Physical Limits in Semiconductor Electronics

Robert W. Keyes

This issue of *Science* attests to the revolution created by the advent of silicon microelectronics. The impact of semiconductor electronics in so many areas is a direct result of its record of providing ever-increasing information processing power per unit of cost. The steadily decreasing cost per digital operation has been achieved by fabricating more and more components—diodes, transistors, capacitors, and resistors—on a single piece, or chip, of silicon. A very primitive approximation asserts that the cost of producing a chip is independent of what devices or circuits are created on it. The number of components that can be fabricated on a single chip has increased from 1 to 30,000 over the last 15 years (*1*). Thus, it is not hard to understand that semiconductor electronics has not only made very large digital systems possible, but is also to be found in such cost-conscious markets as those for automobiles and home entertainment.

Progress to larger substrates, or chips, and miniaturization of components and interconnections have provided the keys to placing more components on a chip (*1*). Modern integrated electronics is based on planar technology, which means that devices and circuits are fabricated by operating on a surface of a chip: modifying it chemically by introducing impurities through masks and depositing layers of conducting and insulating substances in selected regions. The size of a chip is limited by defects, the existence of conditions that prevent the fabrication of an operable component in a certain region. A chip must be small enough to make the probability that it contains no defect reasonably large. Improvements in silicon substrates and in processing techniques have steadily reduced the density of defects per unit area and made progress to ever-larger chips possible (*2*). Chips measuring 0.5 by 0.5 centimeter are now common. Miniaturization involves decreasing all of the linear dimensions of structures fabricated on a chip, such as the widths of interconnection lines and the diameters of openings through which impurities are introduced by diffusion or ion implantation. One wonders how long this kind of progress can continue and whether and at what point laws of physics that limit further advances will be encountered. This article describes the search for such physical limits. Physics offers no reason to anticipate that increases in the size of chips through decreases in the number of defects per unit area will not continue. Advances in size are likely to come slowly, though, as major changes in the size of the substrate that must be processed may require major investments in new process tools.

Lithography

Continued miniaturization depends on improvements in lithography, the art of producing patterns in films that can be used as masks in processing—for example, to define patterns on a substrate where material may be altered by etching or by introduction of impurities. Lithography has received a substantial amount of attention as a limit to miniaturization (*3*). The most common form of lithography is photolithography, in which exposure to light changes the properties of a film of a photosensitive compound, allowing it to be selectively removed for purposes of masking. Exposure to light cannot, in practice, produce patterns with minimum dimensions much less than the wavelength of the exposing light. Thus, not too many years ago, dimensions of 1 or 0.5 micrometer were viewed as the limiting "least count" of microstructures.

Exposure of resists with electron beams has, however, emerged as a method that can produce much finer structures than are possible with optical exposure (*4*). In fact, there appears to be no significant fundamental limit to the resolution that can be achieved with electron beams; single atoms can be seen with the electron microscope (*5*). There are, however, practical limits. A very large number of resolution elements must be exposed to form a microstructure. For example, a chip 3 millimeters square is not large in the light of modern technology, but to expose it with 1-μm resolution implies exposure of 10^7 elements. A feasible production process must expose each of the 10^7 elements in a very short time. The highest possible beam current is desirable.

Thus, a presently accepted view of the limits to electron beam lithography runs as follows (*6*). Spherical aberration in electron lenses spreads the electron beam by an amount proportional to the cube of the incidence angle (α) accepted at the target (see Fig. 1). The diameter of the exposed spot is at least

$$D = C\alpha^3/2 \qquad (1)$$

where C is a constant that characterizes the spherical aberration of the electron lens. The current density in the beam is limited by the brightness of the source, B amperes per unit area per steradian. The current in the beam is proportional to the solid angle accepted, or to α^2, and thus large currents and small dimensions conflict according to Eq. 1. The electrons in the beam arrive randomly in time and are distributed in any specified time t according to Poisson's law. To ensure that statistical fluctuations do not leave any resolution element underexposed, it is necessary that the average exposure of an element exceed some minimum number of electrons, N. For example, if $N = 20$, then there is a probability of .006 that an exposure element receives only 10 electrons. Asking that the probability of an exposure less than half the average be smaller than 10^{-14} requires that N be something like 200. These limits on the beam current and on N imply that there is a minimum exposure time per spot, which depends on the spot diameter. The relation between the exposure time and D is shown in Fig. 1. As the cost of an electron beam exposure tool must be spread over a large number of components, Fig. 1 implies an economic limit on the applicability of electron beam lithography. Only very rough values of the parameters involved are needed to quantify this limit because of the very strong dependence on D. For example, taking the cost of the exposure system as \$100 an hour and the value of the exposed silicon as \$10 per square centimeter, an exposure rate $R = 3 \times 10^{-3}$ cm^2/sec is

The author is a research staff member at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598.

required, and it is found that $D > 5 \times 10^{-6}$ cm.

It would be a mistake to regard such a limit on electron beam exposure technology as "fundamental," however. One possible escape might be found in the suggestion that it is possible to completely eliminate the effects of spherical aberration (*7*). Clearly, the invention of brighter electron sources could directly affect the exposure time. The potential of electron beam lithography is far from exhausted.

There is, though, another very important limit to the resolution of electron beam lithography as practiced today. Electrons pass through the resist, are scattered in the underlying silicon, and may emerge some distance away from the incident beam to expose the resist there (*4*). The range of a typical 25-kilovolt electron in silicon is about 3 μm, so there is a nonnegligible degradation of resolution by this backscattering effect. The effect can be diminished by reduction of the energy of the exposing electrons. Decreasing the electron energy, however, increases the chromatic aberration of the electron lenses, introducing another source of spreading of the spot.

The large-angle scattering that causes electrons to be returned from the substrate to the photoresist could be avoided by exposing with more massive particles—protons or other ions. The low brightness of ion sources has prevented the useful exposure of resists with ions to date. However, the development of more intense ion sources may be another path to pushing back the present limits on lithography.

X-ray lithography is an attempt to take advantage of the very high resolution that is possible in principle in writing with electron beams without being restrained by the long times that are needed to expose a chip by sequential scanning (*3*, *8*). A high-resolution mask made by writing with an electron beam can be used to expose photoresist on a silicon substrate with x-rays; the short wavelength of the x-radiation, 1 to 100 angstroms, preserves the high resolution of the mask. Since a mask can be used to expose a large number of substrates, a long time and low current can be used to achieve high resolution in the preparation of the mask. The resolution attainable with x-rays is limited by a different effect: a secondary electron is emitted when an x-ray photon is absorbed, and the secondary electron has a range of one to several hundred angstroms, and exposes the resist in an area with this radius.

Thus, it must be emphasized again that the limits of particle beam lithography just described are in no way fundamental; they involve practical and economic considerations. Indeed, fabrication of metal conductors 80 Å wide has been reported (*9*). One can be confident that the march of lithography to smaller dimensions will continue. Physical limits to miniaturization must be sought elsewhere.

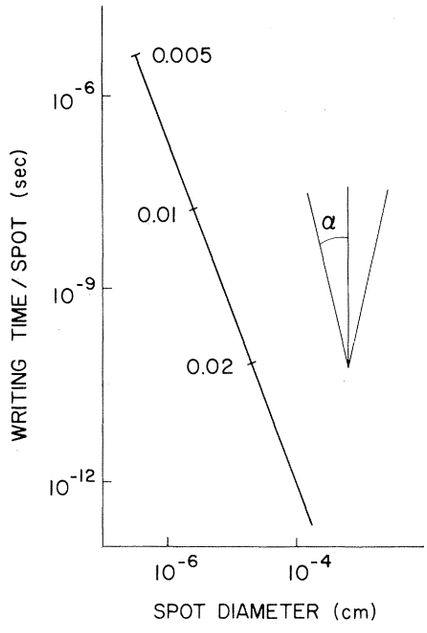

Fig. 1. Minimum time needed to expose a picture element of an electron resist with a focused electron beam as a function of element diameter (*6*). The limit is calculated with the following parameters, which are typical of modern electron beam systems: spherical aberration, $C = 5$ cm; source brightness, $B = 10^6$ amp/cm²-steradian; minimum number of electrons per spot, $N = 200$. The values of α, the beam angle accepted in Eq. 1, are indicated on the limiting line.

Hot Electrons and Breakdown

Another problem of miniaturization of digital devices arises from the fact that voltages cannot always be reduced in proportion to dimensions, so that electric fields in semiconductor devices increase with decreasing size. Voltage levels in digital systems must be large enough to provide a clear distinction between "on" and "off" states of a device, or between informational 0's and 1's. Digital signals must be standardized; a 0 must look like a 0 and a 1 must look like a 1, regardless of their source (*10*). In its application to electrical signals, this means, for example, that if it is intended that a voltage V_1 represents a 1, then a circuit that receives a voltage V_1' less than V_1 must transmit a voltage that is closer to V_1 than V_1' is. Otherwise, there would be a steady deterioration of voltage level as information passes from stage to stage through the processor. When it is realized that a similar statement must apply to voltages near V_0 that are intended to represent 0, it is clear that the standardization requires a nonlinear response.

Voltages applied to electrical devices change the potential of electrons in some spatial regions with respect to those in other regions. Now, electrons are dispersed in energy by an amount of approximately κT by thermal agitation (κ is the Boltzmann constant and T is absolute temperature). Electrical voltages that are small relative to $\kappa T/q$ are thus just a small perturbation on the steady energy distribution of the electrons and produce only linear effects (q is the charge of an electron; $\kappa T/q = 0.025$ volt at $T = 300$ K). Nonlinear effects can be produced by voltages that are large relative to $\kappa T/q$.

This scale of nonlinearity is most perfectly exemplified by the ideal p-n junction, in which the current depends on the voltage as

$$i \propto \exp(qV/\kappa T) - 1 \qquad (2)$$

The junction characteristic as expressed by Eq. 2 is a practical limit to electical nonlinearity. A similar scale even seems to be applicable in biology: neuron voltages are a few times $\kappa T/q$. Various lines of thought agree in establishing $\kappa T/q$ as a practical minimum voltage scale for the production of nonlinear electrical effects. The word *scale* implies that the attainment of the very large nonlinearities needed for reliable logical operation in the presence of such disturbing influences as cross talk, environmental fluctuations, and component variability will require that voltages a great many times $\kappa T/q$ be used in real circuits. Questions such as the necessary degree of reliability and the acceptable amount of component variability lie outside the realm of quantitative physical science, and the only statement that can be made about the voltage is that it must be large relative to $\kappa T/q$.

Thus, voltages are relatively independent of size. Nevertheless, all dimensions of a structure, including such internal device dimensions as base widths in bipolar transistors and the thicknesses of the depletion layers between p- and n-type regions of silicon, are decreased as miniaturization advances. Electric fields and high-field phenomena such as hot electrons and dielectric breakdown grow in importance with decreasing dimensions and form a limit to miniaturization.

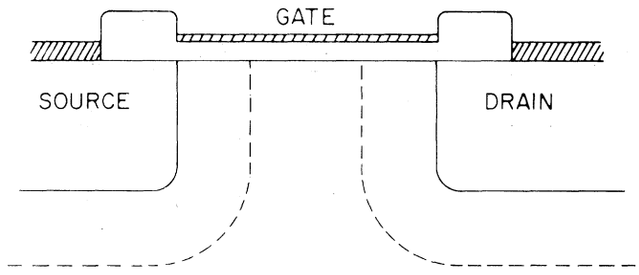

Fig. 2. Structure of an insulated gate field-effect transistor, showing the extent of the depletion regions associated with the source and drain junctions.

A quantitative version of one such limit has been formulated (*11*). Consider the field-effect transistor (FET) shown in Fig. 2. The depletion regions associated with the *p-n* junctions of the source and drain electrodes are shown. The source-drain distance must be greater than the sum of the widths of the depletion layers in order that the gate can exercise control of the conductance along the surface. The width of the depletion layers can be reduced by increasing the doping level of the substrate silicon. The source and drain can then be placed closer together and the transistor made smaller. However, the electric fields in the depletion layers will be increased at any given applied voltage, and the voltage limit of the transistor will be decreased. The same is true of the depletion layer that is formed between the surface and the bulk substrate silicon when a conductive surface charge is induced by a voltage applied to the gate. In this case, the increase in doping results in an increase in electric field in the oxide. Thus, both junction breakdown and oxide breakdown limit the miniaturization of FET's. Since breakdown of silicon and of SiO_2 is a well-studied subject, it is possible to calculate for each voltage the maximum doping level that can be used and the smallest permissible source-drain separation. A calculation of this type is shown in Fig. 3 (*11*). The results presented show that breakdown in the oxide insulator is the limiting factor in the miniaturization of FET's. However, curvature of a junction and nonuniform doping profiles affect the breakdown voltage but are difficult to take into account quantitatively, and it has also been suggested that breakdown at the drain-substrate junction is the more serious limitation.

Related models can be developed for bipolar transistors. The basic requirement is that the base shall not be completely depleted or suffer "punch-through," and that the junctions shall not break down. The punch-through is controlled by heavy doping, which, however, decreases the breakdown voltage of the junctions and the voltage at which the transistor can operate.

The electron temperature, a rough measure of the average electron energy, rises above the lattice or thermal equilibrium temperature at electric fields much smaller than those needed to cause avalanche breakdown. The electrons become "hot." Some of the hot electrons have enough energy to pass from the silicon into the SiO_2 insulator on the surface and become trapped there, simulating a potential applied to the gate and changing the properties of the surface (*12*). Such effects are noticeable in both field-effect and bipolar transistors (*12*, *13*). In FET's hot electrons produced by electric fields in the channel at the surface change the threshold voltage when they escape into the insulating SiO_2. Hot electrons that are produced in the region close to the intersection of a *p-n* junction with the silicon surface increase the leakage current across the junction along the surface when they escape into the SiO_2 and degrade the performance of bipolar transistors. Quantitative studies of such effects in FET's are available and show that undesirable changes in characteristics occur rapidly if the fields in the channel are only slightly greater than 10^4 volt/cm (*12*). Although much more experimental information is needed, that which is available suggests that these hot electron effects will limit reduction of the FET "length," the source-drain separation, to something like ¼ μm. Information concerning the hot electron degradation of *p-n* junctions is even more scanty.

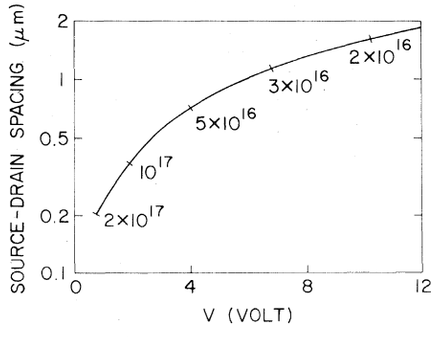

Fig. 3. An example of the limits imposed on a metal oxide semiconductor field-effect transistor as shown in Fig. 2 by breakdown in the SiO_2 insulator (*11*). Heavy doping reduces the widths of the depletion layers illustrated in Fig. 2; the impurity concentrations needed to achieve particular source-drain spacings are indicated. Breakdown in the oxide then limits the supply voltage of a simple inverter to the value shown.

Power Dissipation

An issue closely connected to miniaturization of devices is density. Density means the number of components or number of circuits per unit area. High density is desirable for several reasons. The time taken for a signal to propagate from one circuit to another is reduced as the density is increased. Generally speaking, too, the cost of producing a structure is reduced as the area that it occupies decreases. The most serious limitation on the density of logical processing circuits, if the necessary lithographic techniques have been mastered, is set by power dissipation.

Thus, we turn to another set of problems that are associated with the processing of digital information. Although it seems possible to construct nondissipative information processing systems in principle (*14*), in practice known kinds of information processing devices dissipate power, converting it to heat. There are very basic reasons for the dissipation of power in logical processors. Most important is the irreversible nature of information processing in a general-purpose computer. Information is represented in physical degrees of freedom, such as the charge on capacitors. In information processing different starting points may lead to the same result; the memory of the initial conditions is lost, and there is no way to reverse the physical processes employed to return to the initial state of the system. The logical irreversibility implies physical irreversibility and dissipation (*15*). Although in principle one can preserve a complete history of the computer operations, so that they could all be reversed and all of the energy used eventually recovered (*14*), this is not practical in a modern electrical general-purpose computer. In such a computer, logical processing stages perform their function, pass the information on to a next stage, and are restored to a state in which they are ready to process information again, losing their memory of what was done in the preceding step. Thus, each logical circuit converts a certain amount of electrical power to heat. The heat must eventually be transferred to some fluid, commonly air or water, that carries it out of the system. The heat usually leaves the semiconductor chip itself by conduction across a surface to another solid and is transferred to a fluid across some larger

Table 1. Parameters achieved in packaging an air-cooled computer processing unit with 100,000 circuits (*17*).

Parameter	Value
Average power per circuit	65 mwatt
Circuit delay on chip	0.6 nsec
Dissipation density	0.4 watt/cm^2
Circuits per chip (average)	57
Chip density	0.12 cm^{-2}
Circuit density	70 cm^{-2}
Signal velocity	10^{10} cm/sec

area. There is, however, a finite limit to the rate at which heat can be transferred across any surface, and therefore, a limit to the density at which components that dissipate power can be packed on a planar surface.

At one level the problem of power dissipation is intimately bound to miniaturization. Energy is dissipated in the successive charging and discharging of capacitances that are part of semiconductor devices. Capacitance is dimensionally dielectric constant times a linear dimension and so must scale in proportion to dimension. The energy in the capacitor is proportional to the square of the voltage, V, applied to it. If a circuit element is operated at a rate $1/t$, then the power dissipated in it has the form

$$P = \Lambda V^2 L/t \qquad (3)$$

where Λ is a proportionality constant with the dimensions of a dielectric constant and L is a length parameter, say the square root of the area per element. Then limitation of the dissipation per unit area to a value Q which can be removed by heat transfer means that

$$P/L^2 < Q \qquad (4)$$

Equations 3 and 4 immediately lead to limiting relations between S and t and between P and t. I believe that circuit considerations, the need to send logic signals from one component to another efficiently over transmission lines, limit Λ to something like 1000 times the dielectric constant of free space (*16*). The limits that result from this assumption are shown in Fig. 4.

This thermal limit can be qualitatively understood in the following way. The energy supplied to the circuit is used to charge capacitances. Capacitance is dimensionally electric permittivity times a linear dimension. As circuits are miniaturized, the energy per operation therefore can be reduced in proportion to linear dimension. The number of circuits per unit area, however, increases inversely as the square of linear dimension, and the power per unit area thus increases inversely to dimension at a constant operating rate. Eventually the power density becomes greater than the maximum rate of removal Q; this is the density represented by the solid line in Fig. 4.

A limit also arises from the finite time taken for a signal to propagate from one circuit to another. The higher the circuit power, the farther apart the circuits must be placed to provide area for the transfer of heat, and the longer the propagation time. An estimate of the propagation time limit is also given in Fig. 4.

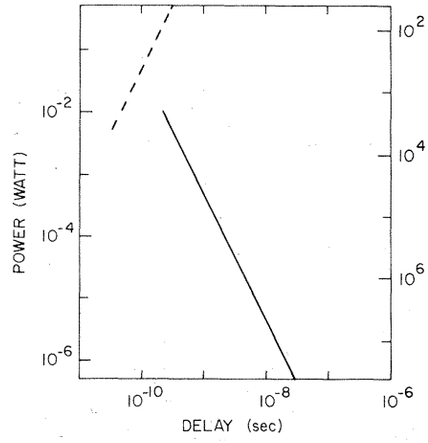

Fig. 4. The thermal limit on logic circuits following from Eqs. 3 and 4, $Pt^2 > (\Lambda V^2)^2/Q$ (solid line), calculated for $\Lambda = 10^{-10}$ farad/cm, $V = 1$ volt, $Q = 20$ watt/cm^2 (*16*). The propagation time limit, $Pt^{-2} > c_1^2 Q/m^2$ (dashed line), is also shown for the parameters: velocity of transmission, $c_1 = 5 \times 10^9$ cm/sec; length of transmission divided by square root of area per circuit, $m = 10$. The number of circuits per square centimeter corresponding to the power is shown on the right.

Packaging

Actually, propagation times become more important at a higher level of packaging. The preceding discussion, in which the circuits are regarded as densely packed on a substrate, is applicable to semiconductor chips. The chips are mounted on larger carriers, frequently made of ceramics, when they are assembled into a larger system. The chip carriers and the substrates or boards on which they are mounted provide for mechanical and electrical attachment of the chip and contain the wiring matrix that interconnects the chip and the power distribution conductors (*17*). Simple mechanics may limit the closeness with which chips may be placed on a large substrate, through problems such as making the chip carrier conveniently replaceable for fault correction or design changes. On the other hand, heat dissipation can also limit the density of chips. The rate at which heat can be removed from boards is much less than the rate at which it can be removed from chips. Chips are cooled by conduction through an interface to some larger solid, and the heat current is limited by the thermal resistance of the solids and the interface; values of 20 watt/cm^2, as used in Fig. 4, can be attained. At the board level, however, heat is transferred to air, which carries it out of the system; 1 watt/cm^2 is difficult to achieve over a large area. Thus the chips may have to be widely spaced to provide for cooling, and the chip-to-chip propagation time can become an important limit on performance.

The delay encountered by a signal that must pass from one chip to another is the sum of two parts, the on-chip circuit delay and the chip-to-chip propagation delay. As just described, if the circuits are driven at a higher speed more power is used, and the chips must be placed farther apart, increasing the propagation delays. The art of mounting the chips on a substrate in such a way that they can be interconnected and cooled is called packaging. Packaging problems comprise one of the most severe limitations on the performance of modern high-speed integrated logic. The many functions that must be taken into account in the design of a package are:

1) Provide mechanical support and attachment.

2) Provide electrical connection to chip.

3) Transform chip contact dimensions to mechanically pluggable dimensions.

4) Contain wiring matrix for chip interconnections.

5) Contain power distribution network.

6) Receive heat from chip and deliver it to a fluid.

7) Protect the semiconductor from the environment.

The emergence of fast devices has required that interconnection lengths be reduced to take advantage of the device speeds. Packages must be made smaller. Thus problems of heat dissipation and mechanical access to the chip to replace failed units or to incorporate design changes have become much more severe. Some of the parameters that have been achieved in a recent large computer are presented in Table 1 (*17*).

Memory

The preceding discussion has been implicitly oriented toward logic circuits.

The same considerations apply, however, to memory, with certain differences in detail. Like logic, memory is digital, and reading, writing, and moving information requires the application of voltages large enough to produce nonlinear effects and is a source of the power dissipation that accompanies the large voltages. Magnetic memories, cores and bubbles, are "nonvolatile" and can preserve information for a long time without dissipating any power. To a degree, semiconductor memories that store information in the charge on a capacitor share the same characteristic, although the stored charge slowly leaks away and must be restored periodically. Active use of these memories, though, involves reading the information through some kind of matrix addressing, or transferring it from site to site until it arrives at some place where it may be read and transmitted to other parts of a system. These nonlinear operations produce heat and subject memory devices to the thermal limitations described above. The principal differences between memory and logic are that memory elements often enjoy a low duty cycle; that is, they are infrequently actively involved in an operation. They are also much simpler physically than logic gates, and so a memory cell occupies less area on a silicon chip than a logic circuit. The result of the lower dissipation of memory cells and their smaller areas is that their dissipation per unit area is not much different from that of logic circuits. Memory is subject to essentially the same thermal limitations as logic, as illustrated in Fig. 4.

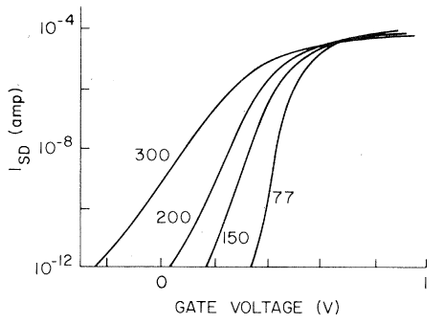

Fig. 5. Source-drain current of a silicon field-effect transistor as a function of gate voltage and temperature, illustrating the sharpening of characteristics as the temperature is lowered (*18*). The curves are labeled with the temperature in Kelvins.

Alternatives

The problems and limitations of silicon electronics have led to a search for new directions that might relax the limits described. Several possibilities have received attention. One of these is abandoning semiconductor electronics in favor of superconducting electronics. A second involves operating silicon devices at low temperatures, typically 77 K, the boiling temperature of liquid nitrogen (*18*). Physics immediately suggests many possible advantages to operating silicon devices at low temperatures. Most obvious is the increase in conductivity; the scattering of electrons by lattice vibrations decreases as the temperature is lowered. The effect of the decreased resistance is probably most important in the metallic interconnections of integrated circuits, since conductivity in the silicon itself is primarily limited by scattering by impurities rather than lattice vibrations. Low temperatures also promise to reduce the power dissipated. As explained above, the sharp transitions between states of a device that are the essence of digital electronics require that voltages large compared to the thermal voltage, $\kappa T/q$, must be used. Thus, the lower the temperature, the lower the voltage required in switching circuits. Figure 5 shows dramatically how the sharpness with which an FET is turned on by the gate potential increases as temperature is decreased. The power dissipation is expected to be proportional to the square of the voltage (*18*). Thus, these rough considerations suggest that 16 times less power will be required at 77 K than at 300 K.

Low temperature may lead to even greater decreases in power in memories in which information is stored in charge on a capacitor. The charge is retained by opening a transistor switch; it gradually leaks away as reverse current through some *p-n* junction and must be periodically refreshed. The reverse current of junctions decreases rapidly with decreasing temperature, so refreshing is needed less frequently at low temperatures. In some circuits the power dissipated in the refreshing operations is an appreciable part of the total dissipation and may be reduced by a large amount (*18*).

The decreased power dissipation means that less area is required to allow for removal of heat. The lowered resistance of metals allows connecting lines to be made narrower and also decreases areal requirements. Thus, significantly higher device densities promise to be achievable at low temperatures. The cost of low-temperature semiconductor electronics should be substantially smaller than that of conventional semiconductor devices. The unanswered economic question is, How large a system is required to justify the added investment in refrigeration by reduction in cost and increase in performance of semiconductor devices?

Another new direction in electronics is the development of digital integrated circuits in other semiconductors. In fact, this direction is not really new, as hope that semiconducting compounds of group III and group V elements would replace silicon and germanium in transistor applications has been entertained ever since the discovery of the compounds more than 20 years ago (*19*). A large share of the enthusiasm for the III-V semiconductors was based on their high electron mobilities, which are in many cases 2 to 20 or more times greater than mobilities in silicon. The search for physical limits reveals one reason why the III-V semiconductors have failed to displace silicon; the limits have little to do with mobility, but concern such things as avalanche breakdown fields and thermal problems. It is necessary to have an energy gap that is large enough to prevent intrinsic conductivity, the thermal excitation of electrons from the valence band to the conduction band, from interfering with device operation. The favorable features of a semiconductor are a high energy gap, which allows the temperature of a component to rise a certain amount and decreases the stringency of the cooling requirements, and high breakdown fields, which are associated with a large gap. A semiconductor with an energy gap smaller than that of silicon is unlikely to play an important role in modern electronics.

An even more important reason for the dominance of silicon, however, was the rapid establishment of a feasible processing technology for silicon. The ease with which an SiO_2 layer can be formed on a silicon surface and the remarkable properties of such a layer as an insulator, a diffusion mask, and a neutralizer of undesirable surface effects are unmatched by any phenomena in the III-V compounds.

Nevertheless, the III-V compounds have had an impact on electronics. The advantages of gallium arsenide as a material for transistors have long been known. It has a larger energy gap than silicon, slightly larger breakdown fields, and a much higher electron mobility. The more difficult technology of GaAs has at last been mastered, and it has become the superior material for microwave transistors (*20*). Field-effect transistors are used to take advantage of the high electron mobility; in bipolar transistors both electon and hole mobilities are important. One other semiconductor, indium phosphide, appears to share the advan-

tages of GaAs, but its technology is still in a more primitive stage.

The technology that makes GaAs microwave transistors possible is now being extended to the fabrication of integrated microwave circuits. One wonders when the application of the proved high-speed microwave capabilities of GaAs to digital circuits will begin. Exploratory attempts have already been made (*21*). A realistic view suggests, however, that the invasion of large-scale digital applications by GaAs will be much more difficult than the conquest of the microwave field. The reason is that one or a few microwave devices with superior frequency response can extend the bandwidth of a system and have great economic value. On the other hand, digital systems use thousands to millions of devices, and low-cost fabrication is essential. Fast devices are less important because the speed of a system is also limited by the delays in the package. The highly developed and versatile silicon technology, optimized by many years of experience, will not be displaced easily.

Summary

Although the limitations of the methods of lithography in use at a particular time are easily recognized and attract substantial attention, experience shows that technological ingenuity keeps pushing them to ever-smaller dimensions. There seems to be no fundamental reason to expect that lithographic limits will not continue to recede. The limits to the advance of miniaturization are to be found in the ability of materials to withstand high electric fields and in the ability of packaging technology to remove heat from active components and provide for power distribution, signal interconnection, and flexible mechanical assembly.

References and Notes

1. G. E. Moore, in *Technical Digest 1975 International Electron Devices Meeting* (IEEE, New York, 1975), pp. 11–13.
2. J. L. Buie, in *1976 WESCON Technical Papers* (Western Electric Show and Convention, El Segundo, Calif., 1976) paper 23/2.
3. G. Pircher, in *Solid State Devices, 1975* (Société Française de Physique, Paris, 1975), pp. 31–72; J. T. Wallmark, in *Solid State Devices, 1974* (Institute of Physics, London, 1975), pp. 133–167.
4. G. R. Brewer, *IEEE Spectrum* **8** (No. 1), 23 (1971).
5. A. V. Crewe, J. Wall, J. Langmore, *Science* **168**, 1338 (1970).
6. T. E. Everhart, in preparation.
7. N. W. Parker, S. D. Golladay, A. V. Crewe, in *Scanning Electron Microscopy/1976*, O. Johari, Ed. (IIT Research Institute, Chicago, 1976), part 1, pp. 37–44.
8. H. I. Smith and S. E. Bernacki, *J. Vac. Sci. Technol.* **12**, 1321 (1975); R. Feder, E. Spiller, J. Topalian, *ibid.*, p. 1332.
9. A. N. Broers, W. W. Molyen, J. J. Cuomo, N. D. Wittels, *Appl. Phys. Lett.* **29**, 596 (1976).
10. A. W. Lo, *IRE Trans. Electron Comput.* **EC-10**, 416 (1961).
11. B. Hoeneisen and C. A. Mead, *Solid State Electron.* **15**, 819 (1972); *ibid.*, p. 981.
12. S. A. Abbas and R. C. Dockerty, *Appl. Phys. Lett.* **27**, 147 (1975).
13. D. R. Collins, *ibid.* **13**, 264 (1968).
14. C. H. Bennett, *IBM J. Res. Dev.* **17**, 525 (1973).
15. R. Landauer, *ibid.* **5**, 183 (1961); *Ber. Bunsenges. Phys. Chem.* **80**, 1048 (1976).
16. R. W. Keyes, *Proc. IEEE* **63**, 740 (1975); *IEEE J. Solid State Circuits* **SC-10**, 181 (1975).
17. For a detailed description of the packaging of a modern computer see R. J. Beall, in *1974 INTERCON Technical Papers* (IEEE, New York, 1974), paper 18/3.
18. F. H. Gaensslen, V. L. Rideout, E. J. Walker, in *Technical Digest 1975 International Electron Devices Meeting* (IEEE, New York, 1975), pp. 43–46; F. H. Gaensslen, V. L. Rideout, E. J. Walker, J. J. Walker, *IEEE Trans. Electron Devices*, in press.
19. D. A. Jenny, *Proc. IRE* **46**, 959 (1958).
20. C. A. Liechti, *IEEE Trans. Microwave Theory Tech.* **MTT-24**, 279 (1976).
21. R. L. Van Tuyl and C. A. Liechti, in *1974 IEEE International Solid State Circuits Conference Digest of Technical Papers* (Winner, New York, 1974), pp. 114–115.

Solid-State Electronics: Scientific Basis for Future Advances

J. A. Giordmaine

The leaps forward in conceptual understanding, the new device principles, the advances in analytical technique, and the achievements in materials preparation that make up the scientific basis for the electronics revolution described in this issue can be readily identified, in retrospect. A list of such contributions would certainly include crystal structure analysis based on x-ray and electron diffraction, the explanation of conductivity in terms of the quantum theory of solids, the growth of ultrapure single crystals of electronic materials with controlled doping, the concept of a semiconductor amplifier, the invention of high-frequency oscillators based on stimulated emission, and the demonstration of quantum tunneling devices.

As Niels Bohr remarked, however, it is very difficult to predict, especially the future. Any attempt to identify a scientific basis for future advances is limited by a number of constraints. The lead time between scientific discovery and utilization in solid-state technology is at least 5 years, frequently more than enough time for the economic factors determining utility to have changed beyond recognition. The important scientific advances rarely emerge in a completely scheduled or planned pattern and not always in response to a perceived need. Progress often occurs in stages of abrupt change in the conceptual basis of the field, following a steady if undramatic accumulation of essential background understanding. Today the challenges of large-scale integration and the pace of change of the technology are making heavy demands on resources previously dedicated to longer-term research. Fundamental research of all kinds is carried on in a climate of increasingly critical scrutiny and diminishing real support.

In spite of the maturing of semiconductor technology and the uncertainties and vulnerability of the research enterprise, I believe that there is considerable ground for optimism about the prospects for continuing innovation in solid-state electronics. From a fundamental point of view, our present degree of control over electrons and their motion in solids may be compared with our ability to manipulate light at the beginning of the 19th century. We are now able to exploit behavior dependent on electron density and current flow, analogous to quantity of optical radiation and radiant intensity. The device utilization of the wave nature of electrons—making specific use of amplitude, phase, and coherence as in the case of light in diffraction, interference, holography, and the laser—has barely begun with the discovery of the Josephson effect.

From a nearer-term point of view, the exponential growth of the scale of in-

The author is director of the Solid State Electronics Research Laboratory, Bell Laboratories, Murray Hill, New Jersey 07974.