On Some Modern Uses of the Electron in Logic and Memory

How silicon MOSFET technology came to dominate the ways in which electrons are used in logic and memory devices; will this dominance continue?

Alan Fowler

The electron plays a role in many facets of modern life—lighting, heating, welding and so on. Here, I concentrate on the use of electrons primarily in logic and memory. I generally skip the early history—that of electronic tubes—and start instead with the modern integrated circuit and then go on to explore possible alternatives to the now-dominant silicon MOSFET technology. (MOSFET stands for metal oxide semiconductor field-effect transistor.) This approach is enforced not only by space considerations, but also by the fact that the passing years have erased triodes, power pentodes, klystrons, traveling-wave tubes and the like from my memory. Not that they are no longer used; clones of the 6L6, EL34 and KT88 are still used in some of the best audio amplifiers. Microwaves and power applications still rely heavily on tubes. Cathode-ray tubes are still the standard display, although their days may be numbered—at least at the large and small ends. Electron microscopes and lithography tools are, in a sense, large demountable vacuum tubes. (They are the subject of Murray Gibson's article on page 56 in this special issue.) However, over the past 40 years, the vacuum tube has been displaced by solid-state devices as the basis of most computer, low-power amplifier, communications and control circuits. (See figure 1, which shows a computer processing unit.)

It is interesting to consider what elements have been critical in this change. In general, electrons travel ballistically and therefore move faster in a vacuum than in a solid. Some 45 years ago, a tube engineer told me that that was why he was going back to tubes after a brief foray into transistors. In retrospect, I realize he had missed an essential point, which was not easy to see at that time because the first transistors were not much smaller than miniaturized tubes. The point was that solid-state devices would eventually be made so much smaller than tubes that the shorter distances traveled by the electrons would result in operating speeds comparable to those of the fastest tubes. And by virtue of their size, transistors could be integrated into complex circuits at very high densities, although no one in the early 1950s foresaw those possibilities. To this day, attempts continue to use the techniques developed in silicon technology to make miniaturized cold cathode devices for emission into vacuum, especially for displays, but so far without much success. An area that may see successful applications is multiple-beam electron lithography, using electron-beam microcolumns fabricated with silicon technology.

ALAN FOWLER is an emeritus IBM fellow at IBM's Thomas J. Watson Research Center in Yorktown Heights, New York.

The rise of silicon

The transistor was not the first semiconductor device, nor was silicon the first semiconductor material of choice. Solid-state rectifiers based on the use of copper oxide and selenium preceded the transistor by at least 15 years, and the use of pyrite and galena crystals for detectors goes back even further. These various materials were used at times when many scientists still thought that silicon was a metal rather than a semiconductor. Silicon was used as a detector material as early as 1905, and it was used for radar detectors in World War II. (See Frederick Seitz's article, "Research on Silicon and Germanium in World War II," PHYSICS TODAY, January 1995, page 22.) However, germanium, rather than silicon, was used for the first transistors in 1947.

The bipolar transistor was not the first active device invented. Academic attempts to observe a field effect in thin metal films date back to about 1900, and patents were filed for field-effect devices at least as early as 1928. Far from practical in performance, the first transistors were made with wire-bonded point contacts, but they were essential to the acceptance and the evolution of the germanium-junction bipolar transistor. Germanium was not displaced by silicon as the main transistor material until 1960–62.

As time went on, there were many challenges to silicon as the basis for electronic technology. Except in special cases and for so-called niche applications, the challenges have been turned back until now. Why has silicon predominated for so long?

Despite silicon's having some advantages over its rivals, the material's outright superiority was not obvious initially. Silicon is more useful than germanium at higher temperatures because its energy gap is about one and a half times that of germanium. (In fact, it was for that reason that the armed services supported work in silicon in the 1950s.) Silicon is one of the commonest elements—unlike germanium and gallium, which are not uncommon, however. And silicon can be grown to high perfection in large crystals, up to ten inches in diameter, which is more difficult to do with gallium arsenide (GaAs). Yet, since the mobilities of holes and electrons in silicon are about half those of germanium, silicon devices are inherently slower for the same dimensions. Moreover, GaAs and

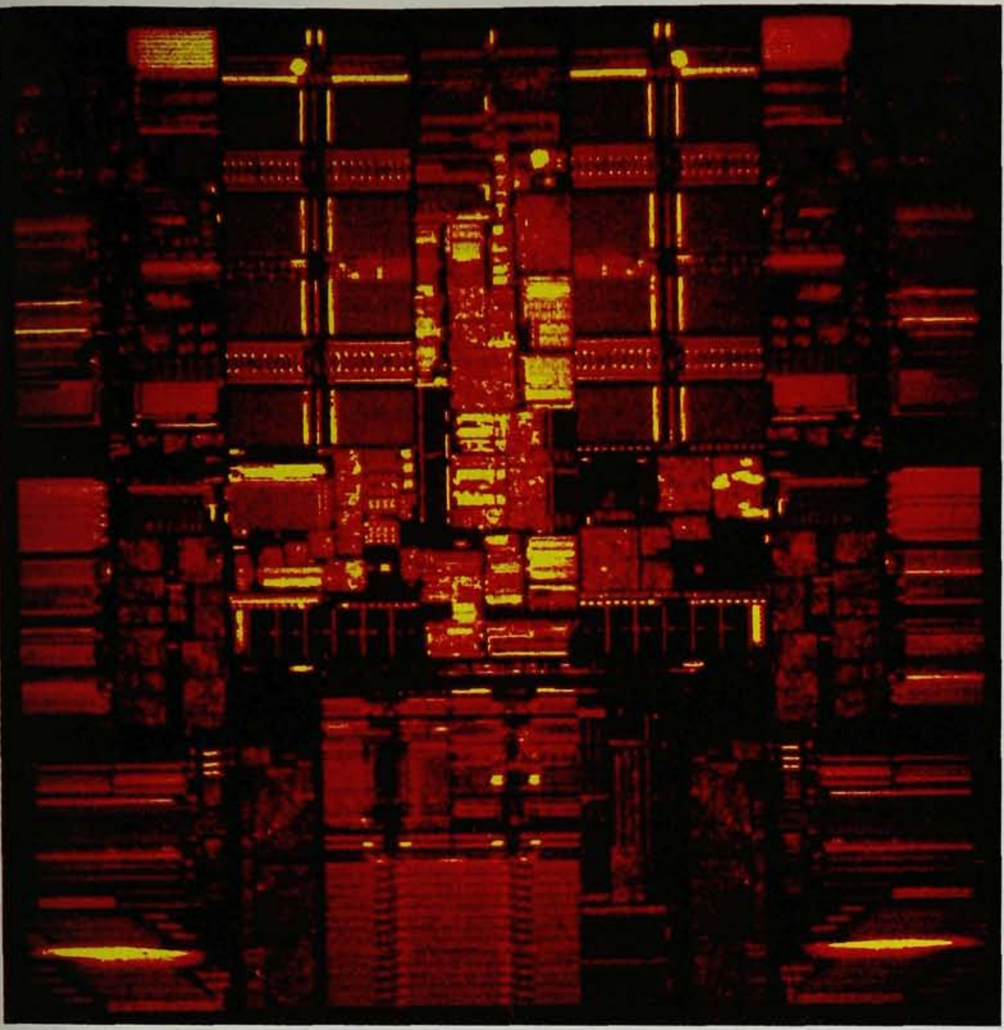

FIGURE 1. A MODERN MICROPROCESSOR—in this case, the chip used in IBM S/390 mainframes—based on the complementary metal oxide semiconductor field-effect transistor (CMOSFET) technology. Measuring 18 mm on a side, the chip contains over eight million transistors and has features as small as 0.3 μm. This image, which was taken in the near infrared by a cooled charge-coupled device camera, shows the weak light emitted by the chip when individual CMOSFET gates switch on and off. Charge carriers passing through the channels of modern CMOSFET gates experience electric fields of over 10^7 V cm^{-1} and can acquire appreciable kinetic energy, some of which is converted to light energy. The particular circuit activity shown in this image is due to the clocks in the chip, and to a signal propagating through a chain of 64 000 latches in the circuit. (Figure made by Jeffrey Kash and James Tsang at IBM's Thomas J. Watson Research Center; originally published in *Science*, vol. 276, p. 1987.)

silicon carbide have even greater gaps than silicon, and the mobility of electrons in GaAs is even higher than that of electrons in germanium.

What makes silicon unique is not its own properties, but those of its native oxide—silicon oxide (SiO_2). This oxide is extremely stable chemically with respect to silicon, and also is stable with respect to other materials used in electronics. It is an exceptional insulator, with both a high bandgap and large barriers against injection of electrons from contacts, and is relatively free from traps, which could trap electrons injected into the oxide and thereby degrade the device or change its operating characteristics. SiO_2 is relatively impermeable to most impurities that could diffuse through it and contaminate the interface. Although the interface with silicon is not perfect—unlike a perfectly matched heterojunction—the "dangling" bonds resulting from the mismatch can be saturated by diffusing hydrogen up to the interface to produce a surface relatively free of interface states. Furthermore, SiO_2 is readily etched, enabling it to act as a mask for the diffusion or ion implantation of dopants into the silicon. Technologically, SiO_2 confers immense advantages on silicon. Some of them can be extended to certain other—but not all—semiconductors by using deposited SiO_2.

By taking advantage of the insulating, passivating and masking properties of SiO_2, planar technology was invented in the 1950s and developed to produce today's very-large-scale integrated (VLSI) circuits. (Planar technology is so-named because the devices are made by diffusing or ion-implanting impurities into the plane of the silicon wafer, with insulators and wires being deposited in successive layers.) The growth of the circuit complexity and speed of chips has been phenomenal. Since 1970, when it was about 10 μm, the size of a feature on a chip has been halved about every five years. At present, devices with 250 nm features and gate lengths of about 200 nm are in production. Gate length is roughly inversely proportional to device speed, and when feature sizes are halved, circuit density increases by a factor of three or four. Devices with features as small as 30 nm have been made using electron-beam lithography. However, optical projection lithography is currently the only lithographic method used to manufacture large integrated circuits. At the same time, quality control of the technology has improved to the degree that perfect integrated circuits can be produced with high yield over much larger areas. As a result, complete memories now approaching 1 gigabyte can be fabricated on one chip, although the usual DRAM (dynamic random-access memory) currently has 16 or 64 megabytes of memory.

From bipolar to field effect

For semiconductors' first 20 years, the predominant device was the bipolar transistor (see figure 2a.) It continued to dominate high-speed switching, logic and memory applications until about 1990, after which it was gradually displaced by complementary MOSFETs, called CMOSFETs, which use far less power. Complementary switches use both an n- and a p-MOSFET in series in such a way that when one is on the other is off, drawing power only during switching. At first, field-effect transistors were used primarily for low-speed circuits and in some areas of memory, especially the ubiquitous DRAM. The complexity and size of an ultra-high-scale integrated circuit may be seen in figure 1. A logic circuit was chosen to illustrate that point, rather than an even more populated memory chip, because it is far more complex. This particular logic circuit, or chip, has about 8 million devices, but logic circuits with 40 million devices have been fabricated. (A more simply wired 256-megabyte DRAM chip would have over 400 million devices.) Specifically, the chip shown in the figure contains two entire mainframe central processing units (two instruction units, two fixed-point units, two floating-point units and so on), a register unit to compare the results from the two processors and a 64-kilobyte cache memory. Together, they constitute the central processing unit of an IBM S/390 mainframe computer—all on a chip that measures about 18 mm in size. Designed to run at 350 MHz, the chip has five levels of connecting metal wires. The picture itself was made by sampling light emitted when the MOSFETs were switching. This method allows us to study the operation and design of the circuit in detail and to find flaws. In special cases, switching shifts of 20–30 picoseconds can be observed.

Next-generation MOSFETs

How much further can this silicon-based MOSFET technology be extended? Like most phenomena with exponential growth rates, it cannot be expected to expand forever. (Whether the demand for increasing function goes on forever is another question, probably more connected to the dynamics and sociology of consumerism than to need.)

Potentially, there are several physical limitations. Among them is the lithography used to define the patterns on the chip. Optical projection lithography remains the universally used means of transmitting such complex patterns at the rates needed for chip production. At present, it is the major practical technique for transferring the master image in parallel rather than in series. Twenty-five years ago, it was thought unlikely that optical projection lithography could be extended down to 1000 nm. Now it is being used down to 248 nm and will be extensible to 193 nm in 4–5 years. (These are the wavelengths: The features are somewhat smaller.) This reduction has been accomplished by moving into the deep ultraviolet with conventional optics. Extension below 100 nm may be possible by reaching further into the UV, where mirrors—as opposed to lenses—must be used. Ultraviolet projection lithography overlaps with x-ray contact lithography, which is the only tool that has demonstrated the capability of manufacturing large circuits at 100 nm, and may be extensible to about 50 nm in 10 or more years. Making masks has been a problem for x-ray lithography because x rays penetrate exposed film—a problem that might be mitigated for optical systems below 100 nm.

Electron-beam lithography has also been used for many years. So far, however, it has been too slow for manufacturing. Its main uses have been in making masks for photolithography, in customizing circuits and for building advanced test chips and devices ahead of the more slowly advancing optical techniques. The practical limit of electron-beam lithography is probably around 30 nm, although special techniques can push it to about 5 nm. There have been many attempts to produce higher throughput electron-beam lithographic tools. Most of them have been based on some form of projection system using a mask. None has so far been fast enough or good enough, but new approaches offer some hope. Another approach is to develop electron-beam microcolumns, which could be integrated and operated in parallel in large numbers.

All these approaches, including the use of x rays, require a large investment of time and money, and the results are unpredictable. Of the approaches mentioned, one or more may eventually be useful to 20–50 nm. If operating an electron-beam system is difficult, those difficulties vanish compared with using an atomic force microscope to move one atom at a time or to do lithography.

Another way to understand the scale of some of these problems is to consider the cost. A modern silicon fabrication facility now costs more than $2 billion. A step-and-repeat camera, which aligns the image of a mask on the wafer in a precise position, costs about $10 million. High tech has a high price. Initial investment in new factories has grown exponentially with time. So far, the cost per unit area of silicon chips has remained fairly constant, whereas content per unit area has increased exponentially, so that the large investment has been repaid. But nothing guarantees that this will continue to be true.

Potentially, there are also device limitations. The physics of electron devices below 50 nm is only partially understood. In this regime, the size of the device becomes comparable to the electron wavelength, and such elementary questions as the size of the electron wavepacket become important. The scaling rules used by engineers for many years to relate insulator thickness, gate lengths and silicon doping have run out of steam. Nonetheless, devices as small as 30 nm have been fabricated, although they lack some qualities of good switches, such as failure to turn off. Recent proposals, however, may solve this problem, at least at 30 nm. The scaling of MOSFETs has reduced oxide thicknesses over the last five years from 10 nm to 5 nm. At 2 nm, tunneling through the oxide becomes a problem, especially for DRAMs.

MOSFETs and physics

Before turning to alternative technologies and devices, it is useful to note the contributions of the silicon MOSFET to understanding physics. Since the electrons in the surface are squeezed within 0.5–10 nm of the surface, they are quantized perpendicular to the surface and so can be considered a quasi two-dimensional electron gas (or insulator, for that matter). Furthermore, since the Fermi energy and electron density can be varied, we have a model system for studying many of the properties of two-dimensional systems over a large range of parameters. For instance, the electron density can be varied by two orders of magnitude. The first observations of two-dimensional behavior were made in silicon MOSFETs, as were those of two-dimensional localization, one-dimensional localization and, of course, the quantum Hall effect. Recently, the first reasonably good evidence of a two-dimensional metal–insulator transition in the absence of a magnetic field has been found in silicon MOSFETs.

Beyond silicon

Whither the technology after silicon? Any replacement must be as good as silicon in most properties, and distinctly better in at least one if it is to become the new mainstream technology. The use of other semiconductors continues to grow.

Compounds of elements from groups III and V of the periodic table have many uses. They dominate the applications of the injection laser and light-emitting diode because, as direct bandgap materials, they are far more efficient light emitters than silicon. The semiconductor injection laser is essential to fiber-optic communication and the use of laser printers, and the light-emitting diode is essential to compact disc technology. GaAs, either with or without GaAlAs, is used to make the fastest solid-state amplifiers, and finds wide use in cellular phones. But because processing these materials is harder to replicate than silicon, it is difficult to make large circuits optimally.

At times, II–VI compounds have found various applications, and recently have been exploited to make light-emitting diodes in the green and blue, although aluminum gallium nitride–indium nitride seems more likely to be practicable.

A more likely successor to silicon than GaAs may be Ge–Si, which has many advantages, one of the most important being that it can be fabricated in a silicon facility using the same tools. Its electron mobility is higher than that of silicon, so that the devices are inherently faster, but not as fast as GaAs–GaAlAs heterojunction circuits, which have the advantage of a high resistance (or in the usual, but less rigorous terminology, a semi-insulating) substrate. The first Ge–Si devices to have found a use are hetero-bipolar transistors, which fill a niche where silicon isn't fast enough and where the circuits are too large to be made in GaAs.

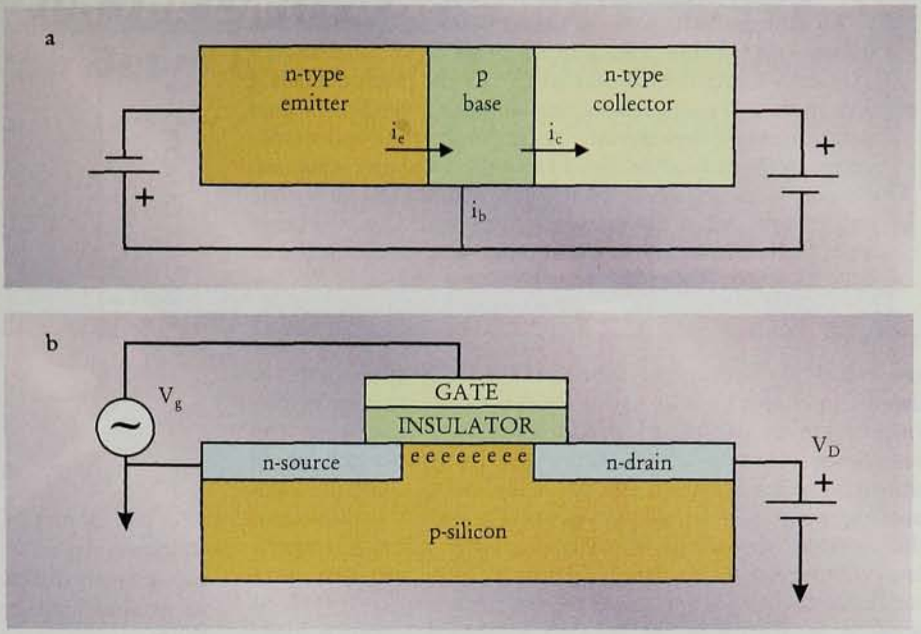

FIGURE 2. BIPOLAR AND FIELD-EFFECT TRANSISTORS. The bipolar transistor (a) functions by emitting electrons across a low-impedance, forward-biased emitter junction into the base region. Most of the electrons are collected by the high-impedance reverse-biased collector junction. The field-effect transistor (b) operates by biasing the gate to induce the presence of electrons in the "channel" in the silicon beneath it and to provide a conducting path between the *n*-doped source and drain regions.

GaAs–GaAlAs heterojunctions have produced a lot of interesting new physics, mainly as a result of very high mobilities of about 10^7 cm^2 V^{-1} s^{-1} at low temperatures. The fractional quantum Hall effect was first observed in such samples. The very long mean free paths of up to the order of 10 μm have been exploited by using patterns defined by Schottky gates on the surface. This technique has allowed surface features to be defined on scales small compared to a wavelength so that quasi one-dimensional and zero-dimensional structures can be made. Probably the most interesting result so far has been the observation of conductance quantization through a narrow aperture or point contact. (See figure 3 and Henk van Houten and Carlo Beenakker's article, "Quantum Point Contacts," PHYSICS TODAY, July 1996, page 22.) So-called quantum dots have also received a lot of attention, as has the transport of electrons through electron "wave guides" and past obstacles.

One way to improve the speed and reliability of silicon MOSFETs is to cool them. So far, the cost of cooling, both in power and money, has not been justified by improved performance. Many of the new devices invented over the past 40 years require low temperatures to operate at all. Of these, the Josephson junction probably has been the focus of the greatest effort to make high-speed logic circuits, despite the fact that it lacks the major attributes of a good logic element—namely, significant gain (current amplification) and good isolation of input and output. Furthermore, it has to be reset at every cycle (although some circuits mentioned below do not require resetting).

Beyond MOSFETs

Without going through all the many tries at replacement technologies over the last 40 years (Esaki diode circuits, magnetic core logic, logic based on Gunn devices, combinations of light sources and photoconductors, parametrically excited nonlinear circuits, to name a few), any good logic device should have certain simple attributes. Gain is necessary because, in general, circuits "fan out," and each device must drive several others. The signal level of a device should reset to a fixed state (as CMOSFET circuits reset to zero or close to the supply voltage). In practice, switching speeds are not limited by the raw switching speed of the device, but by the current that charges the capacitance of lines. For fast circuits, therefore, the "on" resistance must be low—less than 10 kΩ, which is several times higher than the highest MOSFET resistances. Scalability and extensibility have been major assets for silicon technologies, for not all technologies are scalable or extensible. Requirements for memory are usually less rigorous. However, memory and logic technologies should be compatible. One of the failings of the Josephson technology was that large memories proved impossible to make.

Some of the most publicized new technologies are single-electron transistors, quantum cellular automata and all-optical logic. A more nearly complete list would include magnetic bipolar transistors, quantum parallelism, various mesoscopic devices depending on interference or quantized conductance, atomic relays, molecular switches, rapid single flux quantum circuits and various two-terminal resonant tunneling devices. That each technology mentioned is flawed for mass use in logic does not mean that the technologies have no other uses. Although Gunn logic was unsuccessful, Gunn devices were not without their uses. As examples of failures for use in logic, we look at a few of the possibilities mentioned above.

The single-electron transistor (SET) is based on some very elementary physics. If an electron tunnels through a junction with a capacitance *C*, the energy of the system must be changed by the capacitive energy. When the capacitance is small, it can control the tunneling, and in effect produce a voltage barrier to tunneling of *C*/2*e*, where *e* is the electronic charge. The SET has a small island between tunneling contacts, and has a gate electrode that can control the potential of the island, with the result that current through the device can be controlled. Actually, the conductance is a periodic function of the gate voltage, with period *C*/*e*. "SET" is somewhat a misnomer because it implies that only one electron is involved in its operation, whereas it is the granularity of the charge that characterizes the device.

SETs have been fabricated from metals alone and semiconductors. They have many advantages. The all-metal structures are densely integrable, use low power,

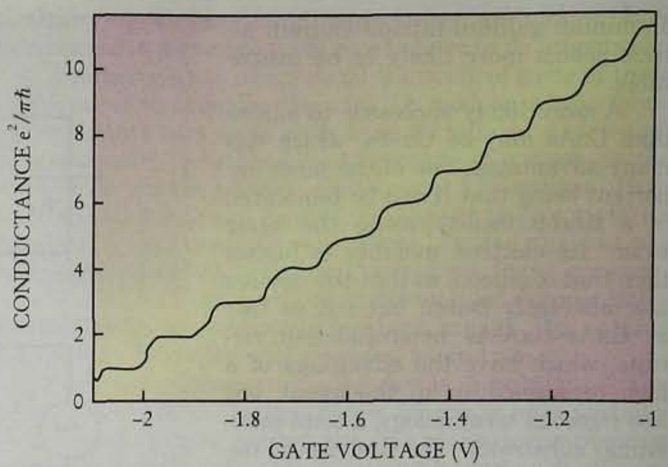

FIGURE 3. VOLTAGE-CURRENT CHARACTERISTICS of a voltage-controllable point contact made on the surface of a GaAs–GaAlAs heterojunction. As the point contact is opened, more and more conducting channels (one-dimensional quantum levels) are opened. Accordingly, the conductance jumps by $2e^2/\pi\hbar$, as expected from the Landauer equation. The effects were first observed in 1988 at about the same time by two groups: one at the University of Cambridge, the other from Delft University of Technology and Philips Research Laboratories. (From Bart van Wees *et al.* Physical Review Letters, vol. 60, p. 849, 1988.)

are relatively simple to fabricate and are integrable in three dimensions—and could therefore be used in multivalue logic (a use of questionable value). However, the resistance must be of the order of 100 kΩ to get sharp characteristics, which leads to problems in charging connecting wires at competitive switching rates. In addition, the devices are sensitive to background charge, require low temperatures at their present size and are very sensitive to fabrication parameters. Most of the work at present concentrates on making lower capacitance junctions to raise the operating temperature. Success will not alleviate the other problems and may in fact exacerbate some. Although the resistance problem cannot go away, it is possible that these circuits will find a use if some of the other problems can be overcome.

Notwithstanding the fact that "mesoscopic" means different things to different people, most mesoscopic devices share some of the disadvantages of the SET, but with few of the advantages. Their resistance is high, although perhaps an order of magnitude less than that of the SET. Reproducibility of the devices is even more of a problem, and very low temperatures are required (less than 1 kelvin). They are not compact and offer no advantage in integrability. Furthermore, it is not clear if any of these devices have voltage gain when significant voltages are applied. So far, no device has been proposed that looks like it would be useful for anything. That does not mean that none will be: The field is young. An intriguing, but as-yet-unrealized proposal, is the possibility of quantum cellular automata. Mesoscopic quantum dots or molecules can be made to be polarizable. Since the polarization of one molecule can affect that of nearby ones, a polarization signal can propagate along a chain (perhaps in both directions.) More complex logical functions could possibly be realized with more complex structures. At this point, it is not clear that there is any density or speed advantage to be gained from quantum dots, and low temperatures are likely to be needed. Quantum dots could be made from silicon (an advantage), but there is no apparent gain. However, despite these and other objections, these structures deserve attention.

Devices that have been rejected do sometimes make a comeback when times are technologically ripe, as proven by the case of thin film transistors. Thin film transistors (related to MOSFETs but using thin film technology) received much attention in the 1960s and later in the 1970s, especially for use over large areas such as in displays. Originally, various II–VI materials were used. With the development of a practicable display medium, liquid crystals, and field-effect devices made with amorphous silicon, thin film transistors became important in making small flat-panel displays. It could be argued that one or more of the devices described above could also have its day.

Beyond binary logic

If a limit is reached in the size of circuits because of lithographic reasons, it is tempting to try to increase effective density by turning to multilevel logic. Compared to binary logic, which has only two states, multilevel logic has proven difficult to implement. The reason is that, as in analog logic—which in principle is far more powerful than binary logic—small errors, resulting from noise or fluctuations in device characteristics, tend to propagate disastrously after many stages.

A hell of a ride

If the discussion above seems unduly pessimistic about the chance of replacing and improving on silicon—coming from an old codger, who participated in and benefited from the silicon revolution—it is not the whole story. It is clear that silicon has its limits. When these limits will be reached is not so clear. The demise or maturation of silicon technology has been predicted many times in the past. But there is certainly a limit. If there is no attempt to find alternatives, they will never be found. Most of the large electronics companies have grown so sophisticated in finding reasons not to work on new technologies, and so bottom-line conscious, that they are unwilling to risk significant money in looking at or for alternatives. Thus, stagnation is probable in the computer and information hardware in a finite time. By then, we may drown in information, if we have not already, and scientific and engineering talent will be turned to other problems. Whatever happens—participating in the last 40 years of semiconductor development has been a hell of a ride and the next 10–20 may be too.

Further reading

1. A survey of where things were about five years ago can be found in Robert W. Keyes, "The Future of Solid-State Electronics," PHYSICS TODAY, August 1992, p. 42.
2. To go back even further, read Rolf Landauer, "The Future Evolution of the Computer," PHYSICS TODAY, July 1970, p. 22.
3. A recent and more optimistic view of the status of superconducting circuits may be found in Konstantin Likharev, "Superconductors Speed Up Computation," *Physics World*, May 1997, p. 39.
4. A discussion of the "law" that has described the growth of integrated circuits for the last 30 years is given in Robert Schiller, "Moore's Law: Past, Present and Future," *IEEE Spectrum*, June 1997.
5. A discussion of the future of SETs is given in Kenji Taniguchi, Masaharu Kirihara, *FED Journal* **7**, 53 (1997). (This journal is dedicated to "Future Electron Devices.")
6. An optimistic view of the future of nano-electronics is given by David Goldhaber-Gordon *et al.*, *Proceedings of the IEEE* **85**, 521 (1997).
7. Vital perspective about extravagant claims may be achieved by reading Rolf Landauer, "Need for Critical Assessment," *IEEE Trans. on Electron Devices* **43**, 1637 (1996) and Rolf Landauer, "Advanced Technology and Truth in Advertising," *Physica A* **168**, 75 (1990). The former contains many useful references. ■