PHYSICS AND THE INFORMATION REVOLUTION

Quantum physics holds the key to the further advance of computing in the postsilicon era.

Joel Birnbaum and R. Stanley Williams

In the fourth century BC, a young man named Pythias was condemned to death by Dionysius, the tyrant of Syracuse, for plotting against him, but Pythias was granted three days' leave to go home to settle his family's affairs after his friend Damon agreed to take his place and be executed should Pythias not return. Pythias encountered many problems but managed to return just in time to save Damon. Dionysius was so struck by this remarkable and honorable friendship that he released them both.

The decades-old friendship between computer technology and physics has also been a remarkable and honorable one, and it, too, has produced salutary results. Present-day experimental and theoretical physicists depend on computing, and have incurred a debt that they have repaid many times over by making fundamental contributions to advances in hardware, software, and systems technologies. (Figure 1 shows an experimental computer and one of its developers.)

In this article, we discuss the physical and economic limits to the geometrical scaling of semiconductor devices that has been the basis of much of the computer industry's progress over the last 50 years. We then look at some of the options that may be available when we come up against fundamental physics barriers sometime after 2010.

Disruptive technology

The first stored-program electronic computer, ENIAC (the Electronic Numerical Integrator and Computer), was built in 1946. A major triumph for vacuum-tube technology, ENIAC could add 5000 numbers in one second. At that rate, it could calculate the trajectory of an artillery shell in only 30 seconds, whereas an expert human with a mechanical calculator would have needed some 40 hours to complete the task. The machine was large (see figure 2)—and expensive. ENIAC . . .

- ▷ Contained 17 468 vacuum tubes
- ▷ Weighed 60 000 pounds
- ▷ Occupied 16 200 cubic feet
- ▷ Consumed 174 kilowatts (233 horsepower).

The amount of energy ENIAC expended to compute a single shell trajectory was comparable to that of the explosive discharge required to actually fire the shell. ENIAC was still the fastest computer on Earth nine years later, when it was turned off because the US Army could no longer justify the expense of operating and maintaining it.

JOEL BIRNBAUM *is chief scientist at Hewlett-Packard, in Palo Alto, California.* STANLEY WILLIAMS *is a senior principal laboratory scientist at Hewlett-Packard Laboratories.*

Even in the early days of ENIAC, though, technologists dreamed of smaller, faster, and far-more-reliable computers. An article by a panel of experts in the March 1949 issue of *Popular Mechanics* confidently predicted that someday a computer as powerful as ENIAC would contain only 1500 vacuum tubes, weigh only 3000 pounds, and require a mere 10 kilowatts of power to operate. Such a machine would be about the size and weight of an automobile, said the experts, with power consumption to match. What was intended to be a bold projection seems quaintly conservative to us now. These days, a palmtop computer is thousands of times more powerful than ENIAC was.

The reason for the experts' now-laughable error is that their prediction was based on the wrong foundation—reasonable extrapolation of the in-place vacuum-tube technology. The transistor, which had already been invented and represented a disruptive technology—that is, a technology that could totally displace vacuum tubes in computers, as electronic calculators later replaced slide rules—was completely ignored.

By 1949, after 40 years of development, vacuum-tube technology was mature, and the associated manufacturing infrastructure was enormous. In 1938 the vacuum tube had still been a decade away from its ultimate accomplishment. But already there was a significant search for something that would be better: a solid-state switch. The development of that switch required a great deal of basic research, both in materials purification and in device concepts.

Even though transistors as discrete devices had significant advantages over vacuum tubes and progress on transistors was steady during the 1950s, the directors of many large electronics companies believed that the vacuum tube held an unassailable competitive position.

Their companies were eventually eclipsed by the ones that invested heavily in transistor technology R&D and that were poised to exploit new advances. As we shall see, there are eerie parallels with the situation today.

Moore's law

Gordon Moore of Intel Corp was the first to quantify the steady improvement in gate density when he noticed that the number of transistors that could be built on a chip increased exponentially with time. (See figure 3.) Over the past 28 years, that exponential growth rate has corresponded to a factor-of-four increase in the number of bits that can be stored on a memory chip in every device generation of about 3.4 years—an increase of 64 000 times!

This exponential growth in chip functionality is closely tied to the exponential growth of the chip market,

FIGURE 1. THE TERAMAC EXPERIMENTAL COMPUTER with one of its developers, Philip Kuekes. "Tera" denotes the fact that the machine performs one trillion gate operations per second (one million gates operating simultaneously at 1 megahertz); "mac" stands for "multiarchitecture computer." Teramac, built at Hewlett-Packard Laboratories in 1995, contains over 220 000 known manufacturing defects, yet operates perfectly.

which has been approximately doubling every five years.

At the present time, there are two recognized factors that could bring Moore's law scaling to an end. The first, according to Moore himself, is economic. The cost of building fabrication facilities to manufacture chips has also been increasing exponentially, by about a factor of two every chip generation. This is sometimes known as Moore's second law. (See figure 4.)

Thus, the cost of manufacturing chips is increasing significantly faster than the market is expanding. At some point, a saturation effect should slow the exponential growth to yield a classic s-curve for expanding populations.

In 1995, to build a single fabrication facility, or "fab," took about $1 billion, or about 1% of the entire annual chip market. By the year 2010, a fab could cost $30 billion to $50 billion—or about 10% of the total annual market at that time—if Moore's second law continues to hold.

The second factor threatening Moore's first law is that the engine that has brought the industry to this point, the complementary metal oxide semiconductor field-effect transistor (CMOS), can only take the technology part of the way to where it needs to go. The Semiconductor Industry Association has established a National Technology Roadmap that sets as a goal the continuation of the current exponential increases in capacity and performance up through the year 2012. (See figure 5.) That projection calls for making chips that are 256 times more capable than current CMOS designs, with no increase in power dissipation. If that goal is attained, the silicon-based integrated circuit will have accomplished a performance improvement of more than six orders of magnitude, using energy as a metric, with a single manufacturing paradigm. Compared to the advances experienced in most human endeavors, that increase would be extraordinary.

By 2010, the individual transistors in the circuits will be turned on or off by the addition or removal of only eight electrons on the gate of a transistor, compared to about 1000 electrons today. The statistics of small numbers will become significant, and the ability to distinguish between zero and one in a digital circuit will be severely compromised.

By 2020, the continuation of geometrical scaling would mean that less than one electron would be available to switch the transistor. That would require getting around a fundamental physical limitation, and not just an engineering obstacle. Yet, many researchers and corporate executives seem to have a blind optimism that somehow that will happen.

We think we cannot risk it. If there is to be any hope of sustaining the economic benefits to the national economy that come from sustaining Moore's law, then we have no choice but to develop quantum switches and the means to interconnect them.

Computational limits

The question of the fundamental limits to computation has been a subject of scholarly attention for decades, but has now become an essential issue. It does not make sense to make the enormous investments in research, development, and manufacturing that will be required to replace semiconductor switches by 2010 if the new technology is likely to perform only marginally better.

Rolf Landauer showed that information is a physical entity, and that therefore computation is a physical process. He proved that a nonreversible computer performing Boolean logic operations requires a minimum energy for a bit operation, $E_0 = kT \ln 2$, where k is Boltzmann's constant and T is the operating temperature of the system. That is the energy cost of throwing away one bit of information and increasing the entropy of the surroundings. At room temperature, the equation predicts that it is possible to perform 3.5×10^{20} bit operations per second with the expenditure of 1 watt of power. Obviously there would have to be a huge number of processes operating in parallel in any real system, but the calculation shows that we have lots of room beyond where

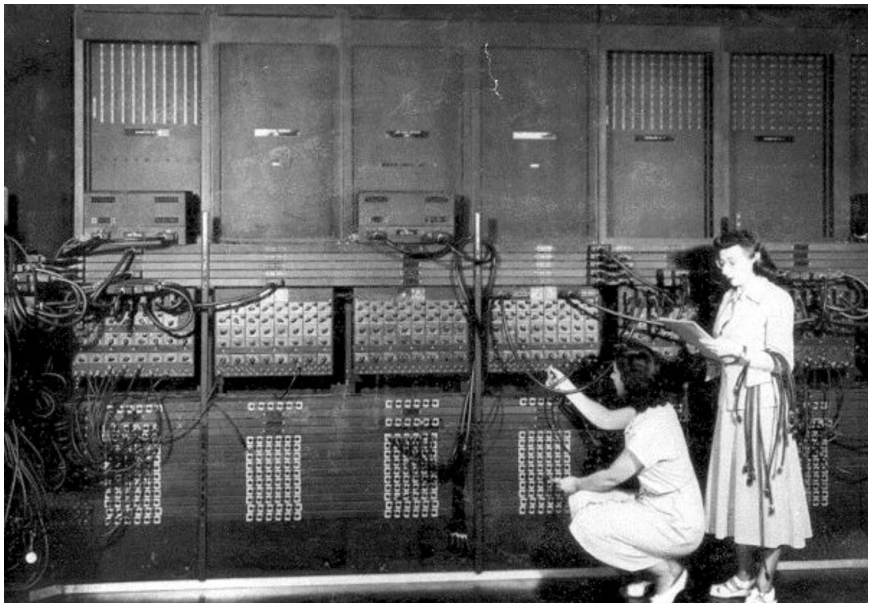

FIGURE 2. ENIAC, the first stored-program electronic computer, circa 1947. The computer, a small section of which is shown here, contained approximately 18 000 vacuum tubes and required 174 kilowatts of power to operate. The Intel 4004 microprocessor of 1971 could perform essentially the same tasks as ENIAC, but required only a few watts of power.

Moore's law will bring us with CMOS.

As with many other issues, Richard Feynman brought exceptional clarity to his analysis of the fundamental limits of computing. Using further thermodynamic considerations, he showed that the minimum amount of energy required—as opposed to the energy now dissipated in a resistive network—to transport a bit irreversibly from device to device in a computational system is $E_t = kT\, d\, \nu/c$, where d is the transmission distance, ν is the operating frequency, and c is the speed of light. This equation is in accord with the usual understanding of nonreversible processes, which cost more energy the faster they occur; but it also shows that smaller systems will expend less energy.

Even for extremely small systems, this energy is very large compared to that determined from the earlier computation. For a maximum information transport distance of 50 nm, Feynman's analysis shows that 10^{18} bit transfers per second will require 1 watt of power. Thus, for optimized nonreversible systems, the energy cost of communication in a computer will swamp that of the actual calculations—which is indeed the situation with today's integrated circuits.

A crude estimate of the energy required to add two 10-digit numbers using an ideal nonreversible computer is 100 bit operations. That implies that 3×10^{16} additions per joule can be performed at room temperature, a factor on the order of 10^9 times the estimated upper limit of silicon integrated-circuit technology in 2010.

Thus, even if the thermodynamic limit of efficiency for a nonreversible computer is never achieved, the fact that such a huge improvement is possible means that the search for new alternatives to the present technology is both prudent and potentially very rewarding. Such efficiency increases would allow either greater computational speed at constant power dissipation or smaller size for constant computational throughput. This is a computer architect's dream: We could build tailored desktop supercomputers or wristwatch-sized replacements for notebooks that would run for a lifetime on one battery.

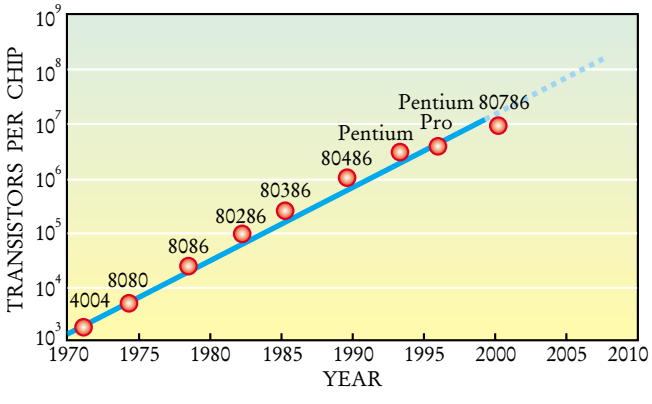

FIGURE 3. MOORE'S LAW is the empirical observation that the number of transistors on a single integrated-circuit chip increases by a factor of four every three years.

To achieve such incredible advances will require a totally different type of computational machinery. The requirement for inventing a new technology paradigm, coupled with the economic rewards that would follow from such a development, has created exciting research opportunities for mathematicians, physicists, chemists, and scientists of many disciplines, as well as for computer technologists. In fact, much of the current interest in interdisciplinary research areas such as nanofabrication, self-assembly, and molecular electronics is being driven by the search for a new archetype computer.

Thus, it seems as though the age of computation has not even begun, because even on a logarithmic scale there is further to go into the future than we have come from ENIAC.

Nanoscale devices

The implementation of some reversibility in a machine would provide even greater efficiency and capability. A number of alternatives to silicon-based field-effect transistors have been proposed, including single-electron transistors, quantum-cellular automata, and molecular logic devices.

A common theme that underlies many of these alternatives is the push to fabricate logic devices on the nanometer length scale—devices that will therefore be dominated by quantum mechanical effects. Such dimensions are more commonly associated with molecules than with integrated circuits, and it is not surprising that chemically assembled configurations, rather than artificially drawn structures, are expected to play an increasingly important role in the manufacture of new devices.

One very significant constraint on trying to manufacture the nanocircuitry of the future will be expense. Given Moore's second law, it is very unlikely that systems with

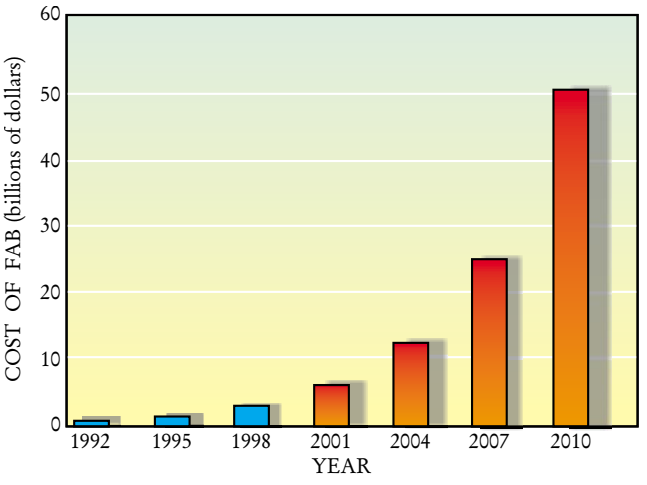

FIGURE 4. MOORE'S SECOND LAW extrapolation. The law is the empirical observation that the cost of fabrication facilities (fabs) for manufacturing integrated circuits has been increasing by a factor of two every three years. Given the huge cost of a fab predicted for the year 2010 (about $50 billion), it is likely that the exponential trend will not continue and that the cost will begin to level off soon.

feature sizes of a few nanometers will be made using traditional lithographic and subtractive processes, because the rules of scaling indicate that the cost of a facility for doing so would be equivalent to nearly the gross national product of the entire world.

Instead, at some point the cost advantage of using chemical assembly procedures to fabricate nanocircuitry will outweigh the disadvantages. At present, chemical assembly processes can produce nanocrystals as small as 10 nm in size directly on a surface, and just a few nanometers in size in solution growth. Islands of quantum domes can reliably be made the identical size, but not yet in a regular array. We can also make individual instances of very much smaller quantum pyramids, which must be imaged at atomic dimensions.

Assume for the moment that various electronic components can be chemically synthesized. How, then, do we connect them to form a relatively ordered configuration?

The interconnection dilemma

For most of the first 50 years of computing, the emphasis has been on the maximization of component count; the wires have usually been treated as though they were free. Today's chip densities are such that the wires consume some 70% of the real estate—hence they account for some 70% of the defects that lower chip production yield. In the world of nanoelectronics, this trend will be exacerbated to the point where it would be apt to rework one of the lines that Shakespeare accorded Julius Caesar: The fault, dear Brutus, lies not in our gates, but in our wires. This problem suggests that the industry would be best served by a computing paradigm that relies on regular structures rather than global wiring. There are many examples of such computing already in existence: the Connection Machine, systolic arrays, computers based on cellular automata, and various special-purpose supercomputers optimized for such problems as genetic algorithms, lattice gas dynamics, and neural networks.

One issue is how to build compilers that can map the applications onto the regular computing elements of these machines in a way that can also achieve general market acceptance. Another obstacle is the enormous heat flux that will build up as devices approach the molecular scale. So we need to develop regular structures of high density that compute with quantum states or whole electrons and that have low power density.

Other problems arise when one attempts to use such assemblies to do computation. Because chemical syntheses invariably have a statistical yield, not all of the discrete devices will be operational. Furthermore, the system will inevitably suffer from some uncertainty in the connectivity of the devices.

Under such conditions, how can one communicate with the system from the outside world in a reliable and predictable way and be assured that the system is performing error-free computations? Furthermore, because one goal of nanoscale technology is to equip a system with a huge number—a mole, for example—of devices, so as to permit parallelism, how can we impose a form of organization that allows the entire ensemble to operate efficiently? Several fascinating possibilities are under very active investigation.

Defect tolerance

Because economic considerations will be a significant constraint on the future of nanoelectronics, it makes sense to examine issues of circuit architecture at this early stage, before settling on a device type that may turn out to be too expensive to fabricate.

Even if the rate of defects in a chemically fabricated nanocircuit were only one per billion components, which exceeds the current best practice in chip fabs, it would still result in a million defects in a system containing 10^{15} components. How can we build a computer that can tolerate such a level of defects?

The largest defect-tolerant computer built so far is Hewlett-Packard's experimental machine known as Teramac. Although Teramac was constructed by means of conventional technology, many of its problems resemble the challenges that face scientists who are exploring nanoscale paradigms. Teramac was built from a large number of components that had significant defect probabilities. To keep the construction costs reasonable, the builders knowingly used components that were defective and inexpensive. Furthermore, the techniques used to connect all the components together were error-prone. It is truly a junkyard computer.

Teramac is a reconfigurable multiarchitecture computer with 10^6 gates that operate at 1 MHz, or a total of one trillion bit operations per second. Teramac is based on field-programmable gate arrays (FPGAs). They are essentially lookup tables connected by a huge number of wires and switches that are arranged to form crossbars, which allow you to connect any input with any output. In principle, FPGAs substitute memory for logic whenever possible. As the number of resources available in a computer increases, it makes more sense to store as many intermediate results as possible and just look them up when needed.

Less than a third of the gate arrays are used for lookup tables; most are used only for their crossbar switches, to provide the massive interconnectivity that defines the six-level hierarchy of the architecture. Because of the very high degree of connectivity in Teramac, it was possible to access nearly all of the good components in the system while ensuring that none of the bad ones were used.

Perhaps the most amazing fact about Teramac is that it was comatose at birth. Three-quarters of the FPGAs

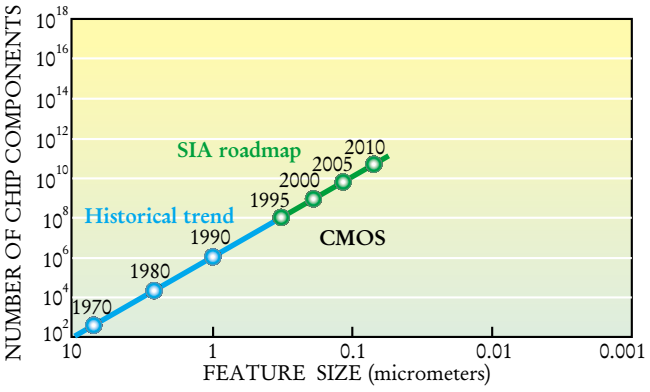

FIGURE 5. SCALING OF ELECTRON DEVICES as a function of the size of the smallest feature on a chip. In 1997 the Semiconductor Industry Association published a roadmap detailing a series of technological milestones that the semiconductor industry must reach to continue the scaling given in Moore's law through the year 2012. However, many of the problems that the milestones represent have no known solution at the present time.

contain defects that would be fatal for an isolated chip. In fact, the manufacturer gave those chips to the creators of Teramac at no cost, charging only for the perfectly functioning ones. Teramac contains a total of 220 000 wiring and gate defects—a total of 3% of all of its resources. For the first 24 hours of its existence, Teramac was connected to a workstation that performed a series of tests to find out where the defective resources were. Those locations were then written to a configuration table as being "in use," to ensure that the defective components would not be accessed by a running program.

A huge configuration command sets all the switches to create a particular architecture that is optimized for a specific computational problem, avoiding all of the defective components in the process. All of that underpinning is invisible to the user, who sees only a perfectly functioning machine that can perform most calculations about 100 times faster than a top-end workstation. The important lessons of Teramac for nanotechnology are that a system does not have to be perfect to be very powerful, and that the more defects a system can tolerate, the cheaper it will be to build.

Thus, perhaps the search for a way to make nanostructured devices possible should concentrate on wires and switches, because those are the components that will allow highly defect tolerant systems to be built. The most desirable types of wires would be those that could conduct information without having to conduct electric current (perhaps the information would be in the form of the phase of a charge density wave). The switches should be a form of nonvolatile memory that requires the expenditure of power only to open or close a circuit, but not to maintain the state of the switch.

An architecture similar to Teramac's can also be the basis for highly efficient reversible computing that relies on chemically self-assembled components. A system can be envisioned in which bits are never created or destroyed but are stored in lookup tables and transported from place to place as needed. For a nanotechnology in which a system contains more than 10^{15} resources, the need for logic may actually be small for most applications.

Quantum logic

So far, we have talked only about machines that substitute quantum switches for semiconductor switches but execute classical algorithms by means of Boolean logic. A further great increase in performance could ensue from developing reversible machines that execute what has come to be known as quantum logic; in principle, very clever algorithms could exploit the inherent parallelism of the superposition of quantum states.

If we could solve knotty problems of decoherence, programming, and input/output (to name a few of the most difficult), quantum logic would make it possible to solve some classes of computationally intractable problems, such as factorization and search, that are important in cryptography and Fourier analysis. As Feynman pointed out, quantum logic machines would be ideal for simulating quantum mechanical systems. For some applications, the reversibility and the inherently parallel nature of quantum logic represent a leap far beyond what ideal nonreversible computing can offer, perhaps by still another nine orders of magnitude or more.

Quantum logic is a fascinating prospect, but it does not seem likely to us that it could become a reality in any widespread practical sense before 2025, and many scientists are less optimistic than that. In any case, barring some currently unimagined breakthrough, it is even more unlikely that an entire system would be built that way.

However, we should not despair, for we have seen that there are tremendous advances possible for computing, even if quantum logic never becomes a reality. A physicist's workstation of the future may well run Windows 17 on a Decium, with lots of RAM, but with a reconfigurable, application-specific, quantum-switch-based supercomputer attached.

Will history repeat itself?

Winston Churchill observed that the further back you can look, the further forward you are likely to see. It is possible that history is about to repeat itself, with the introduction of a new disruptive technology for computation in the 21st century.

Today, we have the silicon field-effect transistor, but we speculate that a quantum-state switch could be better. Many laboratories are now engaged in basic research on fabricating materials into arbitrary shapes and sizes. They are searching for the device concept that will lead to a disruptive new technology.

Breakthroughs will require significant advances in the understanding of fundamental issues and will undoubtedly act as the foundation for new mathematical and scientific disciplines. Those companies that convert the breakthroughs into a new, manufacturable technology will be the survivors of the quantum age of information processing.

It is a noble quest. But like Damon, we computer technologists are being held hostage because of our obligations—to the laws of physics. We can only hope that once again physicists, just like Pythias, will arrive in time to save the day.

This article is based on a speech that one of us—Birnbaum—gave at the American Physical Society's centennial meeting in Atlanta, Georgia, on 22 March 1999. ■