Rebooting Computing: New Strategies for Technology Scaling

Thomas M. Conte, Georgia Tech
Elie Track, nVizix
Erik DeBenedictis, Sandia National Laboratories

Year-over-year exponential computer performance scaling has ended. Complicating this is the coming disruption of the "technology escalator" underlying the industry: Moore's law. To fundamentally rethink and restart the performance-scaling trend requires bold new ways to compute and a commitment from all stakeholders.

0018-9162/15/$31.00 © 2015 IEEE

Society's increasing reliance on and demand for computing power over the past half-century has been ably met by exponential performance increases, yet the heady march toward smaller, faster, cheaper, and more energy-efficient computing technologies is slowing as Moore's law is being disrupted. In fact, some argue that exponential performance scaling ended a decade ago. But before society confronts this sober new era of limitations in which performance supply no longer meets demand, we must rethink—even reboot—computing technology to revive historic exponential performance growth.

This issue of *Computer* explores some of the most promising new technologies, architectures, and engineering strategies to fuel continued computing improvements and what needs to be done to make that happen.

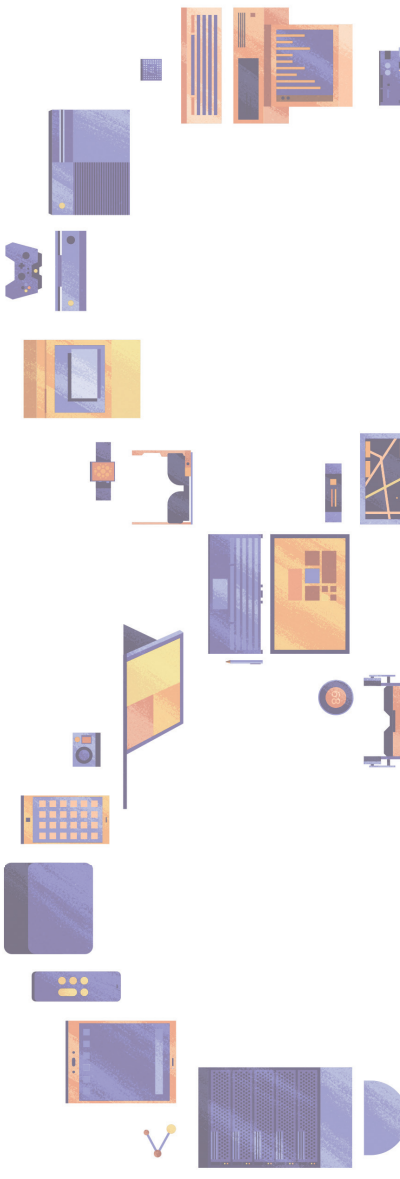
Moore's law has had a really good and relatively long run as far as technology trends go. Reducing dimensions on the surface of a 2D chip enabled more capable devices, with higher energy efficiency and higher clock rates. It characterized a golden era in which software could be developed independent from architecture, architecture independent from implementation, and implementation independent from devices. Each computing domain—software engineering, computer design, semiconductor electronics, and so on—thrived with only an occasional need for interdisciplinary collaboration.

In 2005, physics intervened. The power densities of a CMOS-based microprocessor made continued single-thread performance scaling uneconomical. Multicore processing was only a partial solution: parallel execution accelerates some problems, and using parallelism requires rewriting software instead of simply recompiling. Meanwhile, if Moore's law is to resume, it must be based on an architecture other than multicore.

What new technological escalators are within reach, and how might they impact computing performance? These were questions that inspired the launch of IEEE's Rebooting Computing Initiative (http://rebootingcomputing.ieee.org) in 2012. Thus far, the initiative's work supports the notion that we are nowhere near the end of computing performance growth; in fact, we are entering an era of even more powerful escalators. To get there, though, there must be a real commitment from both government and industry to maximize the remaining benefits of Moore's law; furthermore, research supporting a technological paradigm shift is necessary to reboot computing progress.

IN THIS ISSUE

As John M. Shalf and Robert Leland argue in the first article, "Computing beyond Moore's Law," ideas to reboot computing will require substantial funding and resources to reach the marketplace. The authors were influential in developing the National Strategic Computing Initiative (NSCI), launched by the Obama administration in late July 2015 and tasked with developing technology that could continue scaling for the next decade. Shalf and Leland describe the need for a new



and sustained R&D agenda to evaluate emerging semiconductor materials and device physics.

How will we move from powerful numerical computing to the next level—one at which we can not only analyze vast datasets but learn from them as well? The next five articles reflect our reasoned judgment of the "top 5" new computing approaches that have the greatest likelihood of driving performance improvement. There is an important caveat: each article is written by just one group of researchers, and many other qualified groups in the various fields have different and very valuable perspectives. It is this community of insightful researchers—who both challenge and cooperate with one another—that collectively empowers our faith in continued computing progress.

Three of the articles advocate the reorganization of computing at the lower technology layers. They outline research strategies aimed at improving systems that look similar to those in use today—utilizing existing programming languages, software, and manufacturing methods. The other two articles present alternative approaches that are far more disruptive and reflect a fundamental change in the current computing model. These methods offer greater potential to achieve performance gains to justify less compatibility with traditional programming methods or legacy software.

In "Energy-Efficient Abundant-Data Computing: The N3XT 1,000×," a cross-institutional team of authors introduce Nano-Engineered Computing Systems Technology (N3XT), which reinterprets Moore's law to be more amenable to 3D manufacturing: instead of reducing dimensions in a chip's *XY* plane, their approach retains fixed dimensions in the *XY* plane and increases the number of features in the *Z* dimension. N3XT is based on carbon nanotubes and memristors, but other possibilities exist. The resulting computers could demonstrate benefits when programmed with existing languages, but they could also be programmed with new neuromorphic methods.

IF MOORE'S LAW IS TO RESUME, IT MUST BE BASED ON AN ARCHITECTURE OTHER THAN MULTICORE.

In "Superconducting Computing in Large-Scale Hybrid Systems," D. Scott Holmes, Alan M. Kadin, and Mark W. Johnson challenge the assumption that all computers must use the same underlying technology. Highlighting the speed and energy-efficiency advantages of superconducting electronics, they propose Josephson junctions as the new active elements for logic gates in the implementation of larger installations that could execute much of the existing software base. Of course, the computing industry includes intrinsically small systems such as Internet of Things gadgets and smartphones as well as enormous supercomputing datacenters and the servers that handle smaller systems' backend computing. With trillions of dollars in annual revenue, the industry is large enough to leverage room-temperature technology for mobile devices and faster, more power-efficient cryogenic technology for datacenters to exact the right mix of benefits.

In "Adapting to Thrive in a New Economy of Memory Abundance," Kirk M. Bresniker, Sharad Singhal, and R. Stanley Williams describe how the availability of persistent memory, enabled by novel device physics, could lead to replacement of the current processor-centric computing model based on Moore's law with a memory-driven computing model. Until recently, the relative cost and performance advantages of computing throughput versus accessing data storage has limited the development of data-intensive applications. However, the emergence of 3D memory makes such applications more attractive and could serve as a key technology driver for data analytics.

In "Architecting for Causal Intelligence at Nanoscale," Santosh Khasanvis and his colleagues outline the advantages of randomness in conjunction with new neuromorphic programming methods. Randomness is contrary to current expectations that computers produce deterministic results and that any deviation is an error. Algorithms that make use of random sampling as a fundamental aspect of their function can be executed on conventional computers using software-based random-number generators, but nanoscale devices whose behavior is random in well-understood and controlled ways could be vastly more efficient. The authors propose a

class of such devices that is effective in a probabalistic reasoning context calculation, rather than mere numerical analysis, to address a wide range of problems including machine learning.

Finally, in "Ohmic Weave: Memristor-Based Threshold Gate Networks," David J. Mountain and his colleagues describe how today's Turing-derived machines could incorporate neuromorphic capabilities. Neuromorphic computing research often focuses on understanding the brain's learning and pattern-recognition capabilities and how best to reproduce them in computers, leading to supercomputer clusters that can learn from visual stimuli such as YouTube videos. Although these clusters model enormous numbers of neurons and synapses, they are organized into just a handful of layers. This contrasts with the multiple overlapping hierarchical structures in software that are much more amenable to engineering by groups of people. Integrating these approaches could lead to a computer with brain-like learning structures that are composed hierarchically like programs, subroutines, operating systems, and subroutine libraries.

ABOUT THE AUTHORS

THOMAS M. CONTE is a professor with joint appointments in the Schools of Computer Science and Electrical and Computer Engineering at Georgia Tech. His research interests include computer architecture and compiler code generation. Conte received a PhD in electrical engineering from the University of Illinois at Urbana–Champaign. He is the 2015 IEEE Computer Society president, co-chair of the IEEE Rebooting Computing Initiative, and a Fellow of IEEE. Contact him at conte@gatech.edu.

ELIE TRACK is CEO of nVizix LLC, a startup developing novel photovoltaic technology for solar power based in Stamford, Connecticut. His research interests include innovative high-efficiency solar cells as well as superconducting electronics and its applications in high-performance communications and computing. He received a PhD in physics from Yale University. Track is co-chair of the IEEE Rebooting Computing initiative, past president of the IEEE Council on Superconductivity, and a Fellow of IEEE. Contact him at elie.track@nvizix.com.

ERIK DEBENEDICTIS is a technical staff member in the Non-Conventional Computing Technologies Department at Sandia National Laboratories. His research interests include computing approaches across the entire technology stack, including further scaling of von Neumann architectures through device improvements and packaging strategies, brain-inspired computing approaches, and superconducting electronics. DeBenedictis received a PhD in computer science from Caltech. He is a member of the IEEE Rebooting Computing Initiative, IEEE, the IEEE Computer Society, the IEEE Superconductivity Council, ACM, and the American Physical Society. Contact him at epdeben@sandia.gov.

The direction of US science and technology policy has become increasingly clear during the time in which we prepared this special issue, and the articles that appear here closely align with the NSCI vision. We hope that they will stimulate your own ideas regarding novel strategies to enable continued technology scaling, and we invite you to share your thinking with IEEE's Rebooting Computing Initiative.

Although we alone selected the final lineup of articles for this special issue, hundreds of technical volunteers have shaped our thinking through the Initiative. The various companies, academic institutions, and government research labs that collaborate on the International Technology Roadmap for Semiconductors (ITRS 2.0; www.itrs2.net) have also contributed to the technical discourse in fundamental ways. In addition, the US Office of Science and Technology Policy (www.whitehouse.gov/administration/eop/ostp), under the Office of the President, has played a seminal role by directing federal resources and national attention to important strategic goals.

Selected CS articles and columns are also available for free at **http://ComputingNow.computer.org**.