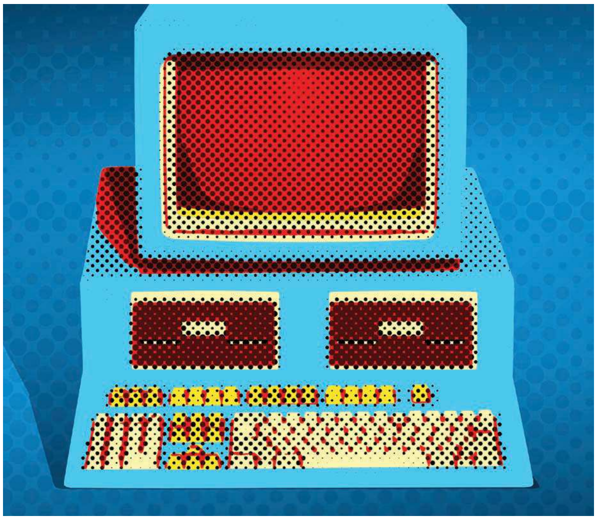

Moore's Law: The First Ending and a New Beginning

Andrew A. Chien, *University of Chicago*
Vijay Karamcheti, *Viridient Systems*

Moore's law has accurately predicted roughly biennial doubling of component capacity at minimal cost for almost 50 years. Recent flash memory scaling exhibits increased density, but reduced write and read lifetimes effectively constitute an ending of Moore's law. However, new resilience techniques, including adaptive management algorithms, and storage technologies based on information theory could maintain progress for a decade or more.

In April 1965 *Electronics*, a popular trade industry magazine, published "Cramming More Components onto Integrated Circuits" by Gordon E. Moore, the then relatively unknown director of R&D at Fairchild Semiconductor (his author biography at the end of the article described Moore as "one of the new breed of electronics engineers, schooled in the physical sciences rather than electronics"). As part of a broader discussion of the future of integrated semiconductor technology, Moore made the observation that "complexity for minimum component costs has increased at a rate of roughly a factor of two per year," going on to predict that "by 1975, the number of components per integrated circuit for minimum cost will be 65,000."[1] (Figure 1 reproduces Moore's original graph illustrating his point.)

MOORE'S LAW

Within five years, the phenomenon Moore described would come to be known as Moore's law, accurately predicting—or, perhaps more to the point, setting the parameters for—a steady series of semiconductor technology advances that have settled into a doubling of transistors per integrated circuit roughly every two years.

Leveraging the drive to much smaller dimensions using the principles of Dennard scaling,[2] the projection Moore outlined in 1965 has proven remarkably prescient. As of 2013, the industry had vastly surpassed his 65,000-component projection, with modern microprocessor designs routinely exceeding a billion transistors[3] and current flash memory chips incorporating 64 billion transistors for 128-gigabit capacity (remarkably these chips store two bits per chip).[4] Such chips power the digital consumer devices used by hundreds of millions of people around the world.

While most researchers and industry experts recognize that Moore's law cannot extend infinitely into the future, roadmaps such as the International Technical Roadmap for Semiconductors (ITRS) include detailed projections for three generations and extrapolated projections to 2025 (www.itrs.net/Links/2012ITRS/2012Chapters/2012Overview.pdf), suggesting that we are as much as a decade from its end.

HOW WILL MOORE'S LAW END?

Discussions based on physical device limitations, manufacturing roadblocks, and even economic constraints often frame the end of Moore's law as a quantum transition—one that will happen uniformly and all at once. Most experts, however, expect the shift will result from a

Published by the IEEE Computer Society

End of Moore's Law

Bob Colwell
Editorial Board Member, *Computer*

There is a subtlety that often goes unnoticed about Moore's law: it isn't about predicting the absolute maximum number of transistors that any given silicon technology generation can provide. Note carefully that Moore's diagram plotted cost against integration density. The more transistors you integrate, the lower the cost per transistor, but only down to a certain minimum. After that, the cost climbed steeply, a function of the lower yield from those additional components. The natural strategy was to design to that sweet spot, thus keeping your manufacturing costs low, and then ship large unit volumes, which grew your user base and set you (and them) up for the next silicon technology two years later. Rinse and repeat.

Looked at this way, it's clear that Moore's law is first and foremost about economics—not computers, not performance, not physics, not electronics, but economics. It's all about unit volumes—whoever ships volume, wins. That's because making the first of a new kind of chip can easily cost a company several billion dollars of upfront investment, but experience typically drives yields up and manufacturing cost down. That leads to price decreases, which drive volumes up even more.

So the game is to ship large unit volumes, which implies large numbers of customers. In turn, this leads directly to a question that virtually all businesses must contend with: how do I design my product so that potential customers will part with their hard-earned cash for it?

Historically, we chip designers traded in our double-transistor bounty per silicon generation for performance. We knew that there would be a demand for that performance boost, if for no other reason than Microsoft's latest operating system required it. We also sometimes found it useful to bundle new processor technology with other system improvements (PCI, USB, graphics) to boost demand for the lucrative CPUs. Prospective computer buyers could readily evaluate what they would be getting for their money, even buyers who were not computer extremophiles.

The few remaining silicon process nodes on the march to 7 nm or 5 nm do not promise the same kinds of improvements that we have seen in the past. In fact, each new generation is less "better" than the one before. There could well come a day, even before the underlying physics fail, when a new silicon process does not facilitate a chip that improves on its ancestors to the degree needed to keep buyers buying. Chipmakers will strive mightily to prevent this, and will probably turn to the same methods they have used in the past, improving other system aspects such as I/O and memory, adding other system-level features, and attempting to market their way out of a tough product sell. In the past, Moore's law rescued them and put them back on the rails one generation later. In the future, the world will continue to demand silicon, but will no longer pay a premium price for what they perceive as commodity chips. Commodity prices do not generate the profit margins needed to keep the silicon industry racing forward.

Think of it this way: isn't it grand that Moore's law has taken us to the point where we all carry smartphones with three times the computing horsepower of yesteryear's Cray supercomputers? It would have been a shame had silicon petered out at, say, 1999 technology, because then we would never have known the joys of mobile computing, car GPS units, and HD video. Perhaps some visionary will see what would have been next, had Moore's law somehow continued another 10 years, and can find an alternate route, such as exotic materials, clever combinations of analog and digital circuits, better hardware/software tradeoffs, better energy storage and management, more optimal networking, and the list goes on. Clearly, there is a lot to do—we better get started!

gradual slowing of the rate of improvement within specific technologies—CMOS logic, DRAM, flash, disk drives—occurring at different rates as we approach the fundamental limits of device scaling (devices no longer behaving properly below a certain size) or manufacturing capability (lithography becoming too expensive or unreliable, for example).

These predictions focus on Moore's law as relating primarily to *cost-effective capacity*, measured by the number of transistors per integrated circuit. Often overlooked, though, is a more subtle but still essential point Moore made in his seminal article: in addition to reducing cost per device, integrated circuits contribute to a *greater reliability factor* for electronics ("In addition, the improved reliability made possible by integrated circuits will allow the construction of larger processing units."), enabling a host of new, even unimaginable applications.

A complete characterization of Moore's law, then, would focus on the capacity to build exponentially larger, more complex assemblies of inexpensive as well as reliable electronics rather than inexpensive electronics alone. In fact,

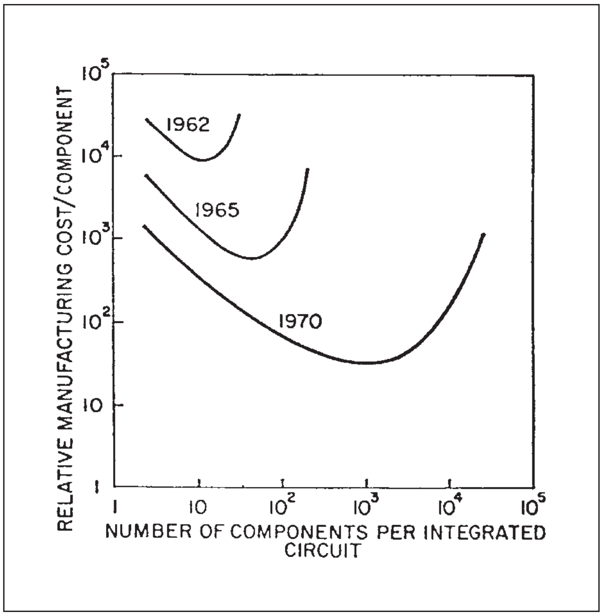

Figure 1. Gordon Moore's original 1965 diagram illustrating the phenomenon now known as Moore's law.

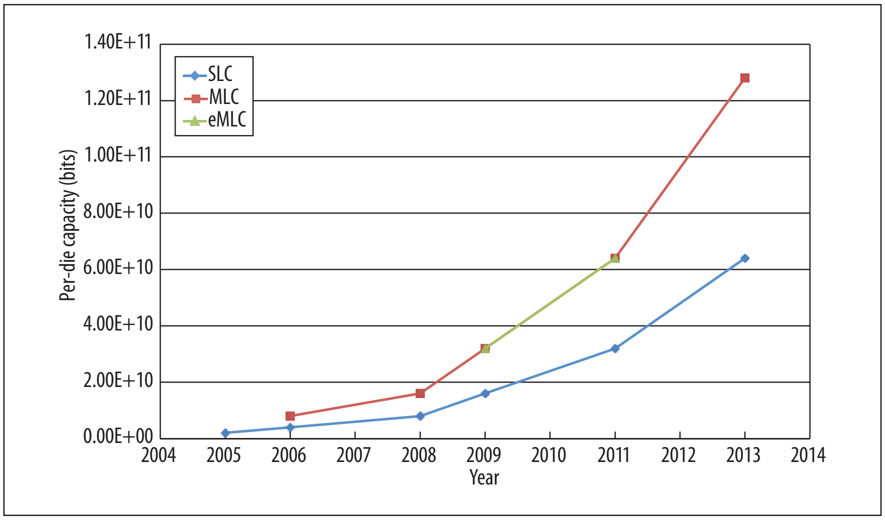

Figure 2. Flash memory chips' pattern of exponential capacity increase over eight years, for single-layer cell (SLC), multilayer cell (MLC), and enterprise MLC (eMLC) memory.

the profound impact of integrated circuitry since 1965 has resulted from a combination of reduced device cost, greater overall capability, power consumption savings, and enhanced reliability.

Therefore, from the perspective of reliability, what we call the "first ending" of Moore's law is already apparent, as we will show, in the evolution of flash memory. And because flash memory has for the past several generations represented the most advanced integrated semiconductor technology—leading in lithography (smallest feature size), density (greatest device integration), and cost (fewest nanocents per semiconductor)—flash technology, as the first victim, is potentially the harbinger for the end of Moore's law as we know it.

MOORE'S LAW AND FLASH MEMORY: THE "FIRST ENDING"

The exponential growth of integration technology predicted by Moore's law has no better exemplar than the 100-fold increase in flash memory chip capacity over the past eight years, as shown in Figure 2. This remarkable progress has sparked a revolution in personal mobile devices, with flash memory now providing workhorse storage for MP3 players, smartphones, tablets, even laptops. Flash has also gained a significant foothold in the enterprise storage domain, in the form of solid-state drives and a new class of high-performance memory arrays that are commonly called storage class memory.

Historically, flash memory has advanced via feature-capacity growth coupled with cell-design shrinkage, classic Moore's law drivers. More recent advances, however, have required an architectural shift from single-layer cell (SLC) memory to a more complex multibit or multilayer cell (MLC) memory model and its enhanced-endurance variant, enterprise MLC (eMLC).

So what Figure 2 actually illustrates is flash memory's exponential improvement in cost-effective capacity, adhering remarkably well to that one aspect of Moore's law—with per-die capacities doubling every two years. But sustaining this generation-over-generation capacity growth has required significant tradeoffs, as shown in Table 1, which traces evolving characteristics of the three different flash device types—SLC, MLC, and eMLC—spanning six semiconductor process generations.

Flash memory cells record information by storing charge in a structure of the semiconductor called a floating gate; MLC devices record two bits of information by storing four different charge levels. At deep submicron device feature sizes, as are required for > 64-gigabit densities, great care must be taken in the charge-storing process (write) as well as in the process used for sensing the amount of stored charge (read).

Despite the most critical level of care during these processes, device endurance limits (the number of times you can reliably inject and drain charge from the floating gate) inevitably decrease at extreme submicron feature sizes. Because of this, currently available flash cells degrade, or "wear out," quite rapidly, with each shrink in cell design negatively impacting the reliability, endurance, and performance characteristics of flash devices. This impact is most clearly evident in MLC devices, where much tighter controls are required to reliably disambiguate, or distinguish dependencies for, the four charge levels associated with each cell's two bits of information.

Consequently, if we consider year-over-year trends for flash capacity in terms of reliability, a quite different picture emerges than that in Figure 2—not exponential progress but actual stagnation. Figure 3 plots the trends for SLC and MLC devices charted in Table 1, using lifetime write and read capacity of flash devices as a proxy for reliable capacity. The write lifetime (number of program/erase cycles) limits quantified in Table 1 are widely acknowledged by systems designers; less recognized are the read disturb thresholds that couple write and read lifetimes, thereby producing limited *usable* read lifetimes for MLC flash memory. (We omit SLC read lifetimes because they are much higher, and not a primary endurance concern.)

In addition, for purposes of comparison with these actual trends, Figure 3 plots projected Moore's law increases for SLC writes and MLC writes and reads, in each case showing a doubling of the number of reliable

Table 1. Characteristics of flash devices across six process generations.

	2005	2006	2008	2009	2011	2013
Geometry	9x	7x	5x	3x	2x	2y
Per-die capacity	2 Gbits	4 Gbits **8 Gbits**	8 Gbits **16 Gbits**	16 Gbits **32 Gbits** **32 Gbits**	32 Gbits **64 Gbits** **64 Gbits**	64 Gbits **128 Gbits**
Error correction requirement	1 bit/512 bytes	1 bit/512 bytes **4 bits/512 bytes**	4 bits/512 bytes **8 bits/512 bytes**	4 bits/512 bytes **12 bits/512 bytes** **24 bits/1 Kbyte**	8bits/512 bytes **24 bits/1 Kbyte** **40 bits/2 Kbytes**	24 bits/1 Kbyte **40 bits/1 Kbyte**
Endurance (program/erase cycles)	100,000	100,000 **10,000**	100,000 **10,000**	100,000 **5,000** **30,000**	60,000 **3,000** **30,000**	60,000 **3,000**
Read disturb threshold (reads/erase block)	—	—	—	— **~1000,000** **~200,000**	— **~600,000** **~150,000**	— **~400,000**
Program time (tPROG, µs)	~300	~300 **~700**	~300 **~900**	~300 **~1000** **~1400**	~400 **~1400** **~1400**	~500 **~1600**
Read time (tR, µs)	25	25 **50**	25 **50**	25 **50** **75**	35 **75** **100**	35 **100**

Values specific to MLC devices are bold, and those specific to enhanced-endurance MLC (eMLC) devices are underlined-bold. Data collated from specifications of devices offered by multiple vendors corresponding to a given process geometry node.

transistors for each generation. This comparison helps highlight the growing divergence of actual flash memory performance from the Moore's law ideal, a divergence that is earlier and more pronounced for MLC and the recent 20-nm process generation of flash memory. Most striking is the total lifetime reads plot for MLC devices, where progress has been essentially flat for three generations. The net increase of *reliable* devices in flash memory has effectively ceased.

For flash memory, then, we have reached the end of Moore's law. This stagnation comes at a point where application demands for flash memory read and write lifetimes continue to increase, driven by ever richer media and analytics applications and made accessible by faster interfaces such as open NAND flash interface (ONFI) and toggle interface modes. Tenfold improvements in the latter over the last three process generations imply that static read lifetimes translate to comparably faster wear-out.

The trends shown in Figure 3 are representative of what lies ahead for all types of flash memory, not just MLC. Increased-endurance eMLC devices may appear to offset some of the lifetime write slowdowns, but just like MLC devices, they do so at the expense of lifetime reads. Moreover, the data in Table 1 show that all types of flash memory are experiencing reductions in write and read performance from generation to generation. Trends for performance-normalized reliable capacity are even worse.

But what does this "first ending" of Moore's law for flash memory portend for advances in semiconductor design

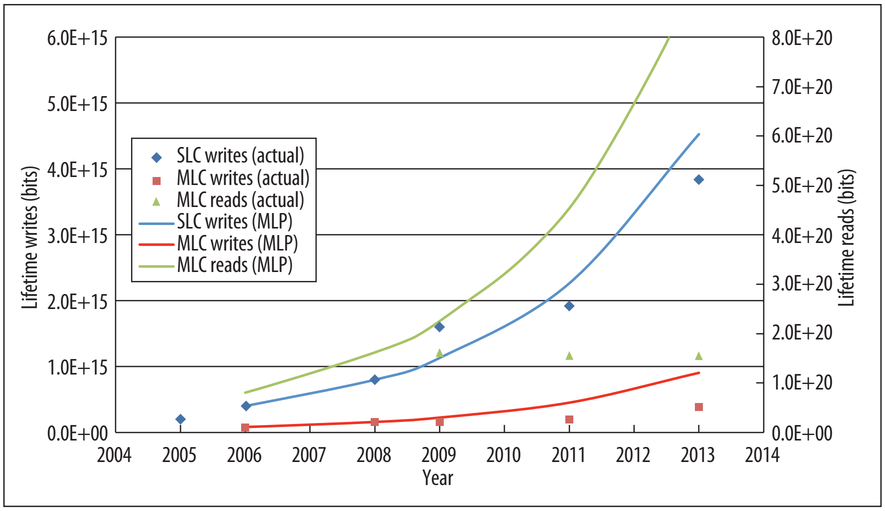

Figure 3. Actual write capacity for SLC and actual write capacity and read capacity for MLC flash memory chips over device lifetime from year of release. Projections based on Moore's law (MLP) over the same time, doubling each generation, are shown for comparison. (SLC read values, which are not a concern here, have been omitted.)

more generally? Will we soon see its ending in other technologies?

A NEW BEGINNING?

We remain optimistic about the future of Moore's law, in some version at least, for several reasons.

First, flash memory still has "legs." In absolute terms, lifetime capacity reliability metrics for flash memory already meet the practical lifetime requirements of many specific applications. As long as absolute read and write values remain adequate for these purposes, such applications will continue to realize benefits from the price-per-bit decreases offered by greater integration.

Second, emerging storage technologies and architectural techniques show great potential. For the next decade or longer, rapid progress in nonvolatile memory will be facilitated by an extraordinary breadth of research and development already under way, spanning new manufacturing techniques, new device structures, and new storage-cell technologies. While as yet still maturing, these cell technologies—such as phase-change memory (PCM) and resistive random-access memory (ReRam)—are particularly promising, offering greater write lifetimes, fine-grained access, lower energy expenditure, and, with multilayer integration, potential densities as high as one terabit per chip.[5–8] While falling short of Moore's law scaling over three decades, terabit-level densities would increase by 16-fold beyond current 128-gigabit densities—offering nearly a decade's worth of exponential growth.

Third, and most promising, sophisticated algorithms and information theory offer significant opportunities for capability advances in spite of degraded scaling. Both research areas have demonstrated sustained exponential improvement for decades,[9,10] and both support continued capability increases in storage systems built from elements that are less reliable over time. In fact, algorithmic advances in wear-leveling, scheduling, write-behind, and other functions already support key performance increases for flash memory.[11]

Promising information theory techniques include error-correcting codes applied to words, blocks, and sectors and other advanced codes such as Reed-Solomon, Turbo, and low-density parity check (LDPC) applied statically and adaptively.[12,13] These powerful fundamental drivers—packaged and delivered in systems abstractions that mask their internal complexity, and formulated to leverage the inexpensive local computations that Moore's law has historically provided in abundance—can certainly bring continued capability growth, perhaps even exponentially.

In the context of flash storage, for example, several companies offer components to manage endurance and read-disturb thresholds, dynamically separating weaker flash memory regions from better behaved ones. Techniques like these can improve reliability by several orders of magnitude, converting a less than reliable MLC device into high-quality enterprise storage while trading only linear amounts of device capacity.

But the question remains whether advanced algorithmic techniques and information theory can have an effect beyond increasing semiconductor storage capability. The spread of new challenges, such as practically finite lifetimes and the increasing possibility for errors, across a range of semiconductor technologies[14,15] requires advanced resilience techniques.[16] It will be exciting to watch as research develops applying these techniques to the design, testing, and management of semiconductor devices beyond the realm of memory.[17–19]

We end by quoting again from Moore's groundbreaking 1965 article:

> Integrated electronics will make electronic techniques more generally available throughout all of society, performing many functions that presently are done inadequately by other techniques or not done at all. The principal advantages will be lower costs and greatly simplified design—payoffs from a ready supply of low-cost functional packages. … The availability of large functions, combined with functional design and construction, should allow the manufacturer of large systems to design and construct a considerable variety of equipment both rapidly and economically.

As we look to the future, Moore's principles, if not his law specifically, will continue to guide us. C

References

1. G.E. Moore, "Cramming More Components onto Integrated Circuits," *Electronics*, vol. 38, no. 8, 1965, pp. 114-117.
2. R. Dennard et al., "Design of Ion-Implanted MOSFETs with Very Small Physical Dimensions," *IEEE J. Solid State Circuits*, vol. 9, no. 5, 1974, pp. 256-268.
3. S. Borkar and A. Chien, "The Future of Microprocessors," *Comm. ACM*, vol. 54, no. 5, 2011, pp. 67-77.
4. "Intel and Micron Extend NAND Flash Technology Leadership, Introduce Industry's Smallest, Most Advanced 20-Nanometer Process," *Intel Newsroom*, 14 April 2011; http://newsroom.intel.com/community/intel_newsroom/blog/2011/04/14/intel-micron-extend-nand-flash-technology-leadership-introduce-industrys-smallest-most-advanced-20-nanometer-process.
5. G.W. Burr et al., "An Overview of Candidate Device Technologies for Storage-Class Memory," *IBM J. Research and Development,* vol. 52, nos. 4-5, 2008, pp. 449-464.
6. P. Wong et al., "Phase Change Memory," *Proc. IEEE*, vol. 98, no. 12, 2010, pp. 2201-2227.
7. S. Cueff et al., "Resistive Switching in Silicon Suboxide Films," *J. Applied Physics,* vol. 111, no. 7, 2012; doi:10.1063/1.3701581.
8. M.J. Kang et al., "Pram Cell Technology and Characterization in 20nm Node Size," *Proc. Int'l Electron Devices Meeting* (IEDM 11), IEEE, 2011, pp. 3.1.1–3.1.4.

9. President's Council of Advisors on Science and Technology, "Progress in Algorithms Beats Moore's Law," *Designing a Digital Future: Federally Funded Research and Development in Networking and Information Technology*, report to the President and Congress, Dec. 2010, p. 71.
10. Federal Communications Commission, *Mobile Broadband: The Benefits of Additional Spectrum*, tech. paper, Dec. 2010.
11. E. Gal and S. Toledo, "Algorithms and Data Structures for Flash Memories," *ACM Computing Surveys*, vol. 37, no. 2, 2005, pp. 138-163.
12. R. Gallager, *Information Theory and Reliable Communication*, Wiley, 1968.
13. D. Roberts, T. Kgil, and T. Mudge, "Integrating NAND Flash Devices onto Servers," *Comm. ACM*, vol. 52, no. 4, 2009, pp. 98-103.
14. P. Kogge et al., *Exascale Computing Study: Technology Challenges in Achieving Exascale Systems*, DARPA-IPTO, 2008.
15. Nat'l Security Agency Advanced Computing Systems, *Inter-Agency Workshop on HPC Resilience at Extreme Scale*, summary report, 2012; http://institute.lanl.gov/resilience/docs/Inter-AgencyResilienceReport.pdf.
16. A. DeHon, N. Carter, and H. Quinn, eds., *Final Report for CCC Cross-Layer Reliability Visioning Study*, 3 Mar. 2011; www.cra.org/ccc/xlayer.php.
17. E. Mintarno et al., "Self-Tuning for Maximized Lifetime Energy-Efficiency in the Presence of Circuit Aging," *IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems*, vol. 30, no. 5, 2011, pp. 770-773.
18. T. Austin et al., "Opportunities and Challenges for Better than Worst Case Design," *Proc. 10th Asia and South Pacific Design Automation Conf.* (ASPDAC 05), vol. 1, IEEE, 2005, pp. I/2-I/7.
19. A.B. Kahng et al., "Slack Redistribution for Graceful Degradation under Voltage Overscaling," *Proc. 15th Asia and South Pacific Design Automation Conf.* (ASPDAC 10), IEEE, 2010, pp. 825-831.

Andrew A. Chien *is the William Eckhardt Professor of Computer Science and Senior Fellow in the Computation Institute at the University of Chicago, and a senior computer scientist at the Argonne National Laboratory. His research interests include programming and computer architecture for exascale computers, ease of use for data-intensive computing tools, and programming models and tools for post-Moore's law computing substrates. Chien received a PhD in computer science from MIT. He is a Fellow of IEEE, ACM, and AAAS. Contact him at achien@cs.uchicago.edu.*

Vijay Karamcheti *is cofounder and chief technology officer of Virident Systems. His research interests include high-performance flash-based enterprise storage solutions and grid computing. Karamcheti received a PhD in electrical and computer engineering from the University of Illinois at Urbana-Champaign. Contact him at vijayk@virident.com.*

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