Problems with Ultraminiaturized Transistors

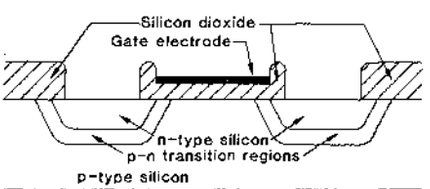
Making extremely small structures is only part of the challenge; new physical phenomena plague microcircuits as components shrink

Transistors are the basic electrical components of microelectronic or integrated circuits. In digital circuits of the type forming the heart of computers, modern telecommunications systems, and "intelligent" office and consumer electronic products, transistors mainly act as ultraminiature switches directing the flow of electricity around the circuit (see box). From the time of the invention of the integrated circuit just over 20 years ago to the present, the number of transistors engineers can pack into one microcircuit has almost doubled every year. The most complex computer memory circuit in production holds more than 65,000 transistors on a thumbnail-sized chip of silicon. This relentless march of increased complexity has been made possible by the ability of engineers to make transistors smaller and smaller, to the point where each occupies an area of only a few hundred square micrometers of silicon. As the size of transistors shrinks, a fundamental question about size gains importance. "Does a small transistor act just like a big one?"

This is the fourth in a series of Research News articles on microelectronics.

In the upcoming generation of microelectronics—called very large scale integrated circuits (VLSI)—the answer is no. Many phenomena that had negligible effect in previous generations of microcircuits are too significant to be overlooked in VLSI-sized transistors. Ultimately, if other factors do not intervene to halt the advance of miniaturization, transistors may no longer function in the usual way, and a new class of device may have to be invented. A National Research Council report released last summer called for the establishment of several new research centers and a generally enhanced level of federal support for the study of phenomena that become important in transistors with dimensions of 1 micrometer or smaller.* The program envisioned by the NRC's Solid State Sciences Committee would cost about $30 million for the first 3 years of operation and would be aimed at exploratory research of the type that the semiconductor industry is unlikely to carry out.

Some of the effects of miniaturization do not, in fact, reflect the appearance of new phenomena but are simply the result of making structures small. The most common form of integrated circuit uses the metal-oxide-semiconductor (MOS) transistor technology. The electrode that turns the device on or off is called the gate electrode. In the first MOS transistors, the gate electrode and the conducting strip that connected the gate to the rest of the circuit were made from a thin film (1 micrometer or so) of aluminum. For reasons having to do with ease of fabrication and more accurate alignment of the electrode with the transistor, engineers have replaced the aluminum with polycrystalline silicon heavily doped with impurities to make it highly conductive. However, the ability of a wire to conduct electricity depends not only on its conductivity but on its cross sectional area. Nearly every speaker or review paper on the challenges of VLSI points out that as the width of polycrystalline silicon conducting strips in integrated circuits decreases to 1 micrometer or less, the resistance of the strip becomes unacceptably large and reduces the performance of high-speed circuits.



Schematic diagram of an MOS transistor. For explicitness, the silicon substrate is assumed to be p-type material. Current flows through the device by way of the electrodes over the two n-type regions at each end of the transistor. The central p-type segment is separated from its electrode (called the gate electrode) by a thin layer of silicon dioxide. With no voltage on the gate electrode, the transistor does not conduct electricity; it is turned "off." But a positive voltage applied to the gate electrode causes those few electrons in the p-type silicon to collect under the electrode, whereas holes are driven away. This forms a "channel" of n-type material, allowing current to flow through the transistor, and the device is "on." This particular form of MOS transistor is called an n-channel device. [Drawing by Eleanor Warner]

An important specification for computer memories, for example, is the access time—that is, the time between asking for a bit of information in a certain location in the memory and receiving it. It is desirable to have as short an access time as possible. Last December at the Washington, D.C., International Electron Devices Meeting, Dean Toombs of Texas Instruments showed calculations demonstrating that the access time for a certain type of memory chip grew smaller as the width of polycrystalline silicon conductors decreased to a dimension of about 1.5 micrometers. At smaller sizes still, the access time began to go back up. One potential solution to the problem is the replacement of polycrystalline silicon with more conductive materials, such as refractory silicide compounds or even multilayer polycrystalline silicon-metal silicide structures.

A second limitation arising purely from making transistors smaller is the problem of computer mistakes due to the passage of high-energy nuclear particles through a microcircuit. These malfunctions are called soft errors because, as soon as the particle has passed through, its effect vanishes and the microcircuit works correctly again. Two sources of such soft errors have been identified. The first is radioactive decay of uranium or thorium that is present in trace quantities in the ceramic package that houses an integrated circuit chip, as first reported by engineers from the Intel Corporation 2 years ago. The decay events produce alpha particles (helium nuclei) that leave behind a wake of electrical charge as they ionize the silicon atoms during their passage through the microcircuit. The second source, reported late last year by James Ziegler of IBM's Yorktown Heights laboratory and William Lanford of Yale University, is cosmic rays. Cosmic rays interact with the nuclei of molecules in the upper atmo-

*Solid State Sciences Committee, National Research Council, *Microstructure Science, Engineering, and Technology* (National Academy of Sciences, Washington, D.C., 1979).

0036-8075/80/0613-1246$01.00/0

sphere to produce a spectrum of elementary particles. Ziegler and Lanford calculated that most important for state-of-the-art computer memories are neutrons, which interact with silicon atoms to produce alpha particles.

In computer memory microcircuits of the type affected by nuclear particles, information is stored in the form of charge on a capacitor. An MOS transistor simultaneously serves both as a switch to allow the capacitor to charge or discharge and as the capacitor; thus the structure is known as a one-device memory cell. (The first random access memory microcircuits required several transistors to store one bit of information. Devising the one-device cell is an example of the type of "circuit cleverness" that has along with miniaturization permitted successive generations of memory chips to store more and more information.) In the newest memory chips that can store 65,536 bits of information, only about 1.5 million electrons constitute the charge that is stored in the capacitor. The passage of nuclear particles, whose charged wake consists of a comparable number of electrons, therefore can cause the capacitor to change its state, from uncharged to charged for example. An attempt to read the content of such a memory cell would result in retrieval of incorrect information. When the memory is tested, however, the chip is found to be operating correctly because testing involves changing the contents of the memory cells, which removes the effect of the nuclear particle.

Problems traceable to radioactive decay first appeared in random access memories storing 16,384 bits (the so-called 16K RAM). Some manufacturers, including Intel, had to redesign their memory chips to account for soft errors due to alpha particles. The problem is more severe with 64K RAM's of the type mentioned in connection with cosmic rays. The effect causes a chip to err once in 1 million hours of operation. Errors can be reduced by "error correcting codes," but this remedy entails a penalty because some fraction of the memory chip must be used to store the codes. Redesigning the cell geometry or using different packaging materials also helps.

A phenomenon that illustrates the onset of new effects due to miniaturization is the problem of heat dissipation. From the point of view of thermodynamics, transistor switching is an irreversible process, and the energy driving the switch (product of total charge moved through the switch and the driving voltage) is lost as heat, which must be removed from an integrated circuit chip because semiconductor devices are extremely sensitive to temperature changes and tend to fail at high temperatures. At small enough dimensions, transistors crammed together on an integrated circuit chip produce so much heat that it cannot all be removed.

"Scaling" is the primary approach to miniaturization. Among the important characteristics of a transistor are its length and width, its thickness, the concentrations of doping impurities in the various parts of the device, the voltage at which switching occurs, and the temperature. As James McGroddy of IBM explains, engineers have devised formulas describing the operation of a transistor that are in dimensionless form; that is,

What Is a Transistor?

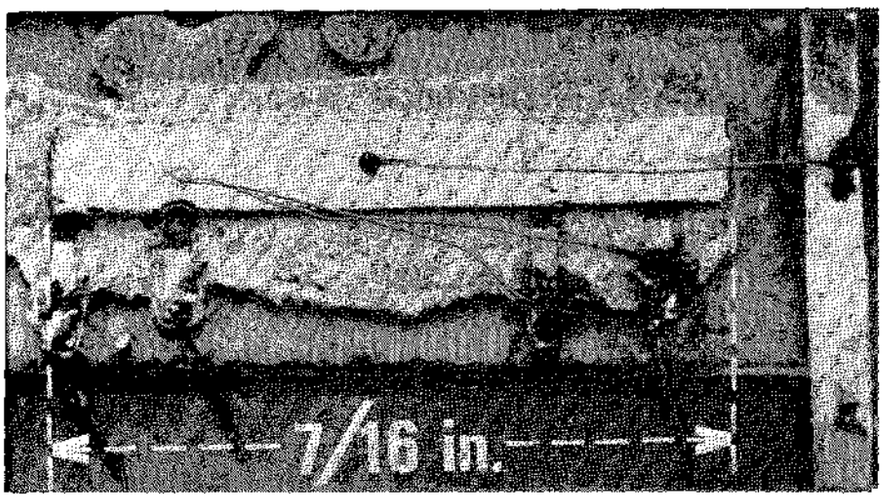
All commercially available integrated circuits are made from the semiconductor silicon. Transistors, the most common circuit element in microcircuits, are made from silicon by the introduction of minute quantities of impurities that determine the electrical properties of the host material. By precisely controlling both the location and the concentration of impurities (called dopants), engineers can build up the transistor structure.

Doping impurities come in two types. The first adds free electrons to the silicon, converting it from a near insulator to a conductor of electricity (although the conductivity is much less than that of a metal). The second type removes electrons from the bonds keeping the silicon atoms in the solid, leaving behind electron vacancies or "holes." The holes behave like positively charged carriers of electricity and thus the second type of dopant also raises silicon's electrical conductivity. Silicon that conducts electricity by way of free electrons is called *n*-type, whereas material that conducts by way of holes is called *p*-type.

Transistors consist of three segments of doped silicon back to back, as it were. The sequence of segments is important; the allowed orders are *n*-type-*p*-type-*n*-type and *p*-type-*n*-type-*p*-type. There are two general classes of transistors, but both can have either the *n*-*p*-*n* or *p*-*n*-*p* sequence of doped silicon segments. The historic first transistor built at Bell Laboratories in 1948 is called a bipolar transistor because electrical current flowing through the device from one end to the other passes through both *n*- and *p*-type silicon and both electrons and holes contribute to the current. Bipolar transistors are also called current controlled because a small electrical current entering the device through the center segment controls whether the device as a whole conducts electricity. A voltage applied only to the two end segments will not cause the transistor to conduct electricity. Viewing the transistor as a switch, one says that the current into the center segment turns the switch on or off.

The second class of transistor is the insulated gate field effect transistor. In this type of device, a thin insulating layer (usually silicon dioxide) is placed between the central segment and its electrode. A voltage applied to the electrode creates an electric field which converts the region of the central segment just under the electrode from one conductivity type to the other (*n*- to *p*-type or vice versa). Thus, field effect transistors differ from bipolar devices in two ways: they are actuated by a voltage applied to the central segment rather than by a current, and all the current is carried by one type of carrier in three segments of the same conductivity type.

With the invention of the integrated circuit in the late 1950's, it became clear that the field effect transistor offered distinct advantages because fewer processing steps were needed to make this type of device and because it took up less space in the silicon. The type of field effect transistor called metal-oxide-semiconductor (MOS) has become the dominant form of commercial integrated circuit. The biggest advantage of the bipolar device is switching speed. Thus, for those applications requiring this capability, such as high-speed logic circuits in computers, bipolar is widely used. Moreover, new forms of bipolar circuits that are more amenable to miniaturization than the older types are being investigated and may well turn out to be as important as MOS microcircuits in the next generation of microelectronics, the VLSI era.—A.L.R.



The first integrated circuit was made in 1958 by Jack Kilby of Texas Instruments.

by scaling all the characteristics by the same factor, the transistor will function the same way independently of its dimensions, or, almost independently. At small enough sizes the dimensionless formulas break down, and this breakdown affects heat dissipation.

At larger dimensions, reducing the size of a transistor reduces the heat produced because less charge has to be moved around and less energy is needed to move it. Offsetting this behavior are more transistors per unit area generating heat. The net effect, according to the scaling principle, is that the total amount of heat produced for a microcircuit chip of a given area remains the same. This felicitous outcome would imply no limitations on miniaturization but for the scaling violation at small dimension.

At some point, as transistors are increasingly miniaturized, the energy needed to drive the switching action becomes so small that it is comparable to the thermal energy of electrons in the device. When this happens, a well-defined switching behavior ceases to be obtainable. Thus, the driving energy (voltage) must be maintained at some fixed level above the thermal energy independently of further miniaturization. As a consequence, the heat generated over the entire microcircuit no longer stays constant but rises with decreasing transistor size. Miniaturization is thereby limited by how efficiently heat can be removed from the integrated circuit chip. One alternative remedy is to operate at cryogenic temperatures where the thermal energy is much lower and the operating voltage can be correspondingly smaller. Less desirable would be to reduce the switching speed. Fewer switching operations per unit time means less heat generated. But this alternative would mean a lower performance.

McGroddy points out that the option of operating at a lower temperature is actually carrying the scaling principle to its logical conclusion. Failure to scale the operating temperature along with other device characteristics leads to transistor failure at small dimensions. MOS transistor switches exhibit a well-defined switching action. A voltage applied to the gate electrode turns the switch on only if it is above a certain threshold magnitude. Plotted in dimensionless form, the threshold behavior should be identical for devices of any length. [The length usually referred to is not the entire device length but the distance between the two electrodes through which electrical current flows when the switch is open, the so-called channel length (see figure on page 1246).] But, says McGroddy, a device with a channel length of 1 micrometer will not exhibit as clear-cut a threshold under the same conditions that a transistor with a channel length of 5 micrometers will. One consequence of the absence of a threshold, which becomes critical at a channel length of 0.8 micrometer or less, is that the charge in the memory cell of a RAM leaks out so rapidly that the device is useless. The reason for the loss of the threshold is that, when the driving energy of the switch becomes comparable to the thermal energy of the electrons in the transistor, the electrons no longer have a clear sense of the force being applied to them.

John Moll of Hewlett-Packard Laboratories describes another consequence of short channel lengths in MOS transistors. Because of errors in the fabrication process, the dimensions of transistors are not all exactly the same. Instead they show a statistical distribution about the nominal values. Device designers must take these variations into account by allowing the output signal from each transistor to have the same sort of variation as the dimensions without disrupting the performance of the microcircuit. As channel lengths shrink from the 10 micrometers characteristic of earlier MOS transistors toward the 1 micrometer expected in future VLSI devices, the margins for error must shrink by a comparable amount. All well and good. The problem is that the device behavior becomes more sensitive to variations that are proportionately the same when the device is smaller. In an example that is exaggerated for clarity, Moll says that two MOS transistors, one with a channel length of 5 micrometers and another with a length of 10 micrometers, will both turn on at the same applied voltage, although they may carry different currents. But two devices, one of 0.5 micrometer channel length and the other of 1 micrometer, will turn on at quite different voltages.

The deviations from the "ideal" electrical behavior characteristic of transistors with large dimensions brought about by shrinking the devices are sometimes called short-channel effects. One of the short-channel effects of most concern at present are hot electrons, which are created when the electric field in a semiconductor becomes very large. A potential of only 1 volt applied to an MOS transistor with a channel length of 1 micrometer would create an enormous field of 1 million volts per meter. In general voltages in integrated circuits have not been decreasing as fast as the shrinking of device dimensions, so each new generation of microcircuit must handle larger and larger electric fields. Hot electrons pick up so much energy from the electric field that it cannot be dissipated in the usual ways (by collisions with vibrating atoms in the silicon crystal, for example). Therefore, the particles have a significantly higher energy than the thermal energy. Looked at another way, the electrons seem to be hotter than the silicon crystal in which they reside.

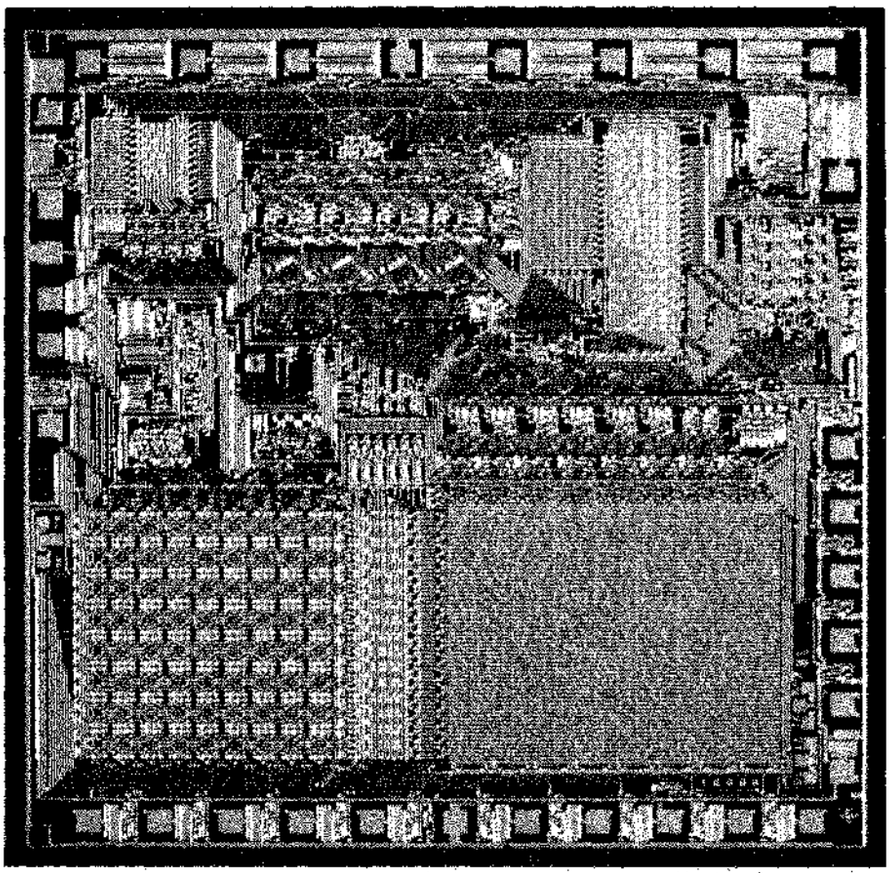
The most important effect of hot electrons as MOS transistor sizes approach VLSI dimensions is that they alter the threshold voltage for switching. Hot electrons have enough energy to jump from the silicon in the region under the gate electrode into the thin insulating oxide layer that separates the electrode from the silicon. The electrons tend to collect in the oxide, where their collective electrical charge distorts the applied voltage seen by the silicon. Although it is possible to design transistor structures that lessen the effect of hot electrons, the main remedy at present is to keep the

electric fields low by reducing the voltage as much as possible. Eventually it will not be possible to reduce voltages because of the limit imposed by the thermal energy.

Some researchers have investigated the possibility of circumventing some short-channel effects by discarding the scaling principle. By reducing channel lengths, device widths, doping concentrations, and operating voltages in different proportions, it may be possible to make transistors that do not exhibit short-channel effects with smaller dimensions than predicted by scaling. One such effort was recently reported by Simon Sze and his co-workers from Bell Laboratories. The investigators empirically determined a formula relating the smallest channel length transistor that would not show short-channel effects to the other device characteristics. Computer extrapolation of the empirical curve suggested the possibility of avoiding short-channel behavior for devices with channel lengths below 1 micrometer.

The ultimate limit on the size of transistors was explored several years ago by Carver Mead and Bruce Hoeneisen of the California Institue of Technology. The researchers concluded that for MOS transistors the minimum channel length was about 0.2 micrometer. A similar limit was placed on the minimum length of the comparable feature in bipolar transistors. The argument of Mead and Hoeneisen depends on the fact that the interface or junction between p- and n-type semiconductors is not perfectly abrupt. The width of the transition region depends on the sharpness of the doping profile (that is, how sharp the transition from p-type to n-type impurities is), on the concentrations of the impurities, and on the applied voltages. As the transistor size decreases, the sum of the widths of the two transition regions between the three segments of the transistor becomes wider than the width of the central segment of the transistor, and the channel disappears. By increasing the concentration of doping impurities, the transition regions can be narrowed, but this procedure simultaneously leads to high electric fields that break down the oxide insulating layer under the gate electrode.

By now there would seem to be more than enough limitations on how much miniaturization can be achieved, but many researchers are looking even farther into the future. It may be that ultrasmall devices may function so differently from today's transistors that some of the limitations discussed are no longer relevant. Investigators believe microfabrication techniques are on the horizon that would allow devices with characteristic dimensions of 100 angstroms or less. In part because of the gigantic electric fields in devices that small, opportunities for completely new types of devices exist. David Ferry of Colorado State University and John Barker of Warwick University in Coventry, the United Kingdom, explained at a NATO advanced study institute in Urbino, Italy, last summer how this could happen.†



Less than two decades later, engineers put a complete microcomputer on a chip. [*Source: Texas Instruments, Inc.*]

Analytical treatment of the transport of electrons in semiconductors assumes, among other things, that the actual time spent in a collision event is negligible as compared to other characteristic times, such as the time between collisions or the time to pass from one electrode to the other. This is a reasonable assumption when the electric field in the semiconductor is not too large. As device sizes decrease and electric fields therefore increase, the velocity of the electron becomes very large and the time between collisions very short. The first effect of miniaturization to very small dimensions, therefore, is that the usual methods of calculating the properties of devices fails. New methods are only now being developed.

One outcome made possible by high fields when the characteristic dimensions of a transistor reach 250 angstroms or less is an altogether new type of transistor action because electrons can pass through the device without ever colliding. Barker cautioned, however, that a rigorously quantum mechanical calculation is necessary to explain the behavior of this "ballistic" transistor, although the concept "is enticingly reminiscent of vacuum tube electronics." One reason for the caution is that, at such very small dimensions, the properties of a device cannot be assumed to be independent of its environment; that is, other nearby devices, insulating structures, chip surfaces, and so forth. According to current practice, a considerable part of the device designer's effort goes to ensuring that each transistor is electrically isolated from its neighbors. But, when such interactions become important, there is the possibility of cooperative phenomena arising among the no-longer isolated devices of the type that occurs in thermodynamic phase transitions. The cooperative behavior is what may make new kinds of electronic devices possible.—ARTHUR L. ROBINSON

†D. K. Ferry, J. R. Barker, C. Jacoboni, Eds., *Physics of Nonlinear Transport in Semiconductors* (Plenum Press, New York and London, 1980).