

BEYOND THE SILICON TRANSISTOR: PERSONAL OBSERVATIONS

Will progress in the miniaturization of silicon transistors soon come to an end? Silicon pessimists and the proponents of revolutionary new devices and circuit architectures have been consistently wrong for decades. The author explores past, present, and possible futures of this elusive aspect of computing technology.

Is the decades-long era of exponentially compounding improvements in the cost and performance of computing devices about to end? Can we sustain progress in information technology hardware by developing some new switching device that can be made smaller, faster, and cheaper than the silicon transistor? These were some of the questions raised by a panel session, “Looking Beyond Moore’s Law, A Technical Perspective,” which I had the pleasure to host during a recent CIA/DARPA conference on the Global Computer Industry Beyond Moore’s Law.

The conference organizers asked each panelist to discuss a particular view of the future of computing. (See the “Session Panelists” sidebar for a list of panelists and their topic areas.) Following the panel’s theme, the participants summarized the current ideas and prospects for further progress in their respective areas, including molecular devices and quantum computing.

Implicit in the choice of panel session titles,

speakers, and topics was the idea that progress in the miniaturization of silicon transistors must soon end. However, this view has been prevalent and consistently wrong during my entire technical career. The silicon transistor has remained the forerunner in mainstream computing. Yet, the future of computing is not set in silicon or any other technology. The panelists’ discussion sparked some of my own observations on where I feel the field is going.

From the Start

When I joined IBM Research in the late 1970s, IBM’s Josephson computer project was in full swing, driven by the idea that silicon would soon reach its limits. The envisioned computer’s central processing unit would be the size of a grapefruit, and this “compactness,” along with the speed and power efficiency of the superconducting Josephson junction switches, would let the central processor run at the then-amazing clock speed of 1 GHz. By the mid 1980s, the Josephson project was dead. Many credited its demise to the lack of a compact Josephson memory device to go with the logic. (This is still a problem for the more modern Rapid Single Flux Quantum [RSFQ] version of Josephson logic.) However, what really killed Josephson logic was the

Session Panelists

Each participant at the panel session “Looking Beyond Moore’s Law, A Technical Perspective” discussed the future of computing from a unique perspective:

- Joel Birnbaum, former chief technical officer of Hewlett-Packard, discussed research at HP on molecular devices and gave an upbeat assessment of the prospects for an alternative computing architecture that might exploit those devices.
- Kostya Likharev, physics professor at the State University of New York at Stony Brook, discussed the status of Rapid Single Flux Quantum (RSFQ) logic based on superconducting Josephson junction devices—extremely promising in terms of switching speed and low-power dissipation but operable only at cryogenic temperatures.
- Keith Miller of the National Security Agency discussed quantum computing, emphasizing the primitive state of current research.
- Richard Lipton, Storey Chair of Computer Science at the Georgia Institute of Technology, gave a measured view of progress and problems associated with DNA computing.
- Tom Knight, senior research scientist at the MIT Artificial Intelligence Laboratory and MIT Department of Electrical Engineering and Computer Science, discussed the need and prospects for alternative architectures for computing.
- Steven Wallach, chair of the High-End Subcommittee of Presidential Advisory Board on High Performance Computing, Communications, and Networking (PITAC), discussed the potential for extending computing capabilities, particularly high-end computing, through architectural advances.

relentless progress of silicon technology.

Alternative device research at IBM shifted to compound semiconductor transistors, especially the new heterojunction field-effect transistors (FETs). By the late ’80s, a small team led by George Sai-Halasz¹ had demonstrated experimental silicon transistors with 100-nm gate lengths. Sai-Halasz felt that his team had pushed silicon about as far as it could go, and, for a while, there was some agreement in the technical community that it had reached some important limits. However, David Frank² published an analysis that definitively showed that much shorter channel lengths were possible, pointing the way to today’s experimental silicon FETs approaching 10 nm in channel length.

Silicon pessimists and the proponents of revolutionary new devices and circuit architectures have been consistently wrong for decades. The economic forces driving incremental improve-

ments in the established technology are strong, harnessing the competitive energy and creativity of thousands of engineers and scientists around the world.

This does not mean that a successor to the silicon transistor will not arise. Indeed, if we take a long historical view, we find that several distinct device technologies have carried computation forward over the years, starting with Herman Hol-lerith’s mechanical tabulators used for the 1890 US census. These were eventually supplanted by tabulators and calculators based on electromechanical relays. ENIAC, the first stored-program computer, was built with vacuum tubes. In little more than a decade, vacuum tubes were replaced by discrete bipolar silicon transistors, and these were in turn replaced by silicon integrated circuits based on either bipolar or FETs.

There is no reason that today’s dominant complementary metal-oxide-semiconductor FETs should be the end of the road. The smallest useable CMOS logic switch that we can currently envision would still contain, if we include the contacts, a few million atoms, and yet we know of no fundamental physical principle that prevents us from inventing and building a useful logical switch containing far fewer atoms.

Steps Forward

What are the properties of a useful logical switch? IBM’s Robert Keyes³ has often noted that all devices that innovators have successfully used to build complex logic circuits—the electro-mechanical relay, the vacuum tube, and the transistor—share important properties. Most importantly, each is a three-terminal device that exhibits amplification. With amplification, one device’s output can reliably drive the inputs of similar devices or charge arbitrarily long signal wires. Signal levels can be repeatedly restored to a reference level (the ground or the power supply voltage), and circuits can therefore be designed for immunity to noise and to the inevitable variability in the operating characteristics of individual devices.

To my knowledge, there is no rigorous proof that three-terminal devices are necessary to build reliable, complex logic circuits, and researchers have proposed logic based on two-terminal diode-like devices from time to time. Perhaps the most ambitious proposals to date are those of Seth Goldstein and Mihai Budiu,⁴ positing a system architecture where much of the computation is done within regions of *nanofabric*. The

nanofabric is to be built of two-terminal bistable switches wired in a dense cross-point interconnection architecture—the scheme being pursued in collaboration between Hewlett-Packard and the University of California, Los Angeles. Each region of nanofabric must be surrounded by CMOS (or some other functionally equivalent three-terminal) logic, so Goldstein and Budiú's argument is not that CMOS will be eliminated but rather that we will reduce the net system cost by replacing much of the CMOS logic with nanofabric.

In contrast, at IBM we are looking for three-terminal amplifying devices that might be made smaller and fabricated more cheaply than the ultimate silicon transistor. Our current focus is on the carbon nanotube transistor, which can certainly be made smaller than the smallest conceivable silicon FET, and which shows some potential to compete on grounds of performance. Hopefully, manufacturing costs will eventually be lower for such devices, since expensive lithographic patterning processes might be reduced in favor of cheaper processes relying on chemical self-assembly.

Some have taken the view that processes involving chemical self-assembly must inevitably result in the fabrication of many defective devices. This would be true only if we restrict ourselves to simple chemical processes such as those that industry commonly uses today. A biochemical process like RNA synthesis makes fewer than one error per 10,000 nucleotides. DNA replication produces fewer than one error per billion nucleotides. Memory and logic designers would be comfortable working with such manufacturing error rates. Nature achieves low error rates in self-assembly by working with a limited set of well-differentiated building blocks and by extensive use of error-correction schemes—enzymatically promoted chemical reaction paths that preferentially remove the unavoidable defects produced at each stage of the process. Our industrial processes will eventually be as subtle.

A three-terminal molecular-scale device such as the carbon nanotube transistor is attractive because it is, to some degree, a drop-in substitute for the silicon FET. Although materials and manufacturing processes would have to change, circuit designs and system architectures would be unaffected. At IBM, many of us believe such a device has a reasonable chance of displacing CMOS logic for some functions and establishing some profitable niche from which it can

grow. Substituting devices that do not exhibit strong amplification will require not just changes in processes and materials but also massive changes in circuit and system architectures. Revolutionizing a large, complex, stratified industry is not easy, and as Mark Horowitz⁵ has pointed out, it is much easier to introduce a change if it affects only one or two layers in the value chain.

Looking Beyond

This contest between still rapidly advancing silicon CMOS technology and potential molecular-scale competitors is only part of the story. In the last few decades, our concept of computing has greatly expanded—hence the inclusion of DNA computing and quantum computing in our panel session. In fact, it is now widely recognized that any dynamical system can be thought of as executing a computation.

But which physical systems are suitable for reliable (very low error rate) computation? To date, all commercial digital computers have been built from logic switches with binary energy states separated by many times the thermal energy (kT), with each switching operation dissipating energy greatly exceeding kT. We now know that reliable computation may be done in systems that operate with much less energy dissipation.

Charles Bennett⁶ showed that reliable computation can be done in the Brownian limit. Logical state transitions in a Brownian computer are driven only by the random thermal agitation of its functional parts, so in the limit of zero thermodynamic driving force, such a computer is as likely to run backward as forward. Bennett pointed out that the molecular apparatus of DNA replication, transcription, and translation appears to be nature's closest approach to a Brownian computer. He envisioned a molecular Turing machine (a model of a general-purpose computer) based on similar chemistry.

Leonard Adelman⁷ showed that we can use

Substituting devices that do not exhibit strong amplification will require not just changes in processes and materials but also massive changes in circuit and system architectures.

DNA chemistry to implement Boolean logical functions, thus introducing the concept of DNA computing. Logical operations in such a DNA computer are implemented as a series of thermally activated chemical reactions in liquid solution. In other words, these logical operations proceed in the Brownian limit. Many proposed implementations of DNA computers involve

some macroscopic apparatus, immune to Brownian motion, to gate the series of reactions, but a recent experiment by Chengde Mao and his colleagues⁸ demonstrates a true Brownian computation. This work is a step toward the general-purpose Brownian molecular computer Bennett envisioned.

Edward Fredkin and Tommaso Toffoli⁹ went beyond Bennett and introduced classical physical models for energy-conserving or ballistic computing. However, these models were less realistic than Bennett's model for Brownian computation. Sensitivity of the classical particle trajectories to initial conditions and to thermal fluctuations during operation would quickly drive the system into an error state.

Any practical implementation would require frequent error correction, thus introducing some energy dissipation.

Such a computer would therefore operate in the extreme limit of the Brownian computer, where instead of being subject to constant thermal perturbations, it would only be perturbed now and then. Although a practical classical computer operating in this limit appears unlikely, this is precisely the limit in which we envision operating quantum computers. Choosing a physical system with favorable quantization of energy states and careful isolation of the system from external perturbations might make a quantum computer viable. Schemes for quantum error correction provide a last essential ingredient in the theoretical framework for a viable quantum computer. Excellent reviews of the recent research are available from Charles Bennett and David DiVincenzo,¹⁰ and Michael Nielsen and Isaac Chuang.¹¹

Will DNA computers and quantum computers compete directly with silicon? I doubt it.

I see a future where these and other emerging information-processing technologies will complement the role of silicon, further expanding the scope of computation in society and the economy. I cannot predict future commercial applications of these technologies or their ultimate importance to society any more than Charles Babbage could predict the eventual applications and impact on society of the stored-program classical digital computer that he had clearly envisioned by 1833. Indeed, we are still discovering the applications and trying to understand the implications of Babbage's invention—just consider the current societal debates regarding the impact of information technology on copyright law and privacy. Furthermore, we still have not determined the best physical system in which to implement Babbage's invention. It should not surprise us, then, that our first attempts at physical implementations of DNA computing and quantum computing are halting, and the ultimate utility of our efforts is unclear.

I suspect the first viable commercial applications will not much resemble our present conceptions of computing but will leverage the peculiar strengths of these emerging information technologies.

For instance, a DNA computer might be used to perform rapid combinatorial chemistry in the service of drug discovery, an application that in turn strongly leverages the enormous capacity for parallelism and the consequent ability to solve complex optimization problems. Early applications of quantum information processing may involve computational tasks that require only a few logical operations. For instance, the feedback control of quantum coherent systems might have important applications in metrology and manufacturing process control.

So I foresee a complex future for computing. The silicon transistor will certainly allow further rapid improvements in price and performance for at least another decade, but it might yet be succeeded by some molecular-scale device that exhibits amplification and is therefore essentially a drop-in in terms of circuits and system architectures. At the same time, the meaning of *information processing* will continue to broaden as the number of distinct physical systems used to process information continues to grow, and information technology, in an

I suspect the first viable commercial applications will not much resemble our present conceptions of computing but will leverage the peculiar strengths of these emerging information technologies.

ever-broader sense, continues to play an ever-larger role in our society.

References

1. G.A. Sai-Halasz et al., "Design and Experimental Technology for 0.1 Micron Gate-Length Low-Temperature Operation FET's," *IEEE Electron Device Letters*, vol. 8, no. 9, 1987, p. 463.
2. D. Frank, S.E. Laux, and M.V. Fischetti, "Monte Carlo Simulation of a 30 nm Dual-Gate MOSFET: How Short Can Si Go?" *IEDM Tech. Digest*, Dec. 1992, p. 553-556.
3. R.W. Keyes, "Physics of Digital Devices," *Rev. Modern Physics*, vol. 61, no. 2, 1989, pp. 279-287.
4. S. Goldstein and M. Badiu, "NanoFabrics: Spatial Computing Using Molecular Electronics," *Proc. 28th Ann. Int'l Symp. Computer Architecture*, ACM Press, 2001, pp. 178-191.
5. M. Horowitz, private communication (notes taken from *Focus Center Research Program [MARCO] MSD-C2S2 Topical Workshop 2001*).
6. C.H. Bennett, "The Thermodynamics of Computation: A Review," *Int'l J. Theoretical Physics*, vol. 21, no. 12, 1982, pp. 905-940.
7. L. Adelman, "Molecular Computation of Solutions to Combinatorial Problems," *Science*, vol. 266, Nov. 1994, pp. 1021-1024.
8. C. Mao et al., "Logical Computation Using Algorithmic Self-Assembly of DNA Triple-Crossover Molecules," *Nature*, vol. 407, Sept. 2000, pp. 493-496.
9. E. Fredkin and T. Toffoli, "Conservative Logic," *Int'l J. Theoretical Physics*, vol. 21, nos. 3-4, 1982, p. 219-253.
10. C.H. Bennett and D. DiVincenzo, "Quantum Computation: Toward an Engineering Era?" *Nature*, vol. 377, Oct. 1995, pp. 389-390.
11. M.A. Nielsen and I.L. Chuang, *Quantum Computation and Quantum Information*, Chapter 10, Cambridge Univ. Press, 2000.

Thomas N. Theis is the Director of Physical Sciences at the IBM T.J. Watson Research Center. He has a BS in physics from Rensselaer Polytechnic Institute and an MS and PhD in physics from Brown University. He is a member of the IEEE and a fellow of the American Physical Society. He serves on advisory boards for the American Institute of Physics Corporate Associates and the National Nanofabrication Users Network. He is also a member of the National Research Council's Board on Physics and Astronomy. Contact him at the IBM T.J. Watson Research Ctr., PO Box 218, Yorktown Heights, NY 10520; ttheis@us.ibm.com.

A LEADING JOURNAL ON TECHNOLOGY IN **MEDICINE, BIOLOGY,** AND **HEALTH CARE** NOW HAS A NEW SPONSOR!

The IEEE Computer Society has joined the IEEE Engineering in Medicine and Biology Society in delivering *IEEE Transactions on Information Technology in Biomedicine*.

Peer-reviewed articles feature topics such as

- **biomedical engineering**
- **virtual reality applications for surgery**
- **visualization and biomedical imaging**
- **ethical issues in biomedical applications**
- **information infrastructures in health and medicine**
- **high performance biomedical computing**
- **biotechnology**
- **broadband technologies in medicine**

IEEE Computer Society members can subscribe for the low member rate of \$25!

Read more about *IEEE Transactions on Information Technology in Biomedicine*
<http://www.ieee.org/organizations/pubs/transactions/titb.htm>

