COVER FEATURE

The Future of Nanocomputing

Reviewing the lessons learned in the semiconductor industry over the past few decades can help us better understand the novel technologies that are beginning to emerge from the research community.

George Bourianoff
Intel Corp.

We're all familiar with the silicon superhighway and the ITRS Roadmap[1] that has tracked technological progress in the semiconductor industry since 1992. We're also very aware of the projected end of scaled CMOS around 2016. Hence, the interest in emerging alternative technologies—and, in particular, the so-called nanotechnologies—that promise to extend Moore's law beyond 2016.

I say so-called because the semiconductor industry had already entered the nanotechnology world in 2000 with the introduction of the 130-nm node with a 70-nm gate-length feature size, followed by the 90-nm node featuring a critical dimension of 50 nm in 2002. We in the industry are confident that we can continue this CMOS scaling for another 12 to 15 years. We are equally confident that from the novel alternative architectures and devices being proposed today, new scalable technologies—hopefully not too many, so that we can reap the benefits of incrementalism—will emerge that will take us through multiple processor generations for another 30 or so years.

Reviewing the lessons learned in the semiconductor industry over the past few decades can help us better understand the confusing array of technologies that are beginning to emerge from the research community. This experience suggests some criteria we will need to apply to bring these early research efforts into the realm of high-volume manufacturing.

LOOKING BACK

Much of the microelectronics technology we rely on today rests on fundamental scientific breakthroughs made in the 1930s and 1940s associated with government-funded efforts to develop solid-state radar detectors during World War II. Of these, three stand out:

- the band structure concept,
- the effect of impurities on semiconductor properties, and
- advances in producing high-quality crystal interfaces in silicon and germanium.

Beginning in about 1946, researchers began to utilize this knowledge base, and most of the basic semiconductor devices used over the past half-century were developed within the next 15 years:

- 1948: bipolar transistors;
- 1953: field-effect transistors (FETs);
- 1955: light-emitting diodes (LEDs);
- 1957: tunnel diodes;
- 1959: integrated circuits; and
- 1962: semiconductor lasers.

Published by the IEEE Computer Society
0018-9162/03/$17.00 © 2003 IEEE

From the many material systems then being proposed, silicon became predominant because of the quality of the Si/SiO2 interface and its scaling properties. Combining silicon with the elegance of the FET structure has allowed us to simultaneously make devices smaller, faster, and cheaper—the mantra that has driven the modern semiconductor microelectronics industry.

Looking forward: The red brick wall

The ITRS Roadmap contains many tables with columns defined by dates and node numbers and rows describing the various technical requirements necessary to meet the design goals associated with that node by that date.[1] Green cells in the Roadmap represent known technology solutions being practiced; yellow, known but not yet being reduced to practice; and red, no known technology solutions. Hence, the term *red brick wall*.

Currently, the red brick wall that separates what we know from what we don't know will occur sometime around 2007-2008. That's not to say we won't get to the other side—only that we don't know how yet. But we're confident that we'll be able to continue scaling current CMOS technology along the same three axes that have gotten us to where we are:

- new materials,
- improved processes, and
- new geometries.

The semiconductor industry has already entered the nanotechnology world with the introduction of 90-nm fabrication processes and 50-nm critical dimensions, and researchers have already fabricated 15-nm working transistors.

Since the 1960s, when semiconductor products were basically silicon, silicon dioxide, and aluminum, numerous new materials—copper, titanium, and tungsten, among others—have enabled scaling to continue. Going forward, we will continue to see the introduction of new

- high-κ gate dielectrics and gate materials,
- structures and channel materials such as silicon germanium or strained silicon,
- source-drain materials, and
- dopants.

We will also introduce new geometries. Figure 1 shows a trigate transistor structure that has improved properties over bulk-planar transistors. In this case, the device has a vertical channel that connects the source to the drain. The gate then goes on three sides of the channel, providing very good control of the gate current within the channel flow, very good current-voltage (I-V) characteristics, and very good subthreshold slopes.

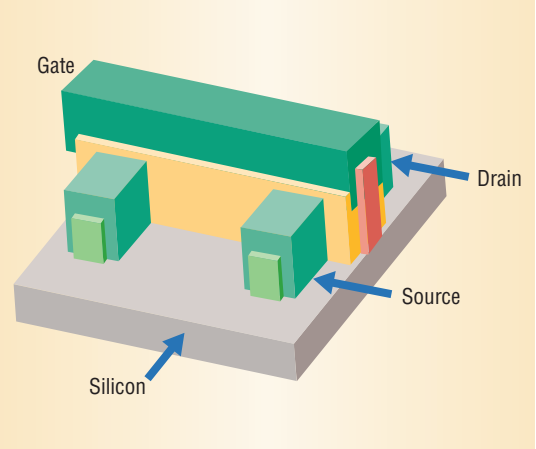

Figure 1. Intel's prototype trigate transistor structure. A vertical channel connects the source to the drain, and the gate is on three sides of the channel. This structure has improved properties over bulk-planar transistors, providing very good control of current flow within the channel, very good current voltage characteristics, and very good subthreshold slopes.

Along with new geometries, there will be improved processing techniques in the way of better etching, anneals, dopant emplacement, and implantation. But better patterning is clearly something that will be necessary to get transistors down into the 15-nm range. A prototype extreme ultraviolet lithography exposure tool currently in use at the Berkeley labs can print 50-nm lines, but because it uses light with a wavelength of 13 nm, it's fairly clear we can get to the 15-nm range.

Technology limits

An analysis based on fundamental quantum mechanical principles reveals that heat dissipation will ultimately limit any logic device using an electronic charge.[2] This analysis of an arbitrary switching device made up of a single electron in a dual quantum well separated by a barrier and operating at room temperature shows that the amount of power pulled off the material surface ultimately limits its device density and operation frequency. That limit is about 100 watts per square centimeter for passive cooling techniques with no active or electrothermal elements. This calculation demonstrates that even though nanotubes and nanodots may allow the fabrication of smaller and faster devices, power dissipation requires spreading them out and slowing them down to keep from vaporizing the chip.

The conclusion is that scaled CMOS devices manipulate electronic charge as well as any research and development nanotechnologies. This fundamental technical limit has increased interest in advanced alternative technologies that rely on something other than electronic charge—like spin or photon fields—to store computational state.

At least partially in response to this observation, the ITRS technical working group for Process Integration, Devices, and Structures formed a committee to study emerging research devices and advise the industry on alternative state variables.[1]

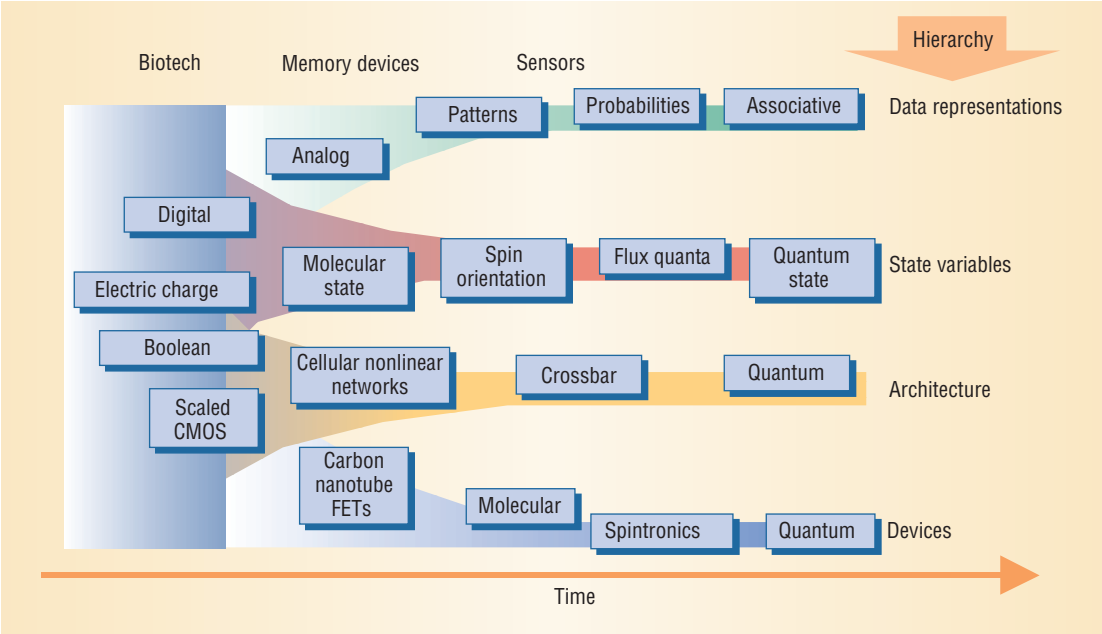

Figure 2. Emerging devices taxonomy. The taxonomy consists of a four-level hierarchy composed of devices, architectures, state variables, and data representations, and it is populated with several prototypical devices.

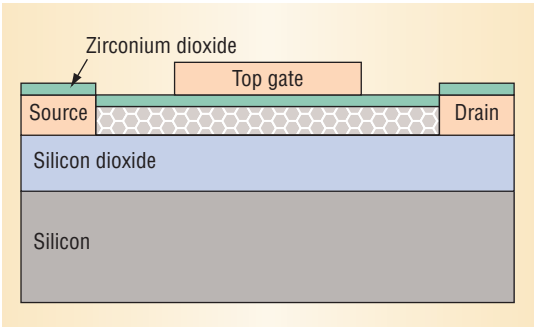

Figure 3. Prototype CNT-FET device. A CNT connecting source, drain, and high-κ zirconium dioxide insulation provide performance characteristics that approach theoretical limits.

NANOCOMPUTING AND BEYOND

A Semiconductor Research Corporation task force working on Emerging Reseach Devices determined that a taxonomy was needed to give some organization to this growing field of research in which the problem domains have not yet been well differentiated. To categorize the rather confusing collection of research efforts shown in Figure 2, it helps to ask, Is it a device or an architecture? Does it introduce a new state variable or data representation?

The working group proposed a hierarchy consisting of four levels: devices, architectures, state variables, and data representations. Using a familiar example, a basic CMOS device such as a FET uses Boolean logic gates (AND/OR gates, XNOR gates, and so forth) in a Boolean architecture, and the state variable is associated with the presence or absence of electric charge differentiating the 1 and 0 state in a binary digital data representation.

Categorizing the devices, architectures, variables, and data representations that populate this space of emerging alternative technologies is even more challenging. Although it is likely to undergo numerous revisions, this taxonomy provides a starting point for developing an organizational structure to frame future discussions.

Devices

Devices are the lowest level in the taxonomy and are the most tangible entities. You can look at a single device through a microscope, put probes on it, and define its operation in terms of inputs and outputs.

Researchers are actively exploring the use of molecular materials such as carbon nanotubes (CNTs) and silicon nanowires in prototype molecular electronics devices such as FETs. Such materials are attractive to researchers because their small molecular structures will conceivably enable scaling (miniaturization) beyond what even the most advanced lithographic techniques permit. While this approach may offer some advantages for some applications—for example, memory—the fundamental limit calculation reveals that potential molecular or CNT devices would have to operate more slowly than scaled CMOS if they are smaller than scaled CMOS.

Carbon nanotubes. To form carbon nanotubes, an atomic planar sheet of carbon atoms is bonded together into an array of hexagons and rolled up to form molecular tubes or cylinders with 1 to 20 nm diameters and ranging from 100 nm to several microns in length. The details of fabrication (the chirality) determine a nanotube's electronic properties—that is, whether it is a semiconductor or a metal.[1] Researchers at several institutions have fabricated CNT-FET device structures with the standard source, drain, and gate, but with a CNT replacing the silicon channel as Figure 3 shows.

Recent studies at Cornell show that it is possible to form a low-leakage gate dielectric by atomic layer deposition of a zirconium dioxide thin-film (8 nm) high-κ gate material on top of an array of CNT-FETs.[3] The zirconium dioxide gate insulation provides high capacitance, allowing efficient charge

injection and reducing current leakage. Performance characteristics for this prototype device show good switching characteristics as measured by the subthreshold slope. The subthreshold slope of 70 mV per decade for the p-type transistors approaches the room-temperature theoretical limit of 60 mV. Subthreshold swing is a key parameter for transistor miniaturization since it measures how well a small swing in gate voltage can cut off current flow. Low cutoff current directly translates into low standby power—a major challenge for end-of-the Roadmap devices.

The availability of a deterministic supply of bulk CNT materials limits CNT devices. All the known processes for producing CNT materials produce CNTs of all types and sizes mixed together. To make devices, the nanotubes must be separated into groups of similar size and chirality. Currently, a laborious manual process using a scanning tunneling microscope is the only way to separate the tubes. The other significant issue with fabricating CNT devices is the problem of contact resistance. Right now, this represents a serious problem relative to more conventional CMOS devices. Even the theoretically "best" value of 6 kOhms is high and will limit the maximum current.

Clearly, such prototype devices show promise, although whether they can actually become competitive remains to be seen.

Silicon nanowires. Recent experimental results, although not conclusive, give some indication that quantum confinement in the transverse direction of silicon nanowires results in greater mobilities than in bulk silicon.[4] Although the effect is not well understood, it seems to be related to the quantum-confined nature of the wire, which limits the density of available phonon states and hence reduces the probability of an electron phonon-scattering event—that is, it reduces drag. Coupling this observation with the desirable characteristics and vast experience base associated with silicon makes using silicon nanowires as a replacement for bulk silicon channels an attractive option.

Silicon nanowire devices have been fabricated in several geometries. Figure 4 shows three nanowire devices with silicon nanowire connecting the source and drain contact points, but with different gate structures: a back gate, a metallic gating structure separated by an oxide, and a coaxial structure. The I-V characteristics of the back gate device have values of 250 to 500 mV/decade, which is understandable because a very thick gate oxide was used in the studies.[4] Experiments are under way to measure the I-V characteristics of coaxial structures with oxide thicknesses of a few nanometers, and these are expected to have I-V characteristics that are much closer to the theoretical limit.

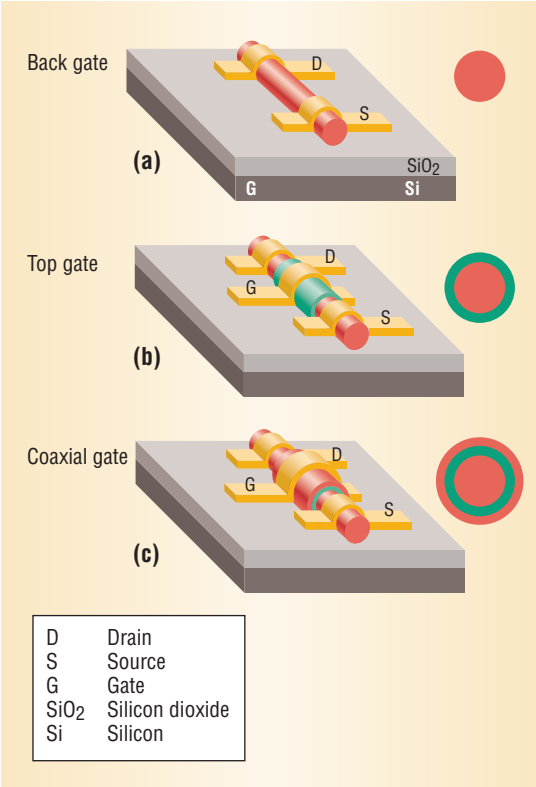

Figure 4. Nanowire devices: (a) back gate, (b) metallic gating structure separated by an oxide, and (c) coaxial structure. Silicon nanowire connects the source and drain contact points in the different gate structures.

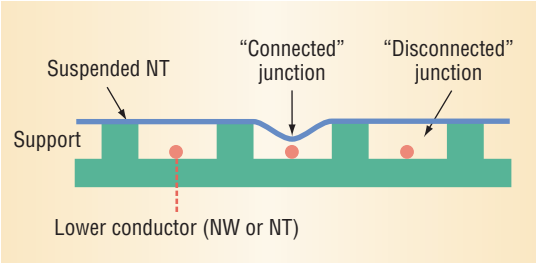

Figure 5. Suspended nanotube switched connection. One array of parallel nanowires is overlaid on a second array oriented at right angles to the first. Depending on the device details, the cross-points of the arrays can either store or switch information.

In addition to offering some advantages compared to bulk silicon in the fabrication of FET structures, silicon nanowires also can be used in other structures that may be better suited to the characteristics of nanowires such as length-to-diameter ratio. Cross-bar arrays are one example of these alternative structures. In such structures, one array of parallel nanowires is overlaid on a second array of nanowires oriented at right angles to the first array. The cross-points of the arrays can be used to either store or switch information depending on the device details.[5,6]

In Figure 5, the crossed nanowires (or nanotubes) act as a switch with bistable positions open or closed. The mechanical equilibrium of the wires maintains the neutral (open) position. Applying opposite charges to the wires pulls them toward each other until they touch, at which time molecular forces hold them in the closed position.

The potential of nanotubes lies in achieving increased density and reducing fabrication costs.

Applying similar charges to the two wires forces them apart, resetting them to the initial position.

The performance of these isolated devices cannot in general compete with scaled silicon on speed. Their potential lies in achieving increased density and reducing fabrication costs. Proponents of cross-bar architectures argue that arrays of these devices can be "self-assembled" using fluidic assembly and Langmuir-Blodgett techniques. The major problems are in providing the gain necessary for signal restoration and fan out and to connect the self-assembled modules to global control lines. This area currently is attracting a great deal of interest, and the next few years will determine if the reality compares with the promise.

Other novel devices. NEC has introduced a room-temperature single-electron transistor (SET) that implements an "island" surrounded by TiOx barriers—a quantum dot—where the presence or absence of a single electron controls the current flow from source to drain via the Coulomb blockade effect. Connected through tunneling barriers, the conductance of the dot exhibits strong oscillations as the voltage of a gate electrode is varied. Each successive conductance maximum corresponds to the discrete addition of a single electron to the dot. This prototype device currently has problems associated with stray charges in the substrates, but there may be ways to handle that. In addition, several concepts of single-electron memory have been experimentally demonstrated,[7] including a SET/FET hybrid.[8] Two major disadvantages of all single-electron memories reported so far are very low operating temperature of 4.2-20 K and background charges.

Another novel device is a molecular switch constructed by Mark Reed's group at Yale.[9] Molecular memory is a broad term combining different proposals for using individual molecules as building blocks of memory cells in which one bit of information can be stored in the space of an atom, molecule, or cell. Placing 100 or so molecules between the source and drain actually changes their molecular state and thus their I-V characteristics. This molecular memory stores data by applying external voltage that causes the transition of the molecule into one of two possible conduction states. The molecular memory reads the data by measuring resistance changes in the molecular cell. This operational characteristic is frequently represented as a hysteresis plot with voltage on the horizontal axis and resistance on the vertical axis.

All of these devices use electrons as their state variables, but it is possible to envision devices that manipulate other state variables. For example, the spin resonance transistor proposed by Eli Yablanovitch at the University of California, Los Angeles, uses the gate voltage to manipulate the spin of electrons bound to two phosphorous nuclei, embedded in an engineered gate stack that has a variable G factor.[10] Readout is accomplished by applying a time-varying magnetic field to the stack and creating an electron spin resonance when exactly the right frequency is applied. This is one example of how to control something other than electronic charge—in this case, nuclear spin.

Architectures

The architecture, a computer's fundamental organizational structure, constitutes the next level up in the taxonomy, and is just slightly more abstract than devices. Although manufacturers vary the microarchitectural implementations, most microprocessors use the Intel x86 instruction set architecture, which remains constant across all implementations.

As alternative devices are introduced, radically new architectures will be required to support them. Table 1 lists the characteristics of six of these emerging research architectures.

3D heterogeneous integration. The integration of semiconductor devices in 3D arrays is being driven from two distinct directions. The first is associated with the need to integrate dissimilar technologies on a common platform to deliver an optimum information processing solution.

It is clear that emerging technologies beyond scaled CMOS offer the potential for greatly improved performance by mixing and matching technologies for particular applications. The combination of technologies requires the 3D integration of functionally dissimilar technologies beginning with microprocessors, ASICs, and DRAMs and extending to RF, analog, optical, and MEMS. These dissimilar technologies may later include 3D integration of molecular, plastic, rapid single-flux quantum superconductors and other emerging technologies directly on to silicon platforms.

The other principal driver for 3D integration is the need to reduce global interconnect delays to maximize system performance. In certain idealized circumstances, 3D superposition of devices will decrease interconnect delays by substantial amounts t relative to an equivalent number of transistors arranged in a planar arrangement.[11,12] In principle, either low-temperature wafer bonding or

Table 1. Emerging research architectures.

Architecture	Implementations	Advantages	Challenges	Maturity
3D integration	CMOS with dissimilar material systems	Less interconnect delay; enables mixed technology solutions	Heat removal; no design tools; difficult test and measurement	Demonstration
Quantum cellular automata	Arrays of quantum dots	High functional density; no interconnects in signal path	Limited fan out; dimensional control (low-temperature operation); sensitive to background charge	Demonstration
Defect-tolerant	Intelligently assembles nanodevices	Supports hardware with defect densities > 50 percent	Requires precomputing testing	Demonstration
Molecular	Molecular switches and memories	Supports memory-based computing	Limited functionality	Concept
Cellular nonlinear networks	Single-electron array architectures	Supports memory-based computing	Subject to background noise; tight tolerances	Demonstration
Quantum computing	Spin resonance transistors, NMR devices, single-flux quantum devices	Exponential performance scaling, but can break current cryptography	Extreme application limitation; extreme technology	Concept

monolithic integration on a common substrate can achieve 3D integration. Wafer bonding has significant limitations because of alignment issues and is currently limited to accuracies of 1 to 5 microns. A host of typical material integration issues pose a challenge to monolithic integration. All 3D integration implementations must deal with issues of heat removal because 3D integrations have lower surface-to-volume ratios than planar circuits. The most promising application of 3D integration appears to be combining memory with microprocessors.

Quantum cellular automata. In the QCA paradigm, a locally interconnected architecture consists of a regular array of cells containing several quantum dots. Electrostatic interactions, not wires, provide the coupling between the cells. When we inject an electron pair into the cell, this electron pair's orientation defines the cell's state. Two bistable states representing 1 and 0 align themselves in one of two directions, with the electrons populating the opposite corner of the cell. Although electronic QCAs are predominant, magnetic QCAs are another recent development for which the performance cannot yet be assessed. Combining these QCAs allows performing circuit functions that are quite different from using Boolean logic gates.

Archetypal QCA straddles the line between being a new architecture and a new device, clearly demonstrating that new architectures may be required to support new devices. If the cells are arranged in a regular square grid, long-established cellular automata theory can be applied, together with its extension, cellular nonlinear (or neural) network (CNN) theory, to describe the information processing algorithm. This allows applying a large body of theory directly to QCA architectures.[1]

Regular EQCA grids can solve certain types of diffusion and wave equations.[13] Also, it is theoretically possible to use them to solve arbitrary digital logic problems, but such systems would be extraordinarily inefficient in terms of area. However, by departing from the regular grid structure, it is possible to design EQCA structures that could carry out universal computing with moderate efficiency.

In addition to a nonuniform layout, EQCAs need a spatially nonuniform "adiabatic clocking field," which controls the switching of the cells from one state to another and allows them to evolve relatively rapidly to a stable end state. Using clocking adds to the complexity of pure EQCA circuits but greatly extends their possible range of applications. It is possible to construct a complete set of Boolean logic gates with EQCA cells and to design arbitrary computing structures. However, current device and circuit analyses indicate that the speed of EQCA circuits will be limited to less than about 10 MHz.[14,15]

Defect-tolerant architecture. The possibility of fabricating nanometer-scale elements that probably will not satisfy the tolerance and reliability requirements typical of larger-scale systems creates the need for defect-tolerant hardware. Systems consisting of molecular-size components are likely to have many imperfections, and a computing system designed on a conventional zero-defect basis would not work. For a conventional integrated circuit, designers describe the chip function, then they construct the hardware.

The general idea behind defect-tolerant architectures is conceptually the opposite: Designers fabricate a generic set of wires and switches, then they configure the resources by setting switches that link them together to obtain the desired functionality.[16] A cornerstone of defect-tolerant systems is redundancy of hardware resources—switches, memory cells, and wires—which implies very high integration density. Fabrication could potentially be very inexpensive if researchers can actualize a chemical self-assembly and attach global interconnects.

Quantum computers are characterized as something between a device and an architecture.

However, such a circuit would require a laborious testing process, implying a significant overhead cost.

It is important to differentiate between defect-tolerant and fault-tolerant architectures. Fault-tolerant systems are designed to deal with transient faults and usually require some form of redundancy checking. Both defect-tolerant and fault-tolerant systems have an upper limit to the number of defects or faults they can handle before the correction process dominates the overall calculation efficiency. The numerical limit appears to be 20 to 25 percent bad elements or defective calculations.

Phase logic. Although phase logic is strictly defined as an architecture, the term is closely associated with the use of phase as a state variable. The concept is generally credited to Richard Feynman, who received a patent for what he called the Parameteron in the 1950s. Feynman postulated that it is possible to store information as the relative phase of two oscillating analog signals in a tank oscillator circuit. One signal is labeled as the reference signal, and the other is labeled as the control signal. Changing the phase of the control signal relative to the reference signal "changes state" and logic. If the circuit has only two allowed relative phases, it is possible to implement both binary and multivalued logic.

Tunneling phase logic is a more recent implementation of the same concept proposed by Richard Kiehl and colleagues at the University of Minnesota.[17] In a TPL circuit, a resonant tunneling diode and an RC circuit produce the reference signal. The more conventional oscillator circuit produces the control signal, but the circuit still stores information as the relative phase between the two signals. Implementing TPLs presents many problems, primarily associated with manufacturing uniform tunneling diodes and stray background charge.

Quantum computing. Quantum computers are characterized as something between a device and an architecture. There are many different ways to instantiate quantum computing, including a solid-state proposal by Bruce Kane and colleagues at the University of Maryland.[18] Other options include liquid state, nuclear magnetic resonance, linear ion traps, or quantum dots. There are numerous implementation methods, but clearly manipulating quantum information will require novel architectures.

Coherent quantum devices rely on the phase information of the quantum wave function to store and manipulate information. A *qubit,* the phase information of any quantum state, is extremely sensitive to its external environment. A qubit is easily connected or entangled with the quantum states of particles in the local environment, and no physical system can ever be completely isolated from its environment. The same sensitivity, however, can be used to entangle adjacent qubits in ways that physical gates can control.

The core idea of quantum information processing or quantum computing is that each individual component of an infinite superposition of wave functions is manipulated in parallel, thereby achieving a massive speedup relative to conventional computers. The challenge is to manipulate the wave functions so that they perform a useful function and then to find a way to read the result of the calculation.

Essentially, three different approaches have been taken to the implementation of quantum computers:

- bulk resonance quantum implementations including NMR,[19] linear optics,[20] and cavity quantum electrodynamics;[21]
- atomic quantum implementations including trapped ions[22] and optical lattices;[23] and
- solid-state quantum implementations including semiconductors[18] and superconductors.[24]

This discussion focuses on solid-state quantum computing because these implementations appear to offer the highest promise for scaling the complexity of quantum computing for commercial applications.

The qubit concept parallels the bit in conventional computation, but offers a much broader set of representations. Rather than a finite dimensional binary representation for information, the qubit is a member of a two-dimensional Hilbert space containing a continuum of quantum states. Thus, quantum computers operate in a much richer space than binary computers.

Researchers have defined many sets of elementary quantum gates based on the qubit concept that perform mappings from the set of input quantum registers to a set of output quantum registers. A single gate can entangle the qubits stored in two adjacent quantum registers, and combinations of gates can be used to perform more complex computations. Just as in Boolean computation, there are minimal sets of quantum gates that are complete with respect to the set of computable functions.

Theoretically, quantum computers are not inferior to standard computers of similar complexity and speed of operation. More interesting is the fact that for some important classes of problems, the

quantum computer is superior to its standard counterpart. In particular, Peter W. Schor demonstrated that a quantum computer can determine the two prime factors of a number in time proportional to a polynomial in the number of digits in the number.[25] This truly remarkable result showed that for this particular class of problems, the quantum computer is at least exponentially better than a standard computer.

The key to this result is the capability of a quantum computer to efficiently compute the quantum Fourier transform. This result has immediate application in cryptography since it would allow the quick determination of keys to codes such as RSA. It is estimated that a few thousand quantum gates would be sufficient to solve a representative RSA code containing on the order of 100 digits. There are several other applications that are variants of the factorization problem.[26]

State variables

In this context, the term *state variables* refers to the notion of the finite state machine introduced by Alan Turing in the 1930s. The idea is that there are numerous ways to store computational information or state to manipulate and store it. The earliest example of a finite-state storage device was the abacus, which represents numerical data by the position of beads on a string. In this example, the state variable is simply a physical position, and the operator accomplishes readout by looking at the abacus. The operator's fingers physically move the beads to perform the data manipulations.

Thermal effects will ultimately limit dimensional scaling of any logic device that manipulates electric charge as a state variable. The corollary of this observation is that the search for alternative logic devices must embrace the concept of using state variables other than electric charge.

Early core memories used the orientation of magnetic dipoles to store state. Similarly, paper tapes and punch cards used the presence or absence of holes to store state. Recent research activities in alternative state variables that seem to be the most prevalent in the research literature include

- molecular state,
- spin orientation,
- electric dipole orientation,
- photon intensity or polarization,
- quantum state,
- phase state, and
- mechanical state.

The development of research devices that use alternative state variables makes it desirable to include state variables as one of the four major categories that constitute the nanocomputing taxonomy.

Data representations

Recent advances in computational power and the anticipated increase in the near future have led to the paradoxical situation that we can generate far more data than we can use or interpret effectively. The goal of current research into alternative data representations is to compactly represent key features of a data set that are useful for a particular application. Image compression using edge-finding and raster-scanning techniques involves scanning an image of binary data generated by some other means and extracting the compressed data. As we encounter alternative devices, architectures, and state variables, adopting alternative data representations in which to manipulate, store, and visualize information may be advantageous.

Adopting alternative data representations in which to manipulate, store, and visualize information may be advantageous.

The best analogy to illustrate this point is the use of Fourier transforms to analyze time series. It is obvious that a Fourier decomposition of a random time sequence offers a compact way to represent a large quantity of data. It is also true that it is possible to work entirely in transform space and, in some cases, to achieve significant numerical efficiencies.

The concept of alternative state variables is a generalization of this concept that can be applied effectively to more general types of problems such as feature recognition, hierarchical data reductions, and multipole expansions. As such, it constitutes the fourth and most abstract category in the proposed nanocomputing taxonomy.

REASONABLENESS CRITERIA

We can't predict which of the devices, architectures, state variables, or data representations will evolve into the next scalable computing technology. However, we can propose some set of reasonableness criteria that will help differentiate between what might actually make its way into high-volume manufacturing and what might not.

Economic criteria

We can confidently predict that scaling will continue as long as we can meet the economic imperative that incremental cost divided by incremental performance gain must be less than it would be for alternative technologies.

Although this economic relevance criterion is very

easy to state, it is very hard to calculate. To say that the risk of adjusted return on investment of any new technology must exceed that of silicon is trivially true, but as Herb Kroemer, recipient of the 2000 Nobel Prize in physics, cautions, sufficiently advanced technologies will create their own applications. To estimate the total ROI for an emerging technology, in addition to markets that exist today, we also must anticipate future markets—and that's tough.

Technology criteria

The other criteria are technical in nature. CMOS compatibility and energy efficiency are the dominant issues when incorporating any alternative technology in high-volume manufacturing. Other key technology concerns include scalability, performance, architectural compatibility, sensitivity to parametric variation, room-temperature operation, and reliability.

If we can apply the huge CMOS manufacturing infrastructure that we already know how to manipulate, we can make up for the deficit in some other areas. I would recommend applying lessons learned from the experience in this area to help guide future research in the university community.

GOING FORWARD

The types of scientific breakthroughs we need to see going forward have strong connections to the work done in the 1940s and 1950s. These future breakthroughs must occur in three areas:

- the bulk band structure of solids needs to be replaced by geometry-dependent energetic structures of nanostructures, requiring us to analyze their stability and their basic quantum mechanical energy levels;
- doping, a bulk process, needs to be replaced by the precise manipulation and placement of individual atoms; and
- crystal growth, another bulk process, needs to be replaced by the self-organization of matter and self-assembly of complex structures.

As we move to nanotechnology, one of the challenges we face is the integration of hard stuff that is very precise and well defined with stuff that's soft, wet, squishy, and subject to fault tolerance and large variations in capabilities.

There are many good reasons to believe that CMOS will continue to scale for another 12 to 15 years. Beyond that, the details are fuzzy, but it is clear that new scalable technologies will begin to emerge and will be integrated on CMOS by about 2015. These technologies represent solutions to meet specific needs and will hopefully point the way to radically new scalable technologies that will take us into the middle of the century.

For either of these things to happen, nanoscience research is needed to enable the new technologies. ■

Acknowledgment

I would like to acknowledge my colleagues at the Semiconductor Research Corporation, Ralph Cavin, Jim Hutchby, and Victor Zhirnov, for much of the material and many of the views expressed in this article.

References

1. International Sematech, *The International Technology Roadmap for Semiconductors, 2001 Edition*, 2001; http://public.itrs.net/Files/2001ITRS/Home.htm.
2. V.V. Zhirnov et al., "Limits to Binary Logic Switch Scaling—A Gedanken Model," to be published in *Proc. IEEE*, Sept. 2003.
3. A. Javey et al., "High-κ Dielectrics for Advanced Carbon-Nanotube Transistors and Logic Gates," *Nature Materials*; http://www.lassp.cornell.edu/lassp_data/mceuen/homepage/nmat7691.pdf.
4. Y. Cul and C.M. Lieber, "Functional Nanoscale Electronic Devices Assembled Using Silicon Nanowire Building Blocks," *Science*, Feb. 2001, pp. 851-853.
5. A. DeHon, "Array-Based Architecture for FET-Based Nanoscale Electronics," *IEEE Trans. Nanotechnology*, Mar. 2003, pp. 23-32.
6. Y. Huang et al, "Logic Gates and Computation from Assembled Nanowire Building Blocks," *Science*, vol. 294, Nov. 2001, pp. 1313-1317.
7. N.J. Stone and H. Ahmed, "Silicon Single Electron Memory Cell," *Applied Physics Letters*, Oct. 1998, pp. 2134-2136.
8. H. Mizuta et al., "Nanoscale Coulomb Blockade Memory and Logic Devices," *Nanotechnology*, vol. 12, 2001, pp. 155-159.
9. M.A. Reed et al., "Molecular Random Access Memory Cell," *Applied Physics Letters*, vol. 78, 2001, pp. 3735-3737.
10. H.W. Jiang and E. Yablonovitch, "Gate-Controlled Electron Spin Resonance in GaAs/$Al_xGa_{1-x}As$ Heterostructures," *Physical Rev. B*, July 2001, vol. 64, no. 4, pp. 041307/1-4.

11. J.A. Davis et al., "Interconnect Limits on Gigascale Integration (GSI) in the 21st Century," *Proc. IEEE*, vol. 89, 2001, pp. 305-324.
12. K. Banerjee et al., "3-D ICs: A Novel Chip Design for Improving Deep-Submicrometer Interconnect Performance and Systems-on-Chip Integration," *Proc. IEEE*, vol. 89, 2001, pp. 602-633.
13. W. Porod et al., "Quantum-Dot Cellular Automata: Computing with Coupled Quantum Dots," *Int'l J. Electronics*, vol. 86, 1999, pp. 549-590.
14. L. Bonci, G. Iannaccone, and M. Macucci, "Performance Assessment of Adiabatic Quantum Cellular Automata," *J. Applied Physics*, vol. 89, 2001, pp. 6435-6443.
15. K. Nikolic, D. Berzon, and M. Forshaw, "Relative Performance of Three Nanoscale Devices—CMOS, RTDs and QCAs—against a Standard Computing Task," *Nanotechnology*, vol. 12, 2001, pp. 38-43.
16. J.R. Heath et al., "A Defect-Tolerant Computer Architecture: Opportunities for Nanotechnology," *Science*, vol. 280, 1998, pp. 1716-1721.
17. T. Yang, R.A. Kiehl, and L.O. Chua, "Tunneling Phase Logic Cellular Nonlinear Networks," *Int'l J. Bifurcation and Chaos*, vol. 11, no. 12, 2001, pp. 2895-2911.
18. B.E. Kane, "A Silicon-Based Nuclear Spin Quantum Computer," *Nature*, vol. 393, 1998, pp. 133-137.
19. M. Steffen, L.M.K. Vandersypen, and I.L. Chuang, "Toward Quantum Computation: A Five-Qubit Quantum Processor," *IEEE Micro*, vol. 21, no. 2, 2001, pp. 24-34.
20. S. Takeuchi, "Experimental Demonstration of a Three-Qubit Quantum Computation Algorithm Using a Single Photon and Linear Optics," *Physical Rev. A*, vol. 62, no. 3, 2000, 032301.
21. P. Grangier, G. Reymond, and N. Schlosser, "Implementations of Quantum Computing Using Cavity Quantum Electrodynamics," *Progress of Physics*, vol. 48, 2000, pp. 859-874.
22. C. Monroe et al., "Demonstration of a Fundamental Quantum Logic Gate," *Physics Rev. Letters*, vol. 75, 1995, pp. 4714-4717.
23. G.J. Milburn, "Quantum Computing Using a Neutral Atom Optical Lattice: An Appraisal," *Progress of Physics*, vol. 48, 2000, pp. 957-964.
24. D.V. Averin, "Quantum Computation and Quantum Coherence in Mesoscopic Josephson Junctions," *J. Low Temperature Physics*, vol. 118, 2000, pp. 781-793.
25. P.W. Shor, "Algorithms for Quantum Computation: Discrete Logarithms and Factoring," *Proc. 35th Ann. Symp. Foundations of Computer Science*, IEEE CS Press, 1994, pp. 124-134.
26. C.P. Williams and S.H. Clearwater, *Explorations in Quantum Computing*, Springer-Verlag, 1998.

George Bourianoff *is a senior program manager in the Strategic Research Group at Intel Corp. His research interests include quantum computing, advanced devices, and optoelectronics. Bourianoff received a PhD in physics from the University of Texas at Austin. Contact him at george.i.bourianoff@intel.com.*