thermally activated and proceed at a rate that is very temperature dependent and so may vary from one part of a wafer to another in the presence of a small temperature inhomogeneity. The concentration of a reagent that is in the process of being used up could differ from one point on the wafer to another. The intensity of radiation will not be perfectly uniform over a large field and the thickness of the photoresist will vary slightly over a wafer. The small variability of process parameters causes variability in the structures produced, in such things as the size of a chemically etched opening and the thickness of layers formed by chemical reaction across a substrate.

Each masking step in a long series alters the results of preceding steps to some extent. For example, each heating cycle allows impurities introduced in previous cycles to diffuse, constraining the time–temperature treatments that may be used. The deposition of one material on another can strain a substrate and the juxtaposition of materials with differing coefficients of thermal expansion creates strains during heating and cooling that can cause plastic deformation that is retained in the wafer to affect the following steps (Shen *et al* 1996, Noyan *et al* 1999). The strains can distort the wafer and affect the accuracy of focus.

The diversity of conditions on a wafer causes significant variability of the properties of devices and chips across wafers and even within chips. The variability between nominally identical devices must be taken into account in circuit design. It limits the performance of chips in computing circuits (Orshansky *et al* 2000) and a few chips may even fall out of the range of acceptable performance. Processing must strive to maintain uniformity in its products.

6.3. New transistor designs

There is more to the transition to a new product generation than making everything on a chip smaller. Device modifications, such as cleverness, frequently entail increasing the number of process steps. The patterning of doping concentrations mentioned in section 4.3 is one example of added process steps and the cleverness example using a capacitor in section 1.2 supplies another. The additional layers of wire that usually accompany an increase in components on a chip obviously entail more process steps.

In addition, new device designs and the introduction of new materials into devices challenge fabrication technology. Besides finding something that works a method for manufacturing it economically must be developed.

Two urgent thrusts aimed at relieving the approaching power dissipation limit that were introduced in section 4, SOI and the high dielectric constant gate insulator, illustrate the demands made on the development of fabrication processes. Two methods of making wafers for the SOI devices illustrated in figure 12 are practised. One procedure forms a layer of SiO_2 on a silicon wafer and bonds the oxidized surface to another silicon wafer. Silicon is then removed from one wafer to leave only a thin layer separated by oxide from a thicker silicon substrate. Another method uses ion implanation to produce a high concentration of oxygen in a layer beneath the silicon surface. High temperature annealing of the implanted wafer causes crystallization of a silicon layer at the surface which is separated from the bulk by a layer of SiO_2. In either case the thickness of the thin silicon layer is a critical parameter of SOI and of double gate transistors. Thicker layers for partially depleted SOI require less precision than thinner layers.

Figure 12 also shows the double gate transistor in which in addition to the gate above the transistor another gate placed below the thin silicon layer offers significant performance advantages. It allows a thicker silicon layer to be depleted and larger currents than those of the simpler SOI transistor to be controlled. However, fabrication of a double gate transistor presents an even more difficult task than that posed by the single gate SOI transistor. One possible sequence is suggested by Wong (2002). An alternative way of making a double-gate is the fin transistor, also presented in figure 12.

Another theme attacks the limit posed by power wasted by leakage through the gate insulator. Continued thinning of the SiO_2 insulator which accompanies further miniaturization of CMOS transistors is untenable because of the rapidly increasing tunnelling current. Replacing the SiO_2 dielectric with one with a higher dielectric (high-K) constant allows the insulator to be physically thicker while maintaining a large gate–channel capacitance. A high bandgap in the dielectric is also essential to the reduction of tunnelling. Achieving the desired result means that the factory must assimilate a new material into the production process

with a minimal upset of the other results of the process, a requirement that further restricts the choice of a high-K material. Chemical reaction of the new insulator with the silicon surface and the gate material must be avoided and reaction with SiO_2 is also unacceptable. A method for depositing the material reproducibly on a silicon surface must be developed, while the chemical environment in which the deposition takes place has a significant effect on the interface properties (Foerst 2004). The considerable research effort that has been expended on the search for a suitable material has uncovered some promising candidates but none has been adopted for large-scale processing yet.

The oxynitride is a first step towards the use of a material with a higher dielectric constant (high-K) and good insulating properties as the gate insulator. It enables some reduction of gate tunnelling with materials quite compatible with silicon technology: combining SiO_2 and silicon nitride (Si_3N_4) in the insulator. The results depend on how the mixture is processed. One method places a layer of silicon nitride above a thin SiO_2 layer as the gate insulator (Shi *et al* 1998). The valued qualities of the contact of SiO_2 on silicon are preserved. The nitride has a dielectric constant of 7 as compared with 3.9 for the oxide, and silicon nitride is well known in silicon technology. Various mixtures of the form SiN_xO_y, where $3x + 2y = 4$, mixtures known as oxynitrides, may be used as the insulator. A gate–substrate capacitance equivalent to an oxide insulator about 2 nm thinner could be obtained for a given tunnelling current density.

Less pressing is the search for an alternative material with lower dielectric constant to replace the SiO_2 that supports the interconnections. The intent is replacement of the *c* in '*rc* delay' with a smaller value (equation (5.4)).

The replacement of aluminium, long used for wiring of integrated circuits, with lower resistivity copper to reduce rc delays is a success story in the introduction of a new material. Copper is also much less susceptible to electromigration, the movement of atoms carrying large electrical currents, than aluminium. As an example of a new problem posed by a new material it was found necessary to have a lining of tungsten around the copper to prevent reaction with the SiO_2.

6.4. Defects

At any given time the abilities of a silicon fabrication facility cannot produce full wafers of devices and wires completely free of defects that prevent a chip from functioning. This is one of the reasons for dividing wafers into individually testable chips. Ordinarily, chips are tested while still part of the wafer and the fraction that is usable is called *yield*. Yield is closely tracked by a facility as high yield is important to its economic viability.

Tracing the sources of problems to particular locations aids identification of the origins of defects in inoperable chips and is a first step in eliminating the causes. Devices are made in a silicon layer grown epitaxially on the wafer cut from a large single-crystal ingot. Intrinsic crystalline defects (agglomeration of defects and interstitials) and oxygen precipitates in silicon may cause defects (Falster and Voronkov 2000). Precipitation of doping atoms that have been introduced at high concentrations is a source. Dislocations in the substrate, propagated from the original ingot or generated by the strains introduced in processing may cause a defect. Dirt, the finite concentration of particulate matter in clean rooms, is another source. Tools used to move wafers during processing and other equipment malfunction can damage chips or otherwise cause defects.

The number of defects per unit area is a limit on the size and content of chips. Continual attention to identification of defects and their sources has produced a steady decline in defect densities through the integrated circuit era, making ever-larger chips possible in spite of the continuing reduction of dimensions that reduces the size of a fatal defect.

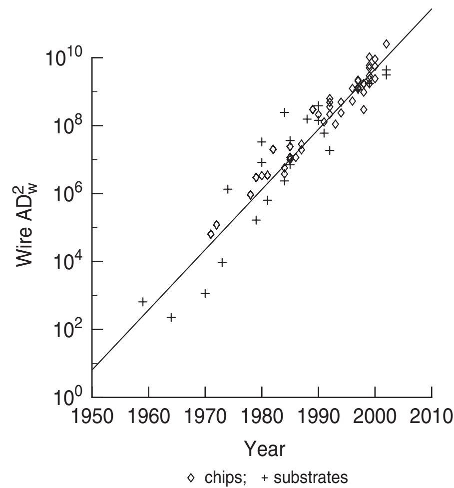

Figure 18. The increase of the wire measure AD_w^2 with time. The equation for the line fitted to the data is $G(t) = 4.3 \times 10^9 \exp(0.406(\text{Year-2000}))$.

The large amount of wire, a kilometre in up to ten layers on a chip, makes wire defects an important source of failures. A model of failure starts from the reasonable assumption that the size of a defect that can cause a loss of function is proportional to the minimum dimension on a part, to the channel width in the case of the wire. Thinking of wires of width w on a substrate with a concentration $C(w)$ of defects per unit area in the size range, $w, w + \mathrm{d}w$, call the minimum size of a defect that can cause a failure w_0, some fraction of w. Then the concentration of fatal defects N_D is

$$N_D = \int_{w_0}^{\infty} C(w)\, \mathrm{d}w. \tag{6.1}$$

Ho *et al* (1982) and Rose (1991) cite evidence that $C(w)$ is proportional to $1/w^3$. Then from equation (6.1) N_D is proportional to $1/w_0^2$. The yield for a part of area A is a function of the average number of defects, $N_D A$, leading to the conclusion that changing all linear dimensions ($A^{1/2}$ and w_0) of a structure by the same factor does not change the yield. For example, yields for chips may be scaled to the structures that support and interconnect the chips, called by names like modules and cards and boards.

In this scaling view, multiplying all dimensions by a factor greater than one would increase the wire pitch and decrease the amount of wire channel in a given part by that factor (equation (5.2)) and would increase the area to be wired by the square of the factor. The wire density D_w and the area that can be wired at that density, A, are related by

$$AD_w^2 = \text{constant}. \tag{6.2}$$

The argument can be used at any particular point in time. In the course of time, however, improvements in many aspects of processing reduce $C(w)$. This steady reduction of defect concentrations has made increasingly complex chips and packages possible, so that the ‘constant’ in equation (6.2) increases with time, making it more useful to write (6.2) with a function of time on the right-hand side:

$$A{D_w}^2 = G(t). \tag{6.3}$$

Figure 18 is presented to determine the function $G(t)$. The product AD_w^2 for the wiring on chips and for packaging substrates is plotted as a function of year. For example, the

IBM Thermal Conduction Module, a water-cooled 9 cm square chip carrier introduced in 1981, had sixteen layers of wire with a pitch 0.5 mm (Blodgett 1983) and an area $A = 81\ \mathrm{cm}^2$, giving $D_\mathrm{w} = 320\ \mathrm{cm}^{-1}$ and $AD_\mathrm{w}^2 = 8.3 \times 10^6$. The line is a least-squares fit to the data for substrates but it is also a good fit to the chips. It corresponds to the equation

$$G(t) = 4.3 \times 10^9 \exp(0.406(\text{Year-2000})). \tag{6.4}$$

$G(t)$ grows by a factor of 10 in a little more than five years and figure 18 shows that wiring on chips and on substrates has indeed advanced together as suggested. Figure 18 and equation (6.4) may be extrapolated and interpreted as a limit to the abilities of wiring technology in the future.

6.5. Additional factors

The cost of everything associated with a new lithographic generation is greater than that of the preceding generations. Light sources, cleaner clean rooms, purer reagents and novel lens materials all are more expensive. More steps in processing require more equipment. Larger wafers need larger processing chambers. The rapidly rising capital cost of a silicon integrated circuit manufacturing facilities (a fab) has been called ‘Moore’s second law’ and expressed as a doubling of the cost of a fab every two years or some similar time period and aroused fears that rising costs rather than physics or technology will end the era of increasing miniaturization and integration (Forbes and Foster 2003). It is true that to date each facility produces many more products so that the cost of electronics per transistor has declined by a factor of around 2/3 per year and the cost of a million instructions per second (MIP) of computing power has fallen even faster.

The large size of the semiconductor industry has attracted much attention to it, in particular, to the impact of semiconductor processing on the rest of the world (Ball 2002, Brumfiel 2004). The Roadmap (ITRS 2003) directs attention to such concerns in an ‘Environment, Safety, and Health’ section. The transition to a new generation of technology must continue to meet high standards in the management of waste disposal, resource conservation, emissions, chemical handling and workplace protection.

7. The transistor?

One may wonder why, in view of the physical limits threatening the future of the FET, computer technology does not simply use some other switching device. Indeed, the Roadmap does advance alternatives as replacements for the transistor (ITRS 2003). However, the transistor is unique and will not be soon replaced. Ever since the invention of the transistor physicists and technologists have tried in vain to produce another solid state device that promised advantages over the transistor concept. The invention of another notable semiconductor device, the tunnel or Esaki diode (Esaki 1958), followed the discovery of the transistor and became the subject of early attempts to realize logic with a two-terminal negative resistance device (Gentile 1962). Two terminal devices depend on ‘threshold’ logic: if the sum of a number of signals applied to one of the terminals is greater than some threshold the circuit switches, otherwise nothing happens. Threshold logic is hindered by variability in component parameters and by signal noise (Keyes 1989), and the difficulty of making reproducible tunnel diodes led to a loss of interest in them for logic in the middle 1960s (Holmes and Baynton 1966).

Nevertheless, attempts to use embodiments of threshold logic to replace three-terminal transistor circuits persist and include Josephson junctions (Matisoo 1967, Anacker 1977), optically bistable nonlinear interferometers (Gibbs *et al* 1979, Miller *et al* 1979, Bowden *et al* 1981, Garmire 1985) and resonant tunnelling diodes (Goldhaber-Gordon *et al* 1997). None has had any effect on computing technology.

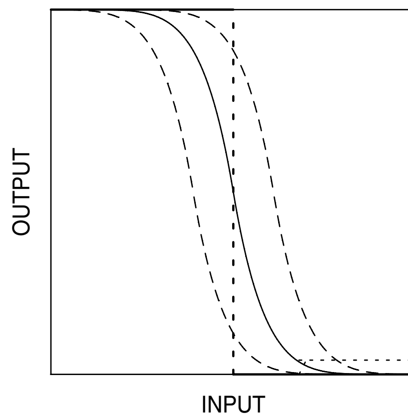

Figure 19. The response of a transistor inverter circuit that restores signals to their intended values.

In view of the Roadmap (ITRS 2003) position it is relevant to ask what magic has made it possible for the transistor (and the vacuum tube) to function in the hostile environment of systems of thousands to hundreds of millions of closely packed electronic logic circuits? The crucial difference that distinguishes the transistor from other solid state devices is gain, the power to amplify signals by large factors. And why is gain essential to devices in a large system? The role of gain can be understood with the help of figure 19, the response of a three-terminal inverting circuit, for example figure 5, to an input pulse. The insensitivity of the response to significant deviations of the input signal from its intended value (0 or 1) means that signals that have suffered attenuation and have been altered by noise or crosstalk between wires produce a corrected output. This is the essence of why binary digital representation is used in computers: there are two standard signal levels in the system. At each step in a computation a signal need only be recognized as one or the other, a zero or a one, to be restored to a known standard value.

Further, as shown by the dashed lines in figure 19 the response of a high-gain transistor to the input signal does not need to be tightly controlled to provide the noise margins, the regions of input that are readily interpreted as zeros or ones, to restore signals. High gain makes it possible for the output signal to vary rapidly in the narrow range of inputs far from the 0 and 1 in any case.

Another important aspect of transistor circuits is the insensitivity of the input to the condition of the output; the three terminals of transistors allow the required isolation of input from output. No signal is sent back to affect preceding logic stages. Charge can be amplified to provide inputs to many receptors by allowing current to flow through a transistor for a long enough time to charge or discharge all of the following stages. The circuit of figure 5 shows how a multiplicity of inputs can be accepted. The circuit can switch in either direction in comparable amounts of time, no separate resetting operation is needed. And the inversion which is necessary for a complete logic system is available.

The question for device technology is not ‘What can be done with a few devices in a laboratory?’ but ‘What is necessary to make a device that operates in a large system with temperature gradients, noisy signals and device to device variability?’ (Keyes 1989, 2001).

The high gain that results from the ability of a small number of elementary charges to control the flow of a far larger number of mobile charge carriers is the key to the stunning success of transistors. The gain made possible by the Coulombic attraction between two signs

of charge is shared only with the amplifying vacuum triode, in which the fixed charge is held on the grid. No other physical phenomenon provides such a powerful force for amplification and no other electronic device has matched the gain available with vacuum tubes and transistors. That is the reason that tubes and transistors have dominated electronics, including not only logic circuits but also electrical amplification of frequencies from the audio to the near infrared range, for almost a century.

8. Limits

Throughout the last half-century the faith of both observers of and participants in the advance of solid state electronics in a long-term continuation of historic progress in the power of information processing machines very far into the future has been sorely tried. The history of the subject is littered with announcements of limits that will mean the end of progress. As early as 1949 John von Neumann is reported to have said that 'It would appear that we have reached the limits of what is possible to achieve with computer technology' (Kurzweil 1999). Needless to say, none of the many other visions of a limit has turned into a dead end for computer technology.

In fact von Neumann revealed his skepticism concerning limits by concluding the above sentence with a rare note of caution 'although one should be careful with such statements, as they tend to sound pretty silly in five years' (Kurzweil 1999). Later authors announcing a limit omit any such reservation, but still today those who speak of limits should be prepared to look back at their pronouncements five years later with some chagrin.

Enduring limits are hard to find because limits only apply within some physical context. Referring to von Neumann's 1949 remark, at that time the context did not include transistors. What is it that we do not know today? Inventors and development laboratories are working hard to evade apparent limits by expanding the context in which thinking can take place. Contemporary estimates of limits are mainly based on a silicon-SiO_2 view of the world. This was once an eminently sensible view, as the introduction of any new material into the well-established silicon technology is fraught with danger. Unforeseen chemical reactions or material transport on microscopic scales can cause problems that are only revealed after the large scale use of a product is well underway.

Present limits to continuing miniaturization and scaling are being forced back by changes in the silicon-SiO_2 context. The introduction of new materials into the well-established arsenal of silicon technology has begun: traditional aluminium wires have been replaced with copper to decrease resistance. Germanium is being added to silicon to form a substrate with a larger lattice constant to take advantage of the effects of strain on hole and electron mobility and to contain a high-mobility FET channel.

The most pressing limit to further miniaturization today is the reduction of the high power dissipated on chips. The replacement of the SiO_2 gate insulator with a thicker layer of a material with a higher dielectric constant is one part of the attack on power. Revolutionary physical FET configurations (figure 12) are another part. These radical departures from a straightforward extension of the present methods strain fabrication technology.

The provision of denser and denser wiring on large chips is the other serious present-day limit to increasing levels of integration. Narrower lines yield increasing wire densities at the cost of increasing wire resistance and slower signal propagation while clock frequencies increase (figure 14). Some relief is obtained from more wire layers, and projections suggest that up to fourteen layers may be needed in the foreseeable future. A search for a low dielectric constant material to replace the SiO_2 that insulates the wires and would reduce transit times is underway.

As in most areas of endeavour though, the admissible physical context is constrained by economics. Physics shows that substantial advantages would accrue to logic circuits by operating them at low temperature, even cryogenic temperatures (Gaensslen *et al* 1977), and were mentioned in section 3.5.5. The advantages have been outweighed by cost factors: a cooling system is needed, portability would be restricted, extra space might be required, questions of reliability would arise and additional demands would be made on customers. The long wire limit could be relieved by the use of wider wires, but that course would increase the size of the chip (equations (5.1*a*) and (5.1*b*)) and the area per device, while the motivation for miniaturization is to decrease the area of devices to maximize the number on a wafer to lower the cost per device.

Physics and chemistry do not pose any insuperable obstacles to realization of the advance technologies envisaged in the ITRS Roadmaps and beyond. The implementation of each new generation of technology will be increasingly costly, however. The limit will be encountered when the cost of producing the advanced chips of the future exceeds the value of the product to the computing enterprise.

References

Anacker W 1977 *Solid State Phys.* (*Inst. Phys. Conf. Ser.* 32) pp 39–55
Ball P 2002 news@nature.com
Bardeen J and Brattain W H 1948 *Phys. Rev.* **74** 230–1
Bardeen J and Brattain W H 1949 *Phys. Rev.* **75** 1208–25
Berger S D *et al* 1991 *J. Vac. Sci. Technol.* **89** 2996–9
Bertrand G *et al* 2004 *Solid State Electron.* **48** 505–9
Blodgett A J Jr 1983 *Sci. Am.* **129** 48–56
Bowden C M, Ciftan M and Robl H R (ed) 1981 *Optical Bistability* (New York: Plenum)
Brinkman W F, Haggan D E and Troutman W W 1997 *IEEE J. Solid State Cicuits* **32** 1858–65
Brumfiel G 2004 *Nature* **431** 621–3
Carlson D M *et al* 1989 *IEEE Trans. Electron. Devices* **36** 1404–13
Carslaw H S and Jaeger J C 1947 *Operational Methods in Applied Mathematics* 2nd edn (Oxford: London)
Christie P 1993 *Proc. IEEE* **81** 1492–9
Christie P 2001 *IEEE Trans. Very Large Scale Integrat. Syst.* **9** 913–21
Christie P, Ennis D J and DiLosa V R 1992 *Synthetic Microstructures in Biological Research* ed I M Schnur and M Peckerar (New York: Plenum) pp 171–82
Christie P and Stroobandt D 2000 *IEEE Trans. Very Large Scale Integrat. Syst.* **8** 639–48
Davidson E E *et al* 1997 *IEEE Trans. Comp. Packag. Manuf. Technol.* **20** 361–74
Davis J A and Meindl J D 1998 *IEEE Circuits Devices* **14** 30–6
Davis J A, De V K and Meindl J D 1998a *IEEE Trans. Electron. Devices* **45** 580–9
Davis J A, De V K and Meindl J D 1998b *IEEE Trans. Electron. Devices* **45** 590–7
Davis J A, Eble J C, De V K and Meindl J D 1996 *Mater. Res. Soc. Symp. Proc.* **427** 23–34
Davis J A *et al* 2001 *Proc. IEEE* **89** 305–24
De V K, Tang X and Meindl J D 1996 *Symp. on VLSI Technology (Tokyo)* paper 20.4, pp 198–9
Deleonibus S *et al* 2000 *IEEE Electron. Devices Lett.* **21** 173–5
Dennard R H *et al* 1974 *IEEE J. Solid-State Circuits* **9** 256–68
Donath W E 1968 *SIAM J. Appl. Math.* **16** 439–57
Donath W E 1979 *IEEE Trans. Circuits Syst.* **26** 272–7
Donath W E 1981 *IBM J. Res. Dev.* **25** 152–5
Doris B *et al* 2002 *IEDM 2002* paper 10.6.1 267-70
Doyle B *et al* 2002 *Intel Technol. J.* 2nd Qtr, 42–53
EE Times 2001 EE Times 12/10/2001
Esaki L 1958 *Phys. Rev.* **109** 603–8
Falster R and Voronkov V V 2000 *MRS Bull.* 28–32
Fischetti M V *et al* 2003 *J. Appl. Phys.* **94** 1079–95
Foerst C J 2004 *Nature* **427** 53–6
Forbes N and Foster M 2003 *Comput. Sci. Eng.* **5** 18–19

Fowler A B *et al* 1966 *Phys. Rev. Lett.* **16** 901–3
Frank D J *et al* 1998 *IEEE Electron. Devices Lett.* **19** 385–7
Frank D J *et al* 1999 *Symp. on VLSI Technology (Tokyo)* pp 169–70
Frank D J *et al* 2001 *Proc. IEEE* **89** 259–87
Gaensslen F H *et al* 1977 *IEEE Trans. Electron. Devices* **24** 218–29
Gardner D S *et al* 1987 *IEEE Trans. Electron. Devices* **14** 633–42
Garmire E (ed) *et al* 1985 *IEEE J. Quantum Electron.* **21** 1339–549
Gentile S P 1962 *Theory and Application of Tunnel Diodes* (Princeton, NJ: Van Nostrand) chapter III.8
Gibbs H M *et al* 1979 *Appl. Phys. Lett.* **35** 451–3
Goldhaber-Gordon *et al* 1997 *Proc. IEEE* **85** 521–40
Gwoziecki R *et al* 1999 *IEEE Trans. Electron. Devices* **46** 1551–60
Heller W R, Hsi C G and Mikhail W F 1984 *IEEE Design Test* **1** 43–51
Heller W R, Mikhail W F and Donath W E 1977 *Proc. 14th Design Automation Conf. (New Orleans)* pp 32–43
Hess K 2000 *Advanced Theory of Semiconductor Devices* (New York: IEEE)
Hergenrother J M *et al* 1999 *IEDM 1999* 75–8
Ho C W *et al* 1982 *IBM J. Res. Dev.* **26** 286–96
Hoeneisen B and Mead CA 1972 *Solid-State Electron.* **15** 819–29
Holmes D P and Baynton P L 1966 *Symp on GaAs* Paper 31, pp 236–40
IBM 2004a *Press Release* 13 February
IBM 2004b *Press Release* 6 December
ITRS 2001 International Technology Roadmap for Semiconductors
ITRS 2003 International Technology Roadmap for Semiconductors
Jacoboni C and Reggiani L 1983 *Rev. Mod. Phys.* **55** 645–705
Keyes RW 1959 *J. Appl. Phys.* **30** 454–5
Keyes R W 1969 *IEEE Spectrum* **6** 36–45
Keyes R W 1975a *Appl. Phys.* **8** 251–9
Keyes R W 1975b *Proc. IEEE* **63** 740–67 (Nikkei Electronics, Nos 125ff) (in Japanese)
Keyes R W 1977 *IEEE Trans. Comput.* **26** 1017–25
Keyes R W 1982 *IEEE J. Solid State Circuits* **17** 1232–3
Keyes RW 1986 *IEEE Trans. Electron. Devices* **33** 863
Keyes R W 1987 *The Physics of VLSI Systems* (Reading, MA: Addison-Wesley)
Keyes RW 1988 *Advances in Electronics and Electron Physics* vol 70 (New York: Academic) pp 159–214
Keyes R W 1989 *Rev. Mod. Phys.* **61** 279–87
Keyes R W 1991 *IEEE Circuits Devices Mag.* **7** 32–5
Keyes RW 1994 *IEEE Circuits Devices Mag.* **10** 28–31
Keyes R W 2001 *Phil. Mag.* B **81** 1315–30
Keyes R W 2002a Unpublished survey
Keyes R W 2002b *IEEE Circuits Devices Mag.* **18** 36–9
Keyes R W and Landauer R 1970 *IBM J. Res. Dev.* **14** 152–7
Kingon A I 2000 *Nature* **406** 1032–8
Koetzle G 1987 *Proc. VLSI and Computers* ed W E Proebster and H Reiner (Los Alamitos, CA: IEEE Computer Society Press) pp 604–9
Kohn W 1957 *Adv. Solid State Phys.* **5** 258–321
Kurzweil R 1999 *The Age of Spiritual Machines* (New York: Viking Penguin) p 169
Landauer R 1961 *IBM J. Res. Dev.* **5** 183–91
Landman B S and Russo R L 1971 *IEEE Trans. Comput.* **20** 1469–79
Laux S E, Fischetti M V and Frank D J 1990 *IBM J. Res. Dev.* **34** 466
Lo S-H *et al* 1997 *IEEE Electron Devices Lett.* **18** 209–11
Lundstrom M 2003 *Symp. on VLSI Circuits* pp 5–8
Magnusson P G 1965 *Transmission Lines and Wave Propagation* (Boston: Allyn and Bacon)
Matisoo J 1967 *Proc. IEEE* **55** 172–81
Meindl J D *et al* 2001 *Science* **293** 2044–9
Meindl J D *et al* 2002 *IBM J. Res. Dev.* **46** 244–63
Miller D A B, Smith S D and Johnston A 1979 *Appl. Phys. Lett.* **35** 658–60
Mizuno T *et al* 2000 *IEEE Electron. Devices Lett.* **21** 230–2
Moore G E 1965 *Electronics* **19** 114–17
Moore G E 1978 *Institute of Physics Conf. Series No* 40
Nguyen T N and Plummer J D 1981 *IEDM Digest* pp 596–9

Nishi Y 1988 *Solid State Technol.* **13** November 115–19
Noyan I C *et al* 1999 *Appl. Phys. Lett.* **74** 2352–5
Noyce R N 1977 *Science* **195** 1102–6
M Orshansky *et al* 2000 *IEEE Trans. Comput. Aided Design* **21** 544–53
Pfeiffer H C and Stickel W 1995 *Microelectron. Eng.* **27** 143–6
Plummer J D and Griffin P B 2001 *Proc. IEEE* **89** 240–58
Price P 1978 *Solid-State Electron.* **21** 9–16
Reid T R 1984 *The Chip: How Two Americans Invented the Microchip* (New York: Simon and Schuster)
Riordan M and Hoddeson L 1997 *Crystal Fire: The Birth of the Information Age* (New York and London: W.W. Norton)
Rose K 1991 *Circuits Devices Mag.* **7** 26–30
Sandberg A 1999 *J Evol. Technol.* **5** 1–34 (p 6)
Seeger K 2002 *Semiconductor Physics: An Introduction* (Berlin: Springer)
Shahidi G G 2002 *IBM J. Res. Dev.* **46** 121–32
Shannon C E 1948 *Bell Syst. Tech. J.* **27** 379–423
Shannon C E 1948 *Bell System Tech. J.* **27** 625–56
Shen Y-L *et al* 1996 *J. Appl. Phys.* **80** 1388–97
Shi Y, Wang X and Ma T P 1998 *IEEE Electron. Devices Lett.* **19** 388–90
Shockley W 1949 *Bell Syst. Tech. J.* **28** 435–89
Shockley W 1961 *Solid-State Electron.* **2** 35–67
Singh R N *et al* 1997 *IBM J. Res. Dev.* **41** 39–48
Solomon P M (ed) *et al* 2002 *IBM J. Res. Dev.* **46** 119–358
Spiller E 1976 *Appl. Opt.* **15** 2333–8
Spiller E *et al* 1992 *Appl. Phys. Lett.* **61** 1481–3
Stein K-U 1976 *Elektrotech. Maschinenbau* **93** 240–8
Stern F 1972 *Phys. Rev.* B **5** 4891–9
Stern F and Howard W E 1967 *Phys. Rev.* **163** 816–20
Sze SM 1981 *Physics of Semiconductor Devices* (New York: Wiley)
Tang T-W, O'Regan T and Wu B 2004 *J. Appl. Phys.* **95** 7990–7
Taur Y and Ning T H 1998 *Fundamentals of Modern VLSI Devices* (Cambridge: Cambridge University Press)
Thompson S Packan P and Bohr M 1998 *Intel Tech. J.* 3rd Qtr, 1–19
Tiwari S *et al* 1997 IEDM **1997** 939–41
Tuckerman B D and Pease R F W 1981 *IEEE Electron. Devices Lett.* **2** 126–9
Wong H S P 2002 *IBM J. Res. Dev.* **46** 133–67
Ziegler J F and Srinivasan G R (ed) 1996 *IBM J. Res. Dev.* **40** 1–128