# The Search for Computing's Secretariat[1]

Erik P. DeBenedictis, Sandia National Laboratories

*The slowing of Moore's law offers IEEE and its members the unique opportunity to influence research toward continued growth in computing performance.*

The doubling of integrated circuit (IC) density every year or two in the past half-century fueled exponential computing performance growth and significantly expanded the global IT economy. Computer architects and programmers were kept busy adapting to more capable chips, with no reason to question the materials and devices on the chips or whether scaling occurred in 2D or 3D. This economic tailwind extended to organizations like IEEE, which published, sponsored conferences related to, and produced standards for semiconductors as well as for the industries enabled by semiconductors.

The slowing of Moore's law offers IEEE and its members an opportunity to influence the technical direction of this important economic sector. Roadmaps such as the International Technology Roadmap for Semiconductors (ITRS; www.itrs2.net) predict that increasing chip density will yield fewer benefits over time. The development of new, exciting products is still possible, but only if greater chip density can be coordinated with other changes, such as a switch to new materials, new computing architectures, and new forms of controlling computers such as machine learning. The ITRS has now been expanded to now include other devices, architecture, and software in addition to semiconductors. Renamed the International Roadmap for Devices and Systems (IRDS), it's become an IEEE "industry connection," offering IEEE members and societies a forum to vet and even influence new research priorities across the industry.

## Computer roadmaps in context

Figure 1 presents Moore's law in the context of overall progress in the computing industry. It also highlights where an IEEE-sponsored roadmap could fit. The area outlined in blue shows computing performance pulling ahead of that of the human brain around 1940 and beginning the exponential climb that continues today [1].
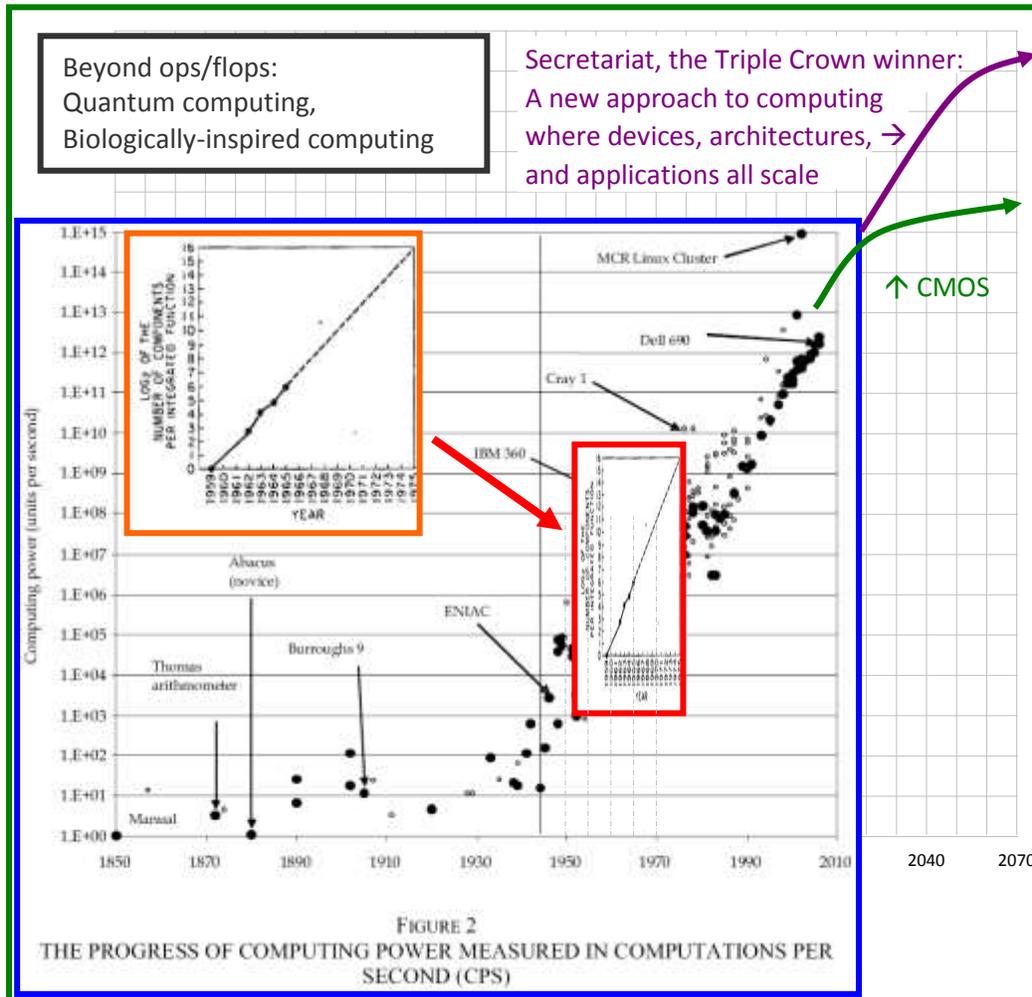
Figure 1: Computing roadmaps: computing performance (outlined in blue),
Moore's law (the original graph is outlined in orange; the same graph scaled to
match years and logarithmic vertical is outlined in red), and future expectations
(purple arrow). The green arrow illustrates CMOS technology reaching full maturity
in a decade or two.

In 1965, Gordon Moore predicted the exponential growth of a handful of parameters related to
IC performance. In Figure 1, the graph from his paper is outlined in orange [2]. This same graph
is also shown outlined in red with the years and the logarithmic vertical scaled to match the
surrounding graph.

The green curve illustrates ubiquitous CMOS technology reaching full maturity in a decade or
two. However, the search is on for a new computing approach that will extend this exponential
growth (purple arrow)—an approach in which the technology and architecture will scale and the
software will serve users' evolving demands. We call the goal of this high-risk, high-payoff
search "Secretariat," named for the racehorse that won the 1973 Triple Crown after a 25-year
interval. Computing's Secretariat might extend the historic growth rates of general-purpose
computing by a few decades but is unlikely to result in computing reaching its ultimate limits.
Future computing generations are still possible, such as quantum computing, biologically
inspired information processing, and other alternative approaches.

## *ITRS Roadmap and Initiatives*

Figure 1 portrays computing's progress as effortless—but there's more to the story. Robert Dennard and his colleagues formalized the insights behind Moore's law in 1974 as mathematical formulae that related changes in density, supply voltage, dielectric (insulator) thickness, energy per signal, and other parameters [3]. As long as the semiconductor industry could produce chips in which these parameters scaled according to the formulae, semiconductor chips would become exponentially denser and more complex and somewhat faster without increased chip cost or energy consumption. But these exponential improvements obviously couldn't continue indefinitely. So the challenge became to identify which parameter would stop scaling first—and what could be done about it?

In 1992, the semiconductor industry began developing a Moore's law roadmap to identify obstacles far enough in advance to fix them before they interfered with business [4]. This effort became the National Technology Roadmap for Semiconductors (NTRS) two years later. The 1997 NTRS contained the following focus areas [5]:
- design and test;
- process integration, devices, and structures (transistors);
- front-end processes (fabrication);
- lithography;
- interconnect;
- factory integration; and
- assembly and packaging.

The NTRS was renamed the International Technology Roadmap for Semiconductors (ITRS) in 1999 as a result of broadening financial support and volunteers

## *Industry and Government Initiatives*

The roadmapping process identified gate oxide thickness as the first limiting parameter for scaling. In simple terms, Dennard's formulae specified a thinning rate for the insulating layer between each transistor's gate and channel [3]. The 1997 NTRS projected the insulating layer would be just a few atoms thick by 2012 (although the crisis actually arrived sooner) [5]. The layer would become so thin that a quantum mechanical tunneling current would adversely affect performance. This projection led to industry-funded R&D to find new materials and a new gate-stack design, ultimately resulting in "high-K dielectrics." Essentially, the insulating layer changed from glass ($SiO_2$) to a material with a higher dielectric constant ($TaO_x$). This allowed a thicker insulating layer, which reduced the tunneling current without affecting the capacitance so important to performance. Chips with the new materials came on the market with multicore microprocessors around 2003.

With the gate oxide-thickness problem and some other issues overcome, roadmaps in the early 2000s predicted the energy-efficiency limitations we're concerned about today. These limitations are due to the maximum amplification possible in transistors—the subthreshold slope—being limited to ln(10) $kT/q \approx 60$ mV/decade for metal-oxide- semiconductor field-effect transistors (MOSFETs). Supply voltage must be reduced to keep chips from overheating as the number of logic gates scales up. With a maxed-out amplification factor and a lowering power-supply

voltage, MOSFETs can't turn off completely when not in use. This results in a "leakage current" and imposes a minimum energy dissipation per gate.

To address the predicted energy-efficiency limitation, the US government authorized the National Nanotechnology Initiative (www.nano.gov) in 2000. In 2004, it formally launched the Nanoelectronics Research Initiative (NRI), which funded research on various device-level alternatives to the MOSFET that CMOS is based on. One such device was a MOSFET-like tunnel field-effect transistor that operates on principles not subject to the $kT/q$ subthreshold slope.

Other devices had no functional relationship to transistors at all, such as interacting magnets. Figure 2 illustrates this research's result [6], summarizing experimental and theoretical energy and speed estimates for two-input NAND gates built from various devices.
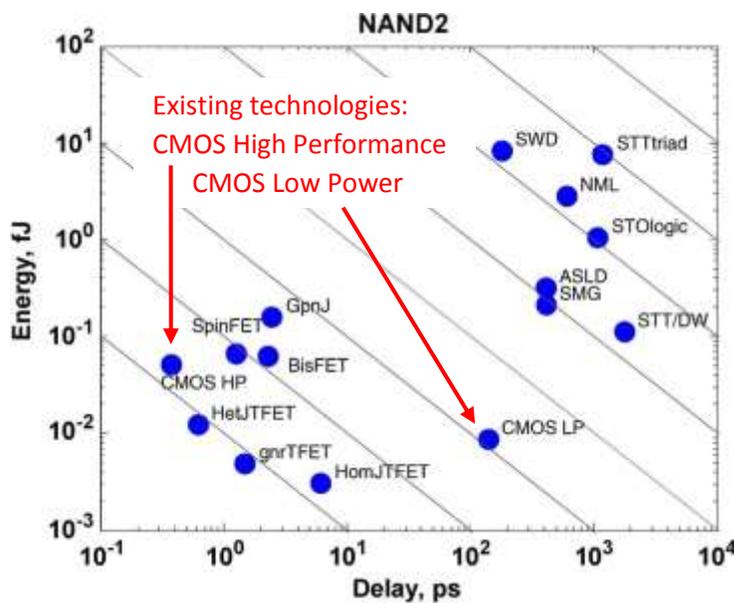


Figure 2: T Transistor alternatives. (Source: D.E. Nikonov and I.A. Young, "Overview of Beyond-CMOS Devices and a Uniform Methodology for Their Benchmarking," Proc. IEEE, vol. 101, no. 12, 2013, pp. 2498–2533.)

What Figure 2 doesn't include is a device that can reinvigorate scaling. Global IT growth depends on declining costs to produce and operate computers. For Figure 2 to present the path forward, it would need at least one device closer to the desirable lower-left corner than the entries for CMOS at high performance and low power. Instead, we see a Pareto frontier: the most viable devices lie above a 45° downward sloping line that avoids the lower-left corner. Some transistor alternatives have lower power but run slowly, while others have higher power but run faster; however, no device has been discovered that offers both low power and high speed, despite NRI's comprehensive search.

### *From NAND Gates to the Triple Crown*

Figure 2's survey of NAND gates doesn't present a path to long-term scaling, perhaps because the solution isn't destined to be based on NAND gates. The information revolution couldn't get started until relays and vacuum tubes were sufficiently mature for von Neumann's computer architecture to implement Turing's theory of computation. The mechanical computers before that time could perform basic functions such as sorting and arithmetic, but mechanical technology was too crude for the more challenging task of general-purpose computing [7]. Figure 2 shows that the industry is in a similar situation today, with the pre-eminent devices unable to make significant headway when merely used as switches in Boolean logic. Computer architecture and software areas such as machine learning present tantalizing ideas, but they won't offer compelling advantages if implemented using any of the NAND gates in Figure 2. The open door is to use a device from Figure 2 as the building block of a new architecture that will support a new generation of applications. This presents three "Triple Crown" challenges at once: new devices, new architectures, and new programming methods. These are all technical areas in which IEEE collectively as a professional organization and IEEE members individually as knowledgeable engineers and scientists can contribute significant leadership.

As a mark of how important the inputs of our community are judged, the US government has asked for our input through a nanotechnology grand challenge that's part of the National Strategic Computing Initiative (NSCI) [8]: "[W]e are therefore challenging the nanotechnology and computer science communities to look beyond the decades-old approach to computing based on the von Neumann architecture as implemented with transistor-based processors, and chart a new path that will continue the rapid pace of innovation beyond the next decade."

The NSCI announcement is similar to the historical ITRS initiatives discussed earlier—if the technology level expands to embrace all three areas instead of just semiconductors. Responding to changing circumstances, the IRDS added the following to the previously listed ITRS roadmapping agenda items in its May 2015 kickoff meeting:

- roadmapping of crossbar computational elements, which is a higher-level computational structure frequently proposed for artificial neural networks;
- workshops on approximate/probabilistic/stochastic computing, superconducting electronics, and quantum computing;
- creation of a computer architecture focus group; and
- creation of a benchmarking focus group for new approaches to computing.

IEEE and the IT industries will continue to develop and use computing roadmaps that facilitate user and supplier planning, identifying and resolving crises before they interfere with business. For the roadmaps to be valid given the changing technological landscape, architecture and applications are being added to the traditional semiconductor focus. This is the charter of IEEE's IRDS industry connection.

Today's roadmaps don't show a long-term scaling path. But with so many of IEEE's members involved with the computer information revolution in one way or another, the organization has the means and motivation to search the combined physics, architecture, and software areas for a new computing approach. It will be a high-risk, high-payoff search not unlike breeding a horse

that will win the Triple Crown. All of us—from individual IEEE members and our businesses and research labs to our professional organizations and the global economy—have a stake in finding computing's Secretariat"

## References

[1]. W.D. Nordhaus, "Two Centuries of Productivity Growth in Computing," *J. Economic History*, vol. 67, no. 1, 2007, p. 128.

[2]. G.E. Moore, "Cramming More Components onto Integrated Circuits, Reprinted from Electronics, Volume 38, Number 8, April 19, 1965, pp. 114 ff," IEEE J. Solid-State Circuits Newsletter, vol. 11, no. 5, 2006, pp. 33–35.

[3]. R.H. Dennard et al. "Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions," IEEE J. Solid-State Circuits, vol. 9, no. 5, 1974, pp. 256–268.

[4]. W.J. Spencer and T.E. Seidel, "National Technology Roadmaps: The US Semiconductor Experience," *Proc. 4th Int'l Conf. Solid-State and Integrated Circuit Technology* (ICSICT 95), 1995, pp. 211–220.

[5]. "The National Technology Roadmap for Semiconductors, Technology Needs, 1997 Edition," Semiconductor Industry Association, 1997; www.ohio.edu/people/starzykj/network/class/ee515/roadmap5.pdf.

[6]. D.E. Nikonov and I.A. Young, "Overview of Beyond-CMOS Devices and a Uniform Methodology for Their Benchmarking," *Proc. IEEE* , vol. 101, no. 12, 2013, pp. 2498–2533.

[7]. W. Isaacson, *The Innovators: How a Group of Hackers, Geniuses, and Geeks Created the Digital Revolution*, Simon and Schuster, 2014.

[8]. L. Whitman, R. Bryant, and T. Kalil, "A Nanotechnology-Inspired Grand Challenge for Future Computing," blog, 20 Oct. 2015; www.whitehouse.gov/blog/2015/10/15/nanotechnology-inspired-grand-challenge-future-computing.