As the transistor has grown smaller and cheaper, engineers have scoffed at theoretical barriers to its progress—so far

THE FUTURE OF THE TRANSISTOR

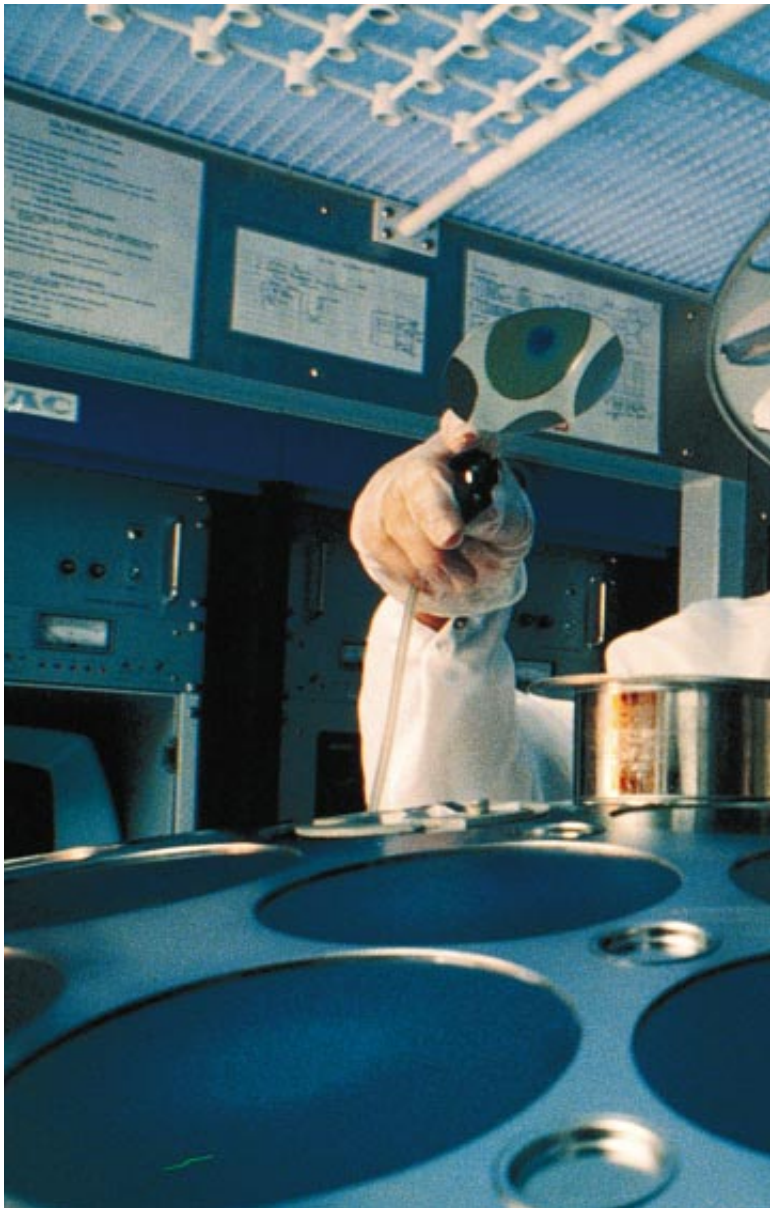
by Robert W. Keyes

I am writing this article on a computer that contains some 10 million transistors, an astounding number of manufactured items for one person to own. Yet they cost less than the hard disk, the keyboard, the display and the cabinet. Ten million staples, in contrast, would cost about as much as the entire computer. Transistors have become this cheap because during the past 40 years engineers have learned to etch ever more of them on a single wafer of silicon. The cost of a given manufacturing step can thus be spread over a growing number of units.

How much longer can this trend continue? Scholars and industry experts have declared many times in the past that some physical limit exists beyond which miniaturization could not go. An equal number of times they have been confounded by the facts. No such limit can be discerned in the quantity of transistors that can be fabricated on silicon, which has proceeded through eight orders of magnitude in the 50 years since the transistor was invented [*see box on pages 50 and 51*].

I do not have a definitive answer to the question of limits. I do, however, have some thoughts on how the future of solid-state electronics will develop and what science is needed to support continuing progress.

Several kinds of physical limitations might emerge as the size of the transistor continues to shrink. The task of connecting minute elements to one another might, for example, become impossible. Declining circuit size also means that researchers must cope with ever stronger electrical fields, which can affect the movement of electrons in many ways. In the not too distant future the transistor may span only hundreds of angstroms. At that point, the presence or absence of single atoms, as well as their behavior, will become significant. Diminishing size leads to increasing density of transistors on a chip, which raises the amount of waste heat thrown off. As the size of circuit elements drops below the wavelength of usable forms of radiation, exist-



MINIATURIZATION has made transistors cheaper than staples by spreading manufacturing costs over millions of devices on each of the hundreds of chips on a wafer. This worker holds a nearly completed wafer. Its components will be connected by the condensation of metal in a vacuum chamber (*foreground*).

ing manufacturing methods may reach their limits.

To see how such problems might arise and how they can be addressed, it is useful to review the operation of the field-effect transistor, the workhorse of modern data processing. Digital computers operate by manipulating statements made in a binary code, which consists of ones and zeroes. A field-effect transistor is operated so that, like a relay, it is switched only "on" or "off." The device therefore represents exactly one binary unit of information: a bit. In a large system, input signals control transistors that switch signal voltages onto output wires. The wires carry the signals to other switches that produce outputs, which are again sent on to another stage. The connections within the computer determine its function. They control the way that the inputs are transformed to become outputs, such as a word in a document or an entry in a spreadsheet.

From Source to Drain

The field-effect transistor contains a channel that interacts with three electrodes: a source, which supplies electrons to the channel; a drain, which receives them at the other side; and a gate, which influences the conductivity of the channel [*see illustration on next page*]. Each part contains different impurity atoms, or dopants, which modify the electrical properties of the silicon.

The gate switches the transistor on when a positive voltage applied to it attracts electrons to the interface between the semiconductor and the gate insulator. These electrons then establish a connection between the source and drain electrodes that allows current to be passed between them. At this point, the transistor is "on." The connection persists for as long as the positive charge remains on the gate. An incoming signal is applied to the gate and thus determines whether the connection between source and drain is established. If a connection results, the output is connected to the ground potential, one of the standard digital voltages. If no connection results, the output is connected through the resistor to the positive power supply, the other standard digital voltage.

Circuits of transistors must be oblivious to the operations of neighboring arrays. Existing concepts of insulation,

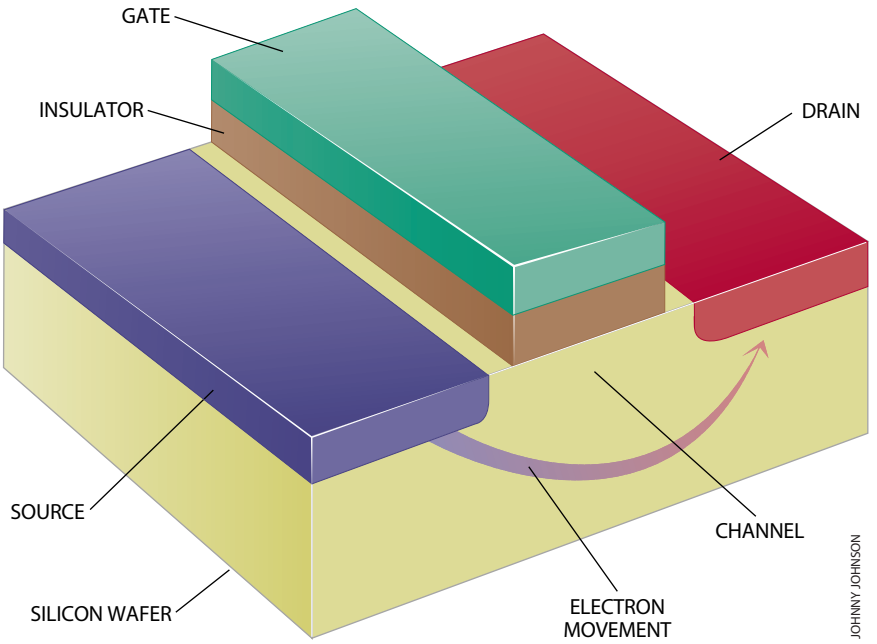
IBM CORPORATION

impedance and other basic electrical properties of semiconductors and their connections should work well enough, for designers' purposes, in the next generation of devices. It is only when conducting areas approach to within about 100 angstroms of one another that quantum effects, such as electron tunneling, threaten to create problems. In laboratory settings, researchers are already at the brink of this limit, at about 30 angstroms; in commercial devices, perhaps a decade remains before that limit is reached.

Another challenge is the strengthening of the electrical field that inevitably accompanies miniaturization. This tendency constrains the design of semiconductor devices by setting up a basic conflict. Fields must continually get stronger as electron pathways shrink, yet voltages must remain above the minimum needed to overwhelm the thermal energy of electrons. In silicon at normal operating temperatures, the thermal voltage is 0.026 electron volt. Therefore, whenever a semiconductor is switched so as to prevent the passage of electrons, its electrical barrier must be changed by a factor several times as large. One can minimize the thermal problem by chilling the chip (which becomes an expensive proposition).

It has only recently been taken for granted that anyone can search for references to anything from kiwifruit to quantum physics.

Even cooling cannot end the problem of the electrical field. Signals must still have the minimum voltage that is characteristic of a semiconductor junction. In silicon this electrical barrier ranges between half a volt and a volt, depending on the degree of doping. That small voltage, applied over a very short distance, suffices to create an immensely strong electrical field. As electrons move through such a field, they may gain so much energy that they stimulate the creation of electron-hole pairs, which are themselves accelerated. The resulting chain reaction can trigger an avalanche of rising current, thereby disrupting the circuit. Today's chips push the limits in the quest for high speed, and electrical fields are usually close to those that can cause such avalanches.



FIELD-EFFECT TRANSISTOR, the workhorse of data processing, is built as a sandwich of variously doped silicon layers. It contains a channel, a source, a drain and an insulated gate. When a positive voltage is applied to the gate, electrons move near the insulation, establishing a connection underneath it that allows current to pass from source to drain, switching the transistor on.

Tricks and Trade-offs

Workers resort to a variety of tricks to mitigate the effects of strong electrical fields. They have designed field-effect transistors, for example, in which the field can be moved to a place where it does not disrupt other electronic functions. This stratagem is just one of many, all of which entail trade-offs with other desired characteristics, such as simplicity of design, ease of manufacture, reliability and long working life.

Miniaturization also increases the heat given off by each square centimeter of silicon. The reason is purely geometric: electrical pathways, and their associated energy losses, shrink in one dimension, whereas chip area shrinks in two. That relation means that as circuits get smaller, unit heat generation falls, albeit more slowly than does the number of units per square centimeter.

Devices already pour out as much as 30 watts per square centimeter, a radiance that one would expect of a material heated to about 1,200 degrees Celsius (this radiance value is about 10 times that of a range-top cooking surface in the home). Of course, the chips cannot be allowed to reach such temperatures, and so cooling systems remove heat as fast as it is produced. A variety of cooling technologies have been devised, including some rather intense ones. But the cost of using them in transistor circuits increases rapidly when the density of heat increases.

The exigencies of manufacturing impose constraints on the performance of electronic devices that might not be apparent from a purely theoretical discussion. Low-cost manufacturing results in small differences among the devices that are made on each wafer, as well as among those that are fabricated on different wafers. This variability cannot be banished—it is inherent in the way solid-state devices are made.

A semiconducting material, such as silicon, is made into a transistor in an integrated process involving many steps. Templates, called masks, are applied to the silicon in order to expose desired areas. Next, various operations involving chemical diffusion, radiation, doping, sputtering or the deposition of metal act on these areas, sometimes by constructing device features, other times by erect-

ing scaffolding to be used in succeeding steps and then torn down. Meanwhile other devices—resistors, capacitors and conductors—are being built to connect the transistors.

Variations intrude at every step. For example, perfect focusing of the source of radiation over a large wafer is hard to achieve. The temperature of the wafer may vary slightly from one place to another during processing steps, causing a difference in the rate of chemical reactions. The mixing of gases in a reaction chamber may not be perfect. For many reasons, the properties of devices on a given wafer and between those on different wafers are not identical. Indeed, some devices on a wafer may be no good at all; the proportion of such irremediable errors places a practical limit on the size of an integrated circuit.

A certain amount of fuzziness is inherent in optical exposures. The light used in photolithography is diffracted as it passes through the holes in the template. Such diffraction can be minimized by resorting to shorter wavelengths.

When photolithographic fabrication was invented in the early 1970s, white light was used. Workers later switched to monochromatic laser light, moving up the spectrum until, in the mid-1980s, they reached the ultraviolet wavelengths. Now the most advanced commercial chips are etched by deep ultraviolet light, a difficult operation because it is hard to devise lasers with output in that range. The next generation of devices may require x-rays. Indeed, each generation of circuitry requires manufacturing equipment of unprecedented expense.

Other problems also contribute to the cost of making a chip. The mechanical controls that position wafers must become more precise. The "clean rooms" and chambers must become ever cleaner to ward off the ever smaller motes that can destroy a circuit. Quality-control procedures must become even more elaborate as the number of possible defects on a chip increases.

Device "Sandwich"

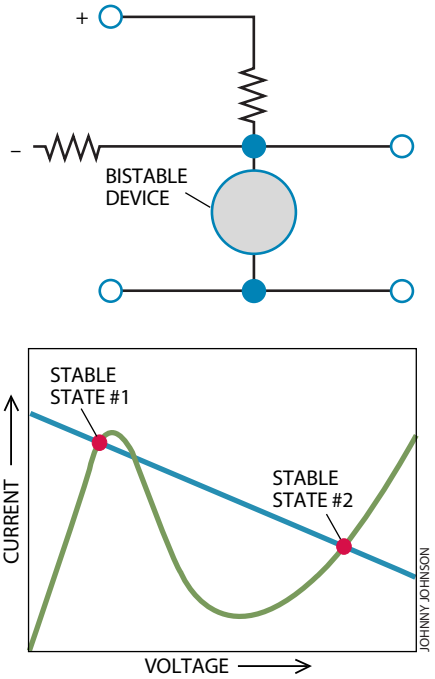
Miniaturization may at first glance appear to involve manipulating just the width and breadth of a device, but depth matters as well. Sometimes the third dimension can be a valuable resource, as when engineers sink capacitors edgewise into a chip to conserve space on the surface. At other times, the third dimension can constrain design. Chip designers must worry about the aspect ratio—that is, the relation of depth to surface area. The devices and connections on chips are built up in the silicon and on the surface as a series of layers resembling a sandwich. Part of the art of making devices smaller comes from using more layers. But the more layers there are, the more carefully controlled each must be, because each is affected by what is beneath it. The number of layers is limited by the costs of better control and more connections between layers.

The formulas that are used to design large devices cannot be used for the tiny transistors now being made in laboratories. Designers need to account for exotic new phenomena that appear in such extremely small devices. Because the effects cannot be accurately treated by purely analytic methods, the designers must have recourse to computer models that are able to simulate the motion of electrons in a device.

A computer follows a single electron through a device, keeping track of its position as time is increased in small steps. Physical theory and experimental information are used to calculate the probability of the various events that are possible. The computer uses a table for the probabilities, stored in its memory, and a random number generator to simulate the occurrence of these events. For example, an electron is accelerated by an electrical field, and the direction of its motion might be changed by a collision with an impurity. Adding the results of thousands of electrons modeled in this fashion gives a picture of the response of the device.

Consider the seemingly trivial question of how to represent the motion of an electron within an electrical field. When path lengths were comparatively long, an electron quickly accelerated to the point at which collisions robbed it of energy as fast as the field supplied new energy. The particle therefore spent most of its time at a constant velocity, which can be modeled by a simple, linear equation. When path lengths became shorter, the electron no longer had time to reach a stable velocity. The particles now accelerate all the time, and the equations must account for that complication.

If such difficulties can arise in modeling a well-understood phenomenon, what lies ahead as designers probe the murky physics of the ultrasmall? Simulations can be no better than the models that physicists make of events that happen in small spaces during short periods. To refine these models, researchers need to carry out experiments on femtosecond timescales.



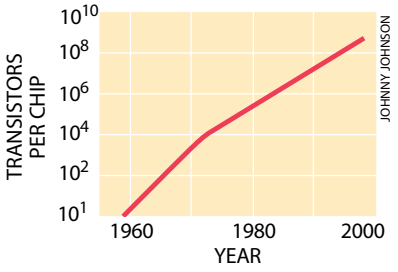
BISTABLE CIRCUIT does the transistor's job by exploiting nonlinear effects. A device such as a tunnel diode is placed at the junction of two main electrodes and a minor one (*top*). If the minor electrode injects some extra current, the circuit will move from one stable state to the other (*bottom*). Such devices are impractical because they cannot tolerate much variation in signal strength.

Remaining Unknowns

Expanded knowledge of solid-state physics is required, because as chips grow more complex they require more fabrication steps, and each step can influence the next. For instance, when doping atoms are introduced into a crystal, they tend to attract, repel or otherwise affect the motion of other dopants. Such effects of dopants on other dopants are not well understood; further experiments and theoretical investigations are therefore needed. Chemical reactions that take place on the surface of a silicon crystal demand a supply of silicon atoms, a kind of fluid flow within the solid lattice; does such motion carry other constituents along with it? These questions did not concern designers of earlier generations of chips, because existing transistors were then large enough to swamp such ultramicroscopic tendencies.

The Shrinking Transistor

Miniaturization is manifest in this comparison between an electromechanical switch, circa 1957 (*background*), and a chip containing 16 million bits of memory (*foreground*). Progress appears in these snapshots (*below, left*): Bell Laboratories's first transistor; canned transistors; salt-size transistors; a 2,000-bit chip; a board with 185,000 circuits and 2.3 megabits of memory; and a 64-megabit memory chip. —*R.W.K.*

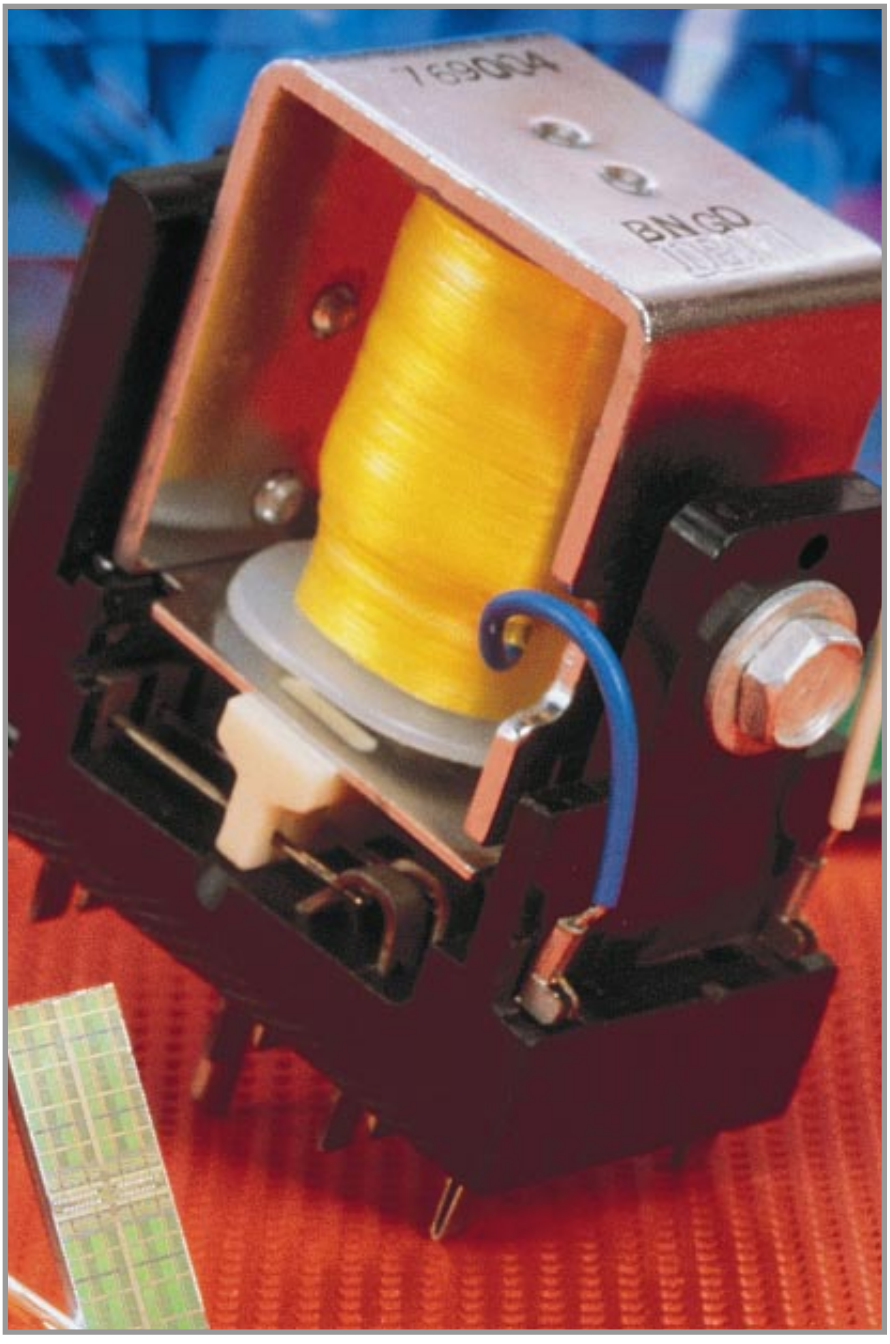


The first transistor
1948

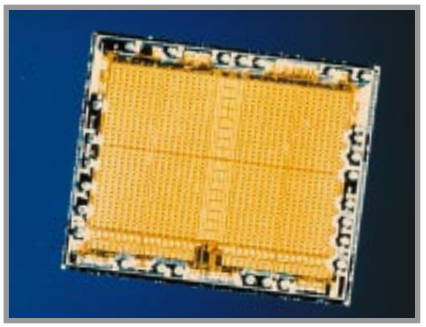
Early commercial transistors
1958

IBM CORPORATION (*photographs*)

Salt-size transistors
1964



Switches, now and then



Early integrated circuit
1973

Circuit assembly
1985

Dynamic random-access memory chip
1997

The prospect of roadblocks aside, the transistor has only itself to blame for speculation about alternative technologies. Its extraordinary success in the 1950s stimulated an explosive development of solid-state physics. In the course of the work, investigators discovered many other phenomena, which in turn suggested a host of ideas for electronic devices. Several of these lines of research produced respectable bodies of new engineering knowledge but none that led to anything capable of even finding a niche in information processing.

Some workers have argued that the transistor owes its preeminence to having been the first off the block. Because of that head start, semiconductors have been the center of research, a position that guarantees them a margin of technological superiority that no rival can match. Yet I believe the transistor has intrinsic virtues that, in and of themselves, could probably preserve its dominant role for years to come.

I participated, as a minor player, in some of the efforts to build alternative switches, the repeated failures of which made me wonder what was missing. Of course, quite a few new fabrication methods had to be developed to implement a novel device concept. But even though these could be mastered, it was difficult to get a large collection of components to work together.

What gave the transistor its initial, sudden success? One difference stood out: the transistor, like the vacuum tube before it, has large gain. That is, it is capable of vastly amplifying signals of the kind processed in existing circuits, so that a small variation in input can produce a large variation in output. Gain makes it possible to preserve the integrity of a signal as it passes through many switches.

Rivals to the transistor may have been equally easy to miniaturize, but they exhibited far less gain. Take, for instance, bistable devices [*see illustration on page 49*], which perform logic functions by moving between two stable states that are separated by an unstable transition. Researchers have produced such a transition by designing circuits having a range of values in which current declines as voltage increases. Any slight disturbance, such as that obtained by injecting extra current through the device, will switch the circuit between its two stable states.

Because this slight input can bring about large changes in the current and voltages, there is a sense in which gain is achieved. Yet the gain is far less useful than that provided by an ordinary transistor because it operates within rather narrow tolerances. A bistable switch thus performs deceptively well in the laboratory, where it is possible to fine-tune the circuit so it stays near enough to the crossover point. A collection of such switches, however, does not lend itself to such painstaking adjustments. Because not all the circuits will work, no complex device can be based on their operation. Negative resistance therefore plays no role in practical data processing.

The same difficulty has plagued the development of nonlinear optical devices, in which the intensity of optical beams replaces the currents and voltages of electrical circuits. Here, too, the operation depends on fine-tuning the system so that a small input will upset a delicate balance. (Such switches have occasionally been termed optical transistors, a label that misconstrues the principles of transistor action.)

Optical switches face a problem even more fundamental. Light, unlike electricity, hardly interacts with light, yet the interaction of signals is essential for logic functions. Optical signals must therefore be converted into electrical ones in a semiconductor. The voltage thus produced changes the optical response of another material, thereby modulating a beam of light.

Useful Interference

Another proposed switch, sometimes called a quantum interference device, depends on the interference of waves. In the most familiar case, that of electromagnetic radiation, or light, one wave is divided into two components. The components begin oscillating in phase—that is, their peaks and troughs vibrate in tandem. If the components follow routes of different lengths before reuniting, the phase relation between their waveforms will be changed. Consequently, the peaks and troughs either cancel or reinforce one another, producing a pattern of bright and dark fringes. The displacement of the fringes measures the relative phase of the system.

Electrons also possess a wave nature and can be made to interfere. If the two components of a wave move at equal speeds over similar paths to a rendezvous, they will reconstitute the original wave; if they move at different speeds, they will interfere. One can manipulate

IBM CORPORATION

IMMENSE AND DENSE: this active-matrix liquid-crystal panel shows that today's electronic structures can achieve great complexity over large areas. Each liquid-crystal pixel is controlled by its own transistor, providing extraordinary resolution.

the velocity of one wave by applying a tiny electrical field to its pathway. The correct field strength will cause the waves to cancel so that no current can flow through the device.

At first sight, this action duplicates a field-effect transistor, which uses an electrical field to control a current through a semiconductor. In an interference device, however, conditions must be just right: if the applied voltage is too high or too low, there will be some current. This sensitivity means that an interference device will not restore the binary nature of a degraded input signal but will add its own measure of noise. Data passing from one such device to another will quickly degenerate into nothingness.

The Only Game in Town

The lack of real rivals means that the future of digital electronics must be sought in the transistor. The search begins anew with each voyage into a smaller scale or a different material. The latest reexamination was occasioned by the introduction of new semiconductor materials, such as gallium arsenide and related compounds, several of which may even be incorporated to achieve some desired characteristic in a single device. These combinations may be used to produce what are called heterojunctions, in which crystalline lattices of different energy gaps meet. Lattices may mesh imperfectly, creating atomic-scale defects, or they may stretch to one another, creating an elastic strain. Either defects or strain can produce electrical side effects.

These combinations complicate the physics but at the same time provide a variable that may be useful in surmounting the many design problems that miniaturization creates. For instance, the dopants that supply electrons to a semiconductor also slow the electrons. To reduce this slowing effect, one can alternate layers of two semiconductors in which electrons have differing energies. The dopants are placed in the high-energy semiconductor, but the electrons they donate immediately fall into the lower-energy layers, far from the impurities.

What, one may ask, would one want with a technology that can etch a million transistors into a grain of sand or put a supercomputer in a shirt pocket? The answer goes beyond computational power to the things such power can buy in the emerging information economy. It has only recently been taken for granted that anyone with a personal computer and a modem can search 1,000 newspapers for references to anything that comes to mind, from kiwifruit to quantum physics. Will it soon be possible for every person to carry a copy of the Library of Congress, to model the weather, to weigh alternative business strategies or to checkmate Garry Kasparov? SA

The Author

ROBERT W. KEYES is a research staff member at the IBM Thomas J. Watson Research Center in Yorktown Heights, N.Y. His interests have focused on semiconductor physics and devices and on the physics of information-processing systems, subjects on which he has written and lectured widely; he has also received eight issued patents. A native of Chicago, Keyes studied at the University of Chicago, where he earned a doctorate in physics. He is an active participant in the programs of the National Research Council, the American Physical Society and the Institute of Electrical and Electronics Engineers.

Further Reading

FIELD-EFFECT TRANSISTORS IN INTEGRATED CIRCUITS. J. T. Wallmark and L. G. Carlstedt. John Wiley & Sons, 1974.

THE PHYSICS OF VLSI SYSTEMS. R. W. Keyes. Addison-Wesley Publishing, 1987.

CAN WE SWITCH BY CONTROL OF QUANTUM MECHANICAL TRANSMISSION? Rolf Landauer in *Physics Today,* Vol. 42, No. 10, pages 119–121; October 1989.

LIMITATIONS, INNOVATIONS, AND DEVICES INTO A HALF MICROMETER AND BEYOND. M. Nagata in *IEEE Journal of Solid-State Circuits,* Vol. 27, No. 4, pages 465–472; April 1992.

THE FUTURE OF SOLID-STATE ELECTRONICS. R. W. Keyes in *Physics Today,* Vol. 45, No. 8, pages 42–48; August 1992.