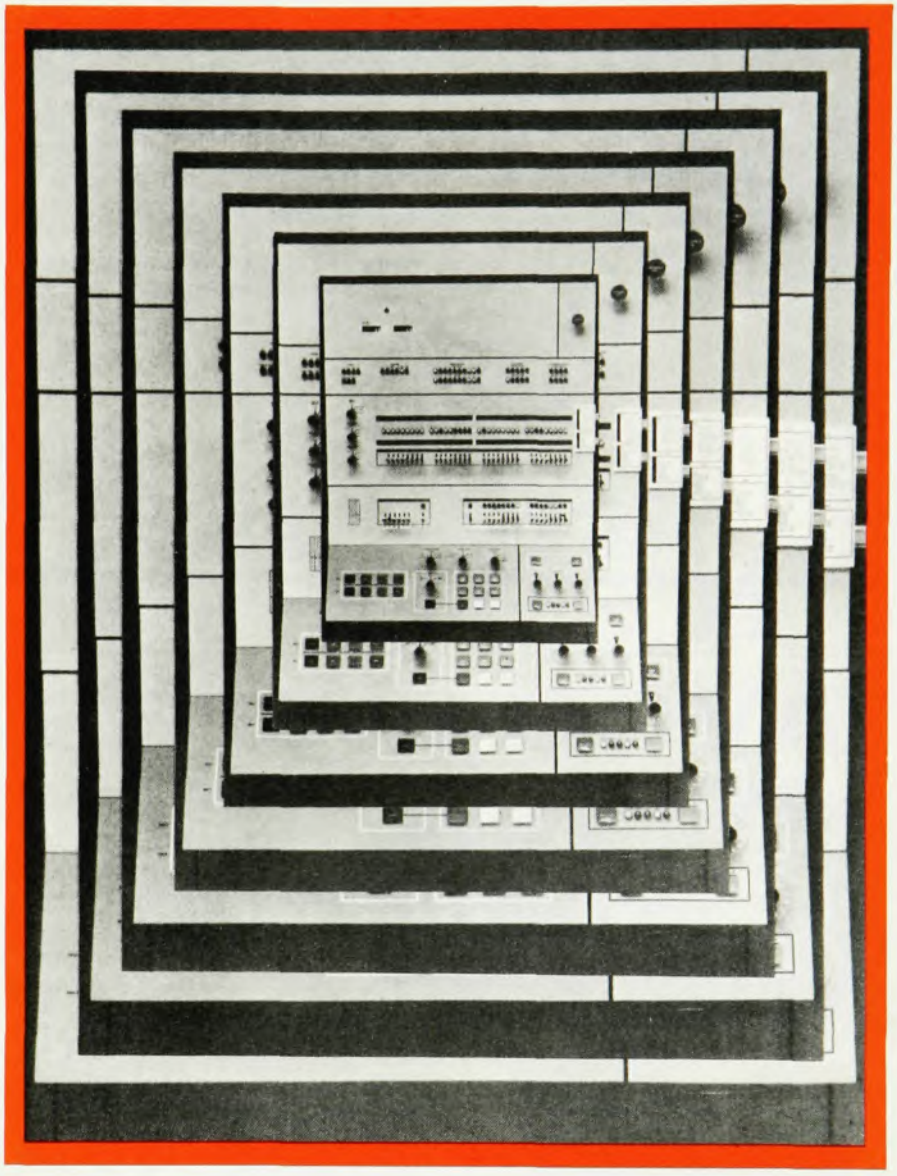
THE FUTURE EVOLUTION OF THE COMPUTER



Internal-logic devices have set the pace in the past and could be the key to further development.

Rolf Landauer

DIGITAL COMPUTERS in physics research

How far can we expect the development of the computer to go? What paths will future evolution take? To answer these questions we should consider both refinements in existing technologies and completely new device principles. The pace of computer evolution so far also leads us to think about the ultimate limitations of the information-handling process, which inevitably depends on the number of real physical degrees of freedom available.

To keep this discussion within reasonable limits I shall deal here only with the devices that perform the internal logic in the computer. To single out these devices for such special attention from the complex system that makes up a complete computer, may seem invidious. It is rather like equating the progress of automotive transport to progress in the design of car engines. After all, many other things have had important effects on transport; they include quality of the roads, driver education, tires, steering systems, atmospheric contamination, and the level of gasoline taxes.

However, gross simplifications are occasionally necessary. For the computer we can claim that the internal logic devices, because of their rapid progress, have been the pace setters. Also their potential for evolution is understood in greater depth than it is for the rest of the system. The rapid progress of logic devices emphasizes the need for clever new system concepts, which can use device improvements to ease more stubborn problems elsewhere.

The computing process

Computing is a process in which streams of information come together and interact with each other nonlinearly. Thus in a typical elementary step in a computer two signals will be fed into a circuit, giving rise to an output that will be some function of the inputs. For example in an "and" device the output is "1" if, and only if, the two inputs are each "1." Even at this level of description it becomes clear that a device with only two leads coming out of it is not well enough

equipped for us to distinguish between "outputs" and "inputs," and it will be hard to use such a device as the main element in a logic stage. Computers have been built out of two-terminal devices, but these schemes have never led to a widespread technology.

Typically we do logic with three-terminal devices, of the kind illustrated schematically in figure 1; fluid supplied through one pipe controls the flow through another pipe. Such fluid-logic elements were taken seriously for general computer purposes as recently as a decade ago, when the attempts to make integrated semiconductors were still in their infancy; by contrast we understood how to mold many of these hydraulic elements, simultaneously, in one block. Figure 1 illustrates that computation does not have to be done by electrical circuits—we can use other physical interactions instead. But if we want these logic events to take place quickly we should choose fast-moving entities, such as phonons, excitons, or spin-waves, rather than bulk matter. Note, however, that we need a nonlinear interaction between the information streams, and that for uncharged excitations this generally takes us to very high energies. By contrast the electron has a very convenient "handle," its Coulomb charge, that makes interactions easy. Nevertheless, handling information through purely radiative degrees of freedom is a subject that continues to attract new proposals. We have laser pulses available today that last only a fraction of a picosecond, whereas the ordinary electrical technology merely takes us down to about 20 picoseconds. Furthermore, we already know a great deal about nonlinear interactions between light pulses. But all of the known suggestions for completely optical logic take us to very high energies, as has been shown in a recent analysis by R. W. Keyes and J. A. Armstrong.[1] Their analysis does not eliminate all further possibilities of low-energy optical logic; rather it looks at some of the existing proposals in the broadest spirit possible and shows that these do require high energies. In addition to the energy requirements, there will probably be a need for precision optical surfaces, both in lasers that may be involved and also at places where energy is being divided (as for example when one logic stage sends its signal on to several other stages). It is nevertheless gratifying to have these very short laser pulses to prove that we can find physical phenomena to take us to time scales very short compared to those currently faced in computers.

I have emphasized the difficulties of replacing the transistor, as an elementary logic engine, by optics. This is by no means to say that optics and lasers have no role in data processing. The laser beam has a very likely role as a pointer, for example, to reach into large memories for reading and writing information.

Before leaving figure 1 let us note one of its basic features. The operating principle of that device can be said to be the conservation of volume surrounding the piston. If more volume is taken up by the control fluid, less is available for the channel whose flow is being controlled. The piston can not be too leaky if the device is to work, and the two fluids have to be recognizably separate.

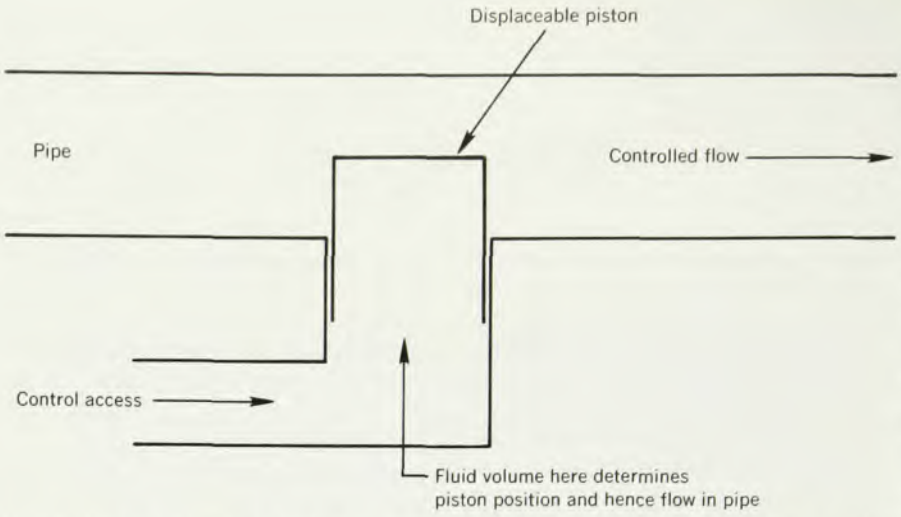
Electron interactions

In our search for new logic devices, we note that our discussion above showed that electron interactions are the most likely basis for these devices. But we are still left with a choice between electrostatic and magnetic interactions. It would be nice if we could easily rule out the magnetic interactions at this point by some simple principle. After all, for two well separated electrons moving at modest velocities, the e^2/r term is large compared to Lorentz forces and spin terms. But we have to remember that in an initially neutral solid structure, such as a computer, charges arise as a result of current flow, and they are generated by the same currents that determine magnetic fields. In electromagnetic waves, and in resonant circuits, electric and magnetic energies are in fact equal. If nature had made electrostatic interactions more easily available, we would have electrostatic motors in our household appliances. The duality between electrostatics and magnetics is shown by the presence of logic devices that use magnetic fields, which remain serious contenders although they have never come to dominate technology. Magnetic fields can control current flow particularly well through their effect on the existence of a superconducting state, and superconducting devices continue to be of interest. Work in Josephson junctions[2] has replaced an earlier interest in controlling bulk superconductivity. Josephson junctions can be switched by means of the current carried by another Josephson junction in a time shorter than a nanosecond, and are thus comparable in speed to transistors. Interactions between magnetic domains also allow logic possibilities, and a particularly interesting set of proposals has been made for such schemes in orthoferrite materials.[3] P. M. Marcus and M. J. Freiser[4] have argued, however, that devices in which electron spins have to relax cannot be expected to switch in less than 3×10^{-10} sec; the fastest exploratory transistors already do better than this. If, in the remainder of this discussion, I slight magnetic interactions it is not as a result of an analysis comparable to that involved in putting optical interactions aside.[1] It is rather because devices based on electrostatic interactions have been far more developed and their potential for evolution is better understood.

Once we specialize to letting current-flow patterns in solid structures interact through Coulomb charges we have arrived at a transistor, admittedly defined in a somewhat generous fashion. It is, however, very useful to define a transistor as a device, analogous to figure 1, in which two charges, Q_m and Q_c, occupy either the same volume or closely adjacent volumes and obey a neutrality condition

$$Q_m + Q_c = \text{constant}$$

One of these, Q_c, is considered to be the *control*, determining how much *moving* charge, Q_m, is available for motion. Just as the piston in figure 1 should not be too leaky, the electrons



HYDRAULIC CONTROL ELEMENT. This three-terminal logic element uses fluids, but it also illustrates the principle of other logic devices. —FIG. 1

should not make transitions between the two classes too quickly. If Q_m and Q_e are contained respectively in the two bands of a semiconductor, we obtain the ordinary "junction" transistor that currently fills our computers and our radios. If Q_m is at the surface of a crystalline semiconductor, as illustrated in figure 2, and the amount of Q_m is controlled by the charge brought into a metal electrode, separated from the semiconductor by a thin insulator, we have an "Insulated Gate Field Effect Transistor" (IGFET), which in the last few years has received continually increasing technological attention. It is a device that requires fewer process steps than the ordinary transistor and appears particularly suited for high device densities in integrated circuits. High device densities, illustrated in figure 3, show integration of ordinary junction transistors not at the laboratory frontier but as already represented in existing commercial computers (IBM System 360, models 85 and 195). Such high densities are relevant for several reasons. First of all, the more devices we can make in one physical piece the lower the processing cost per device. Secondly, to the extent that we can interconnect these devices on the semiconductor chip, the high density reduces the burden imposed on the rest of the supporting and interconnecting structure. Finally, of course, compactness in a computer cuts down the time required for a signal to propagate between different portions of the computer. The IGFET has also exhibited itself as an interesting tool for physics. It gives us a method for confining a two-dimensional electron gas[5] to the surface of a semiconductor, a configuration of particles not otherwise easily obtainable.

Another type of field effect, the Schottky-barrier transistor, eliminates the insulating layer altogether and relies upon biasing the resulting rectifying metal–semiconductor interface in the direction of little current flow. This means that the charge in the metal has the right sign to repel the carriers in the underlying semiconductor. If that semiconductor is thin enough, current flow through it can be cut off by chasing all the carriers out of it. (The Schottky-barrier transistor also eliminates the p–n junctions shown in figure 2.)

Devices of this type have been made[6] that can provide gain up to frequencies of 30 GHz, a far higher frequency than has been attained with conventional junction transistors. The device, shown on the cover of this issue of PHYSICS TODAY, is a gallium-arsenide structure. As well as illustrating the Schottky-barrier transistor, the device also serves to illustrate the potential inherent in semiconductors other than germanium and silicon.

Still other types of field-effect transistors have received considerable attention. These include thin-film transistors and junction field-effect transistors.

Other charge decompositions are possible. In the early 1960's a good deal of attention was paid to a hot-electron transistor, in which Q_e consisted of the normal low-energy electrons (typically in a thin metal), whereas Q_m consisted of much faster electrons, injected at energies high up in the conduction band. Unfortunately the device was never made to work in a convincing and appealing way, and this may have been partly responsible for the very conservative attitude towards the invention of further very novel transistor types.

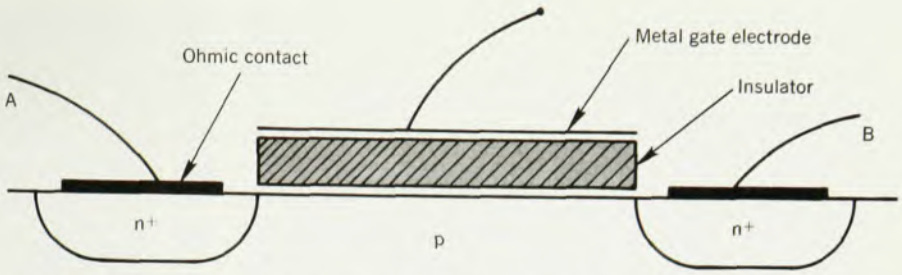
In my opinion the search for other novel transistor types has not been prosecuted with nearly enough zest and energy, perhaps because the basically oriented solid-state physicist has become too divorced from the transistor's technological thrust. In principle any decomposition of carriers into two classes is a hint for a possible transistor structure. Note, however, that as we learn how to make more and more compact structures, in which carriers spend less and less time, the requirements of the interclass transition rates (recombination times), that is, avoidance of the leaky piston, become easier to satisfy.

Naturally, the decomposition of electrons into two classes is not enough; the right means of introducing carriers and taking them out of the structure must also be found. A recent development[7] shows how charges can be passed around a silicon surface, under external control, passing from the area under one control electrode to the next control electrode. The control electrodes here are separated from the silicon by an insulating thermal oxide of silicon, as in the IGFET.

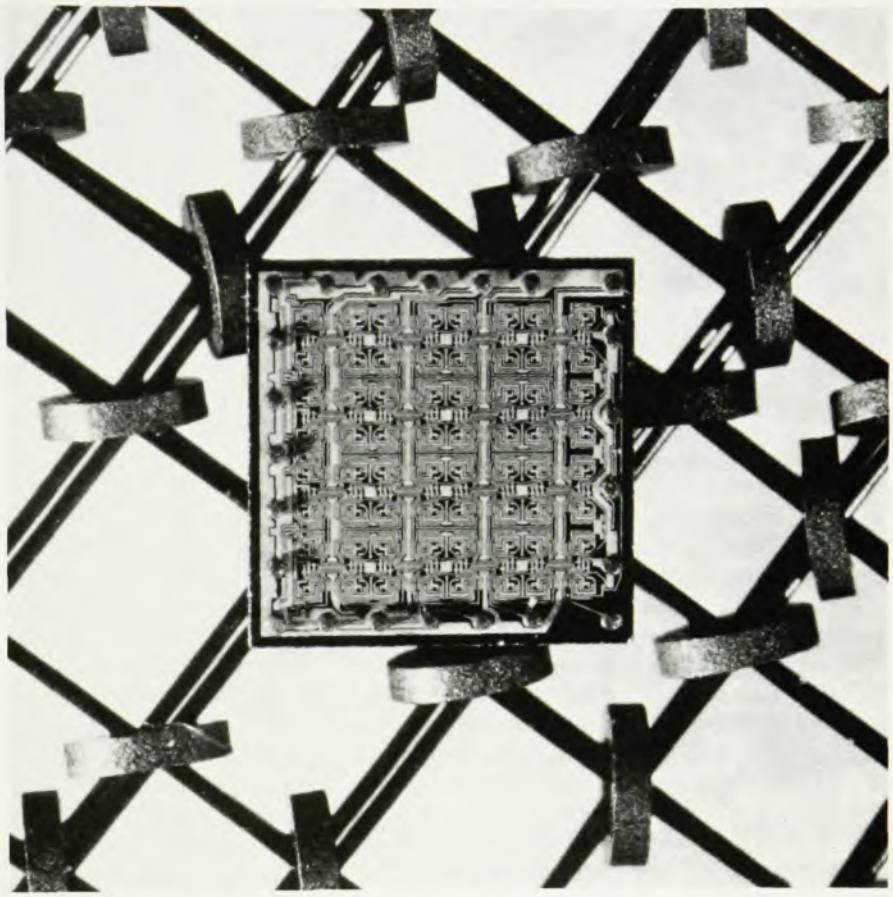
Rolf Landauer took his PhD at Harvard in 1950 and joined IBM two years later. During 1962–66 he was director of physical sciences at the Yorktown Research Laboratory; in 1966 he became assistant director of research, and in 1969 he was appointed an IBM Fellow, a position which has allowed him to return to more personal research activity. His interests have been in the physics of computing devices, transport theory, ferroelectricity, and nonlinear electromagnetic wave propagation.

In this very recent development the charges do not have to leave the silicon surface by metallic connections, so that, in principle, the method of introducing and removing carriers can become extremely simple.

Thus we can either try to invent transistors that utilize the differentiation between spin-up and spin-down electrons, or we could utilize the distinction, in or near a superconductor, between paired superconducting electrons and the normal electrons. A primitive relative of a transistor, which exhibits the high transit velocities obtainable for electrons at the Fermi surface of metals, exists already as part of an experiment unrelated to technology. R. J. von Gutfeld and A. H. Nethercot[8] have exposed one end of a crystal of gallium to short heat pulses by laser irradiation. They then measured the transfer of heat to a fast responding bolometer on the other side of the sample. They have found that they can detect the unscattered ballistic flight of electrons, coming straight through the sample, as the first indications of heat transport. Thus electrons in the velocity classes emanating from the heated spot are favored at the expense of other velocity classes. Figure 4 illustrates the experiment.



INSULATED GATE FIELD EFFECT TRANSISTOR. In this schematic representation, current flow from A to B is small because one of the p-n junctions is reverse biased. When a positive charge is brought into the gate, an n-type layer is established at the semiconductor surface, so permitting current flow. —FIG. 2



INTEGRATED CIRCUIT CHIP measuring 0.284 cm on each side. The chip contains 64 bits of memory and is photographed against a background of magnetic cores containing one bit per core. The chip has 664 components, which represents a density of over 8000 components per square centimeter. —FIG. 3

Transistor size

In the future evolution of the transistor towards higher speeds there are three avenues that we can explore. The first is the invention of completely new transistor types, using novel operating principles, as we have just discussed. The second is to reduce the size of existing structures, and the third is to apply new materials, with perhaps greater carrier mobility, to these devices.

Now let us see what we can achieve by reducing transistor size. Reducing size reduces the transit time for carriers through the structure and reduces the device capacitances. Modern microelectronic elements are made by optical techniques, exposing materials called "photoresists" to light patterns, thus selectively polymerizing (or depolymerizing) the photoresist, which then serves as a pattern in a subsequent etching step. The push toward smaller elements is symbolized by current explorations,[9] illustrated in figure 5, to use electron beams for photoresist exposure, and thus to start at and subsequently go beyond the limitations of optics. Some useful devices[10] (but not transistors) have been made with submicron dimensions, as shown in figure 6.

In typical high-speed machines the delay of a signal as it goes through one logic stage is made up of three roughly equal parts. One part is the time taken to go through the logic stage, if we assume that the stage need drive only one nearby subsequent stage. Another third of the delay comes from the extra capacitances that arise because the signal is typically passed on to several stages. The third contribution to the delay comes from the finite velocity of pulse propagation through the machine. Existing large high-speed machines have a delay of about 5 nanoseconds per stage, or about 2–1.5 nanoseconds, through the isolated, lightly loaded, logic stage. The fastest experimental logic circuits existing today in the laboratory are about ten times faster.

Keyes[11] has calculated that we can go about another factor of ten beyond

that before we will be making our transistors so small that they can not effectively pass their dissipation on to the surrounding crystal. This, then, is how far technology can take the conventional transistor without the invention of new transistor types or the use of esoteric materials.

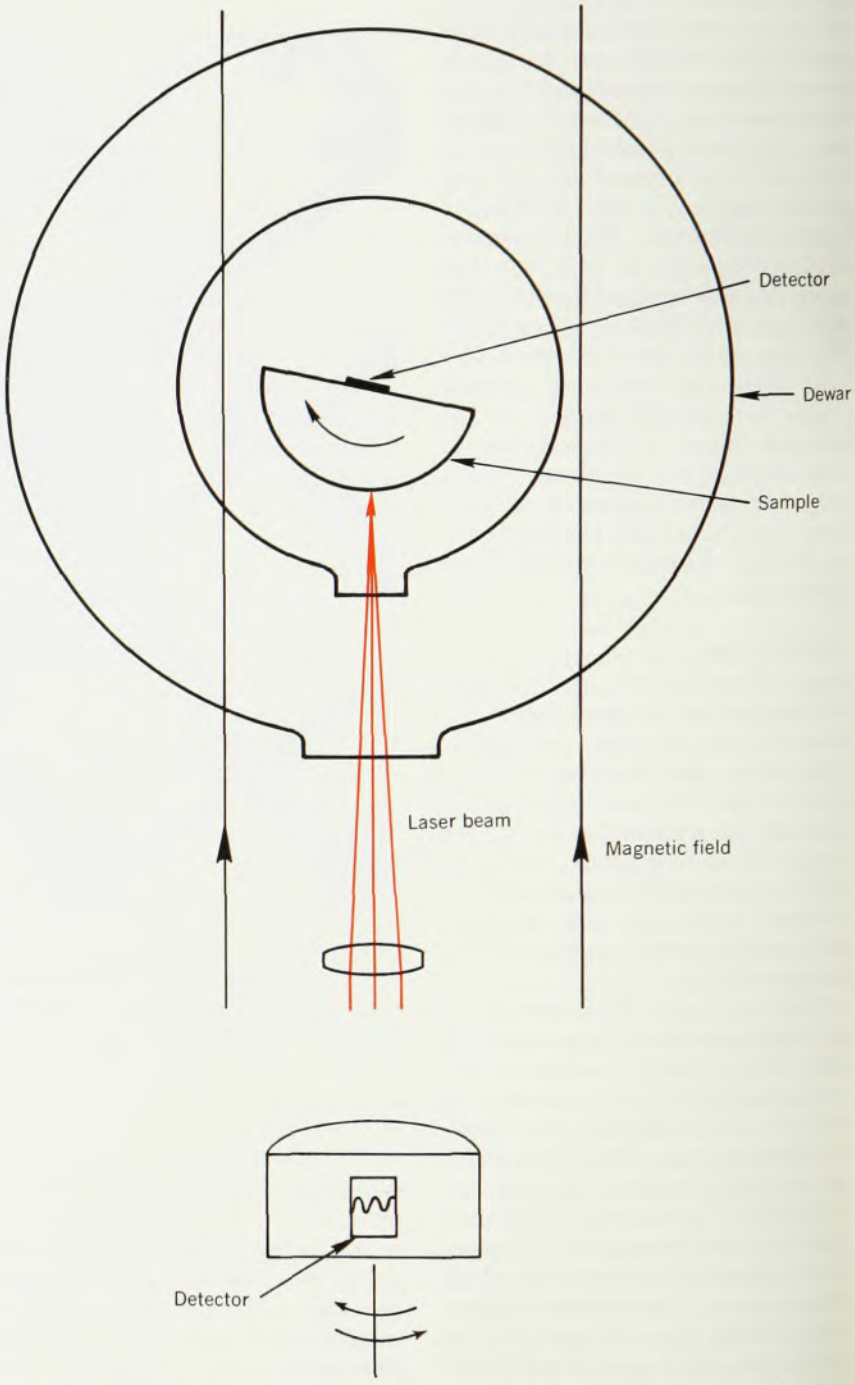
Carrier mobility

Finally the transistor speed is affected by carrier mobility. Figure 7 shows that germanium and silicon, at room temperature, have rather modest mobilities compared with those available in some other materials and at lower temperatures.. The figure shows mobilities at low fields and in pure materials, and thereby overstates the advantage to be had through the choice of materials. High-mobility semiconductors, furthermore, tend to suffer from the fact that they adversely effect the value of thermal conductivity and of dielectric constant. A forthcoming paper by R. W. Keyes, E. P. Harris and K. Konneth analyzes these drawbacks in detail.

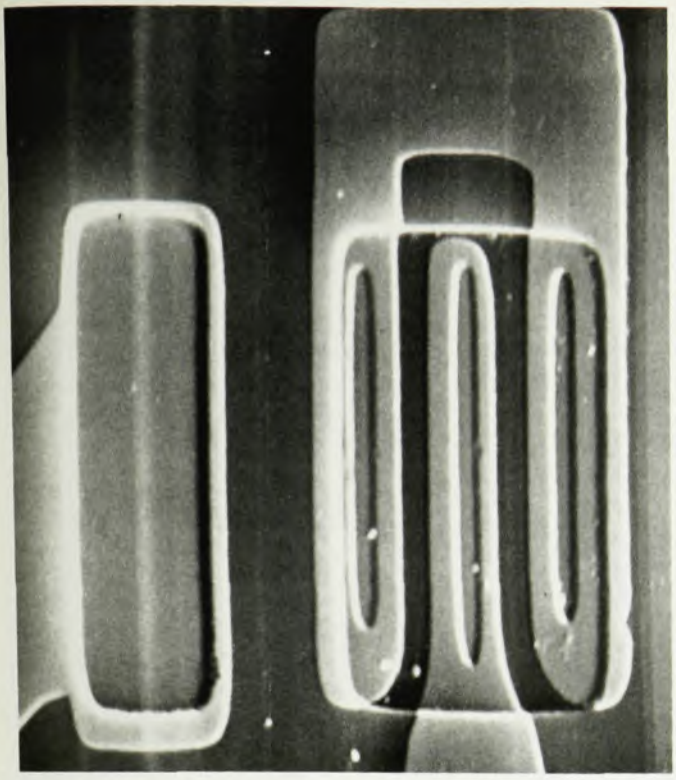
Nevertheless, figure 7 does show one inducement for ingenuity in device design. Remember that our ordinary junction transistor depends for its action on the motion both of holes and of electrons, and therefore on two mobilities. Other transistor types, however, such as the structure shown on the cover, depend on only one carrier mobility and can take advantage of materials that have high electron mobilities. Naturally the introduction of a novel material poses tremendous development tasks; germanium and silicon have had decades of effort to bring the materials under precise control. By contrast, compound semiconductors not only have had far less effort but they are, intrinsically, materials with more degrees of freedom for behavior or misbehavior, and are therefore harder to control.

Whereas figure 7 makes it clear that low temperatures can lead to high mobilities, there are other reasons for expecting, in the long run, more attention to low temperatures for very high-speed processors. The voltage level required to control the flow of thermal electrons, and therefore the power dissipation, can be expected to decrease with kT. Furthermore, metallic conductivities, which in the final analysis determine the size of the computer by their effect on transmission-line attenuation, can be expected to improve modestly with lowered temperature, thereby permitting a more compact computer structure. Finally we have seen evidence of "roadwear" in computers.[12] This term means deterioration as a result of usage. The momentum associated with current flow can be transferred to atoms, dragging them along, and thus changing the physical structure. Such deterioration processes represent thermally activated atomic jumps and can be drastically reduced with temperature.

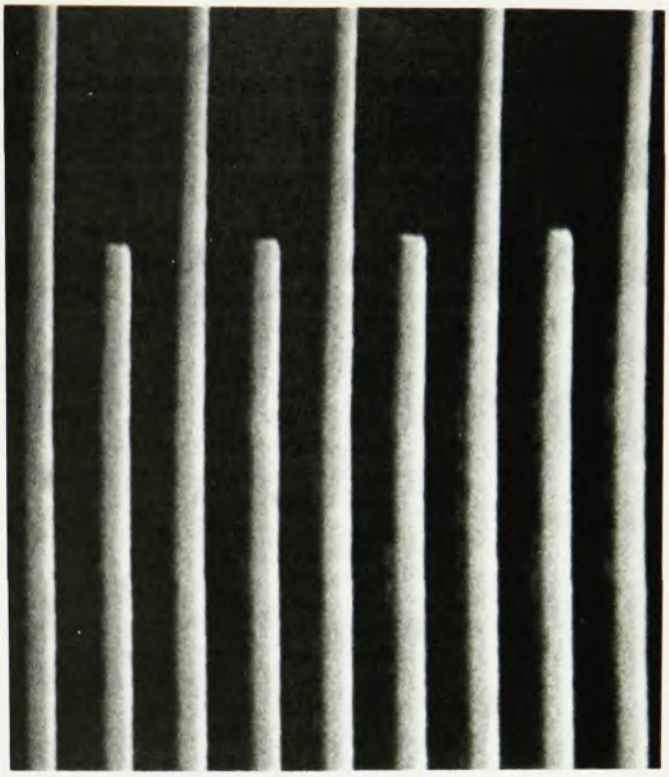
We have now seen a number of avenues that are available for the speedup of the logic process. The chief bottleneck is the investment in know-how represented by the high density of integration, illustrated in



HEAT TRANSPORT experiment. Laser beam is incident on a single crystal of gallium at 1.8 K, generating heat pulses. On the far side of the crystal a serpentine-shaped superconducting bolometer, controlled by the magnetic field, detects heat flowing to that surface. The first indications of heat flow in each pulse represent the unscattered ballistic flight of electrons. —FIG. 4



EXPERIMENTAL TRANSISTOR made by electron-beam photoresist exposure. The narrowest controlled dimension is 0.8 micron. This view is from reference 9. —FIG. 5



INTERLACING ELECTRODES deposited on a lithium niobate ($LiNbO_3$) crystal as a surface-wave transducer. Each electrode is 0.3 microns wide. From reference 10. —FIG. 6

figure 3. Any serious departure from current technologies takes us into new processes with lower yields, and must, therefore, come initially at the price of fewer devices per unit. In the final analysis, of course, it is the work carried out by such a chip that counts and not really the speed of the individual transistor on it.

Ultimate limitations

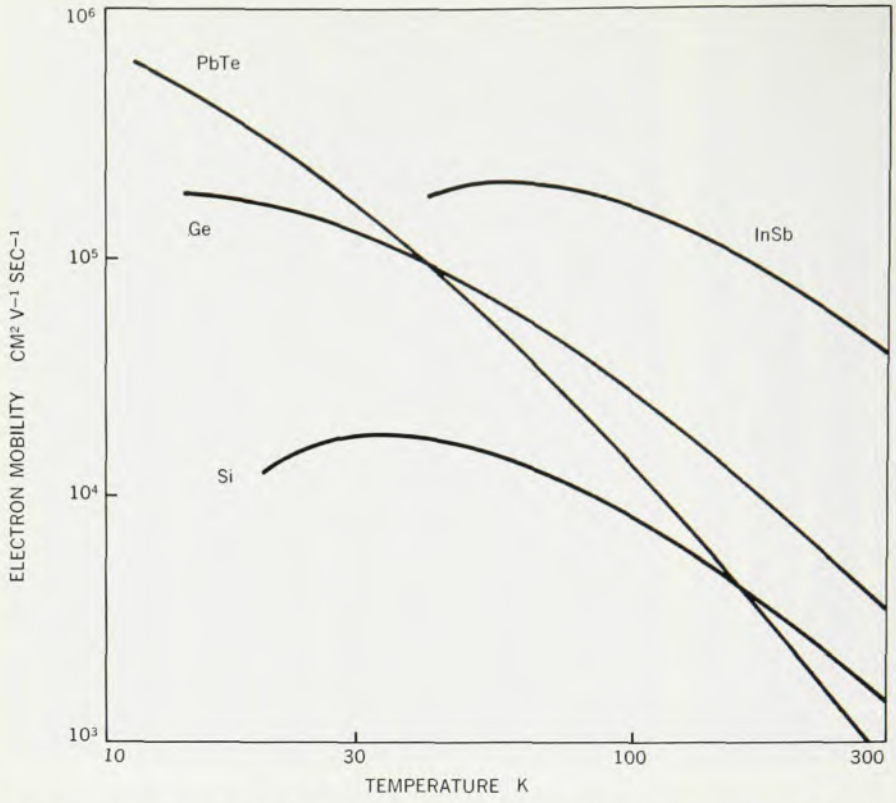
The rich potential for progress sketched above naturally leads to the question: How far can we ultimately go? How is information processing restricted, once we recognize that, whether carried out in the brain, on paper or in the computer, it inevitably utilizes physical degrees of freedom?

Such questions are not new and, at least in spirit, were anticipated decades ago by P. W. Bridgman. Bridgman wrote one particularly engaging paper,[13] in which he extended his operational philosophy to mathematics. It is amusing, 35 years later, to find that the editor of the journal in which this item was published had to put a disclaimer at the beginning of the article: "As in the case of all articles . . . it is understood that this does not necessarily represent the views of the editors." The lapse of 35 years, however, seems to have made the viewpoint not much less controversial. After all, we are telling the mathematician: "Whether you are talking about the physical world or about abstract entities, in either case you are dependent on operations carried out in the physical world and therefore subject to the constraints of that world." This takes us to such basic questions as whether the universe has a finite or an infinite number of degrees of freedom, and whether the processes of handling information can be made arbitrarily immune to the deleterious effects of thermal fluctuation. Naturally, mathematicians are not ready to accept such an attack on their right to make their own rules.

The work in this area of fundamental limits is at a very rudimentary state; it has not gone much beyond developing a spirit for posing questions. Two articles summarize what is known.[14]

What have we actually learned? First of all, it has been argued that computers, in order not to choke on their own intermediate results, must have the ability to throw away information. This logical irreversibility can be tied to physical irreversibility, and thereby to the requirement for energy losses of the order kT per elementary logic step. This is of the order of 10^{10} less than is currently the case, showing that fundamental limits are so far from existing technology as to be useless as a guide to the engineer. Their importance is concerned more with the epistemological questions. There are (hypothetical) computing elements that, if we are satisfied with very slow computing, require dissipations not very much larger that the above-mentioned minimal values of order kT, derived from phase space and entropy considerations.

Fluctuation-dissipation theory ties these energy losses to noise sources; computers must have built-in noise sources. This naturally leads to the question whether computing can be made completely reliable in the presence of this noise. With the above-mentioned hypothetical computing elements, we can show that computing can be done to arbitrary reliability specifications, provided that we are willing to take arbitrarily long, and that we grant the physical realizability of certain interaction potentials required by the hypothetical devices. In other words, if computing can not be made noise-immune, it is not a result of statistical mechanics but



ELECTRON MOBILITIES for some pure semiconductors at low fields. Note that much higher mobilities exist than those of Ge and Si at room temperature. —FIG. 7

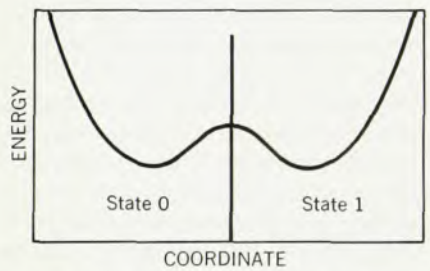
would either be the result of a demand for a minimal computing rate, or the result of the limited selection of interaction potentials available in the universe.

Along a different track there are several papers concerned with the finiteness of both memory content and communication-channel capacity available from a finite storehouse of material. Thus, for example,[15] an energy U used to send a message through a channel, or a set of parallel channels, at temperature T_0 can not transmit more than $U/(kT_0 \log 2)$ bits (binary choices) of information. It is not clear what the effective temperature of the universe is, or its energy content. If we take 10^{56} grams as the mass of the universe and $T_0 \gtrsim 1$K, we find 10^{93} bits available if we burn up the universe to send our message. Regardless of how we choose U and T_0, we are likely to get an exponent of order 100, rather than 10^6. 10^{93} may seem like a large number, but it is small when we compare it with the kinds of numbers available from combinatorial problems, for example the

$$10^{2.4 \times 10^9}$$

different possible variations of the human genetic structure.[16]

We also have the beginnings of a theory that analyzes active dissipative devices, such as transistor flip-flop circuits, and treats their stochastic behavior. There turns out to be a remarkable parallel between such systems and static bistable systems, such as ferromagnets, ferroelectrics, and the ammonia molecule. For the static systems, the energy, as a function of some information-bearing degree of freedom, typically has the form sketched in figure 8. For the static system, the Boltzmann factor $\exp(-U/kT)$ is basic to the behavior of the system, and in particular it tells us what the probability of a thermally activated jump between the two information states is. We have learned now how to generalize the Boltzmann factor to some dissipative bistable systems and thus to calculate the unintentional rate of information loss in them.



ENERGY as a function of the information-bearing degree of freedom for a typical simple bistable system. The two potential wells correspond to the two stable states of the system. —FIG. 8

We thus find ourselves at a point in the technological development of the computer where we are attempting to cope with a physical, rather than a mathematical or philosophical, science of knowledge and information.

* * *

This article was adapted from a talk given at the American Institute of Physics Corporate Associates Meeting in October 1969.

References

1. R. W. Keyes, J. A. Armstrong, Appl. Optics, **8**, 2549 (1969).
2. J. Matisoo, Proc. IEEE, **55**, 172 (1967).
3. A. H. Bobeck, R. P. Fisher, A. J. Perneski, J. P. Remeika, L. G. Van Uitert, IEEE Trans. on Magnetics, **5**, 544 (1969).
4. M. J. Freiser, P. M. Marcus, IEEE Trans. on Magnetics, **5**, 82 (1969).
5. A. B. Fowler, F. F. Fang, W. E. Howard, P. J. Stiles, Phys. Rev. Lett. **16**, 901 (1966).
6. K. E. Drangeid, R. Sommerhalder, W. Walter, Electronics Letters, **6**, 228 (1970).
7. W. S. Boyle, G. E. Smith, Bell System Tech. J., **49**, 587 (1970); G. F. Amelio, M. F. Tompsett, G. E. Smith, Bell System Tech. J., **49**, 593 (1970).
8. R. J. von Gutfeld, A. H. Nethercot, Phys. Rev. Lett., **18**, 855 (1967).
9. M. Hatzakis, S. Magdo, C. H. Ting, Proceedings of the IVth Int. Conf. on Electron and Ion-Beam Science and Technology, Los Angeles, May 1970 (R. Bakish, ed.) Electrochemical Society, New York.
10. A. N. Broers, E. G. Lean, M. Hatzakis, Appl. Phys. Lett., **15**, 98 (1969).
11. R. W. Keyes, IEEE Spectrum **6**, 36 (1969).
12. I. A. Blech, E. S. Meieran, J. Appl. Phys. **40**, 485 (1969); J. R. Black, Proc. IEEE **57**, 1587 (1969); R. Rosenberg, Appl. Phys. Lett. **10**, 27 (1970).
13. P. W. Bridgman, Scripta Math., **2**, 224 (1934).
14. R. Landauer, Proceedings of the conference on "Fluctuation Phenomena in Classical and Quantum Systems," Chania, Crete, Greece, August 1969, (E. D. Haidemenakis, ed.), Gordon and Breach, New York.
15. H. Marko, Kybernetik **2**, 274 (1965).
16. H. J. Bremermann, Progr. Theoret. Biol., (F. M. Snell, ed.), Academic, New York, 1967, page 59. □