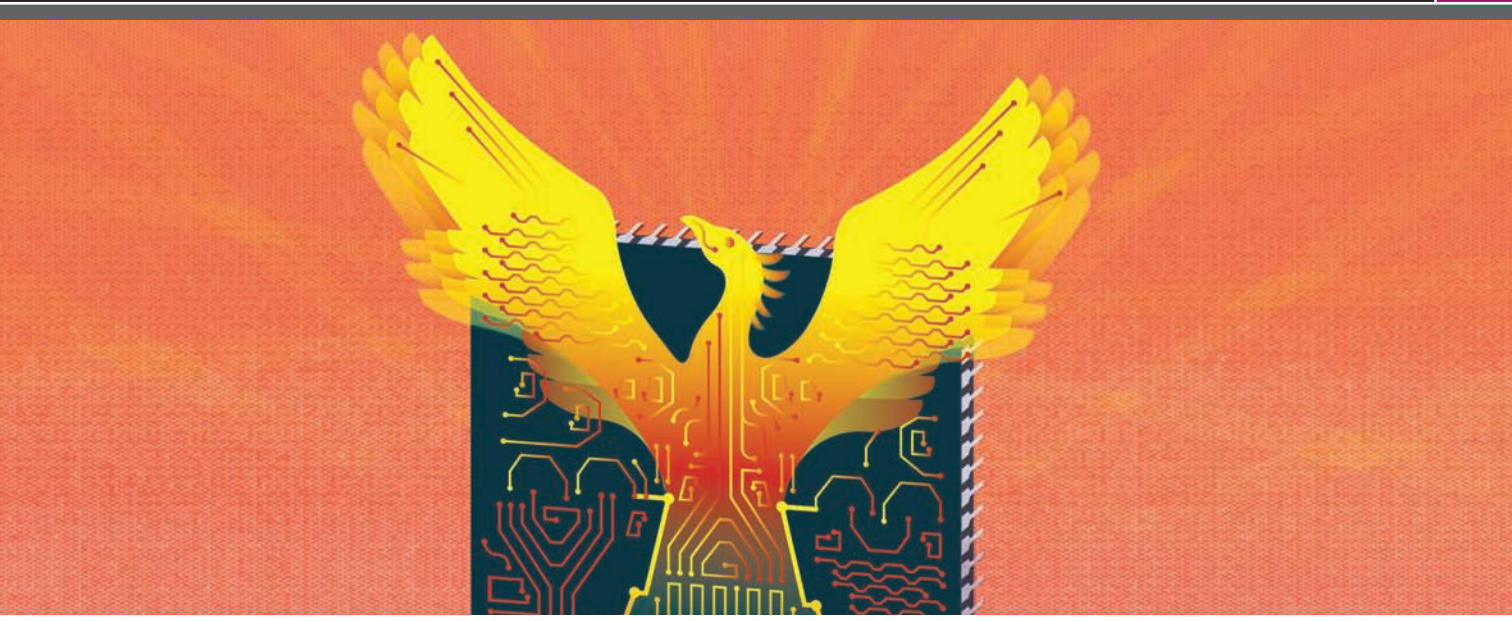

# The End of Moore's Law:
## A New Beginning for Information Technology

**Thomas N. Theis** | Columbia University
**H.-S. Philip Wong** | Stanford University

Far from signaling an end to progress, the gradual end of Moore's law will open a new era in information technology as the focus of research and development shifts from miniaturization of long-established technologies to the coordinated introduction of new devices, new integration technologies, and new architectures for computing.

*"From the end spring new beginnings."—Pliny the Elder*

Gordon Moore's 1965 and 1975 papers[1,2] still shine brilliantly. While the value of continued miniaturization of electronic components was already well understood in the 1950s,[3] Moore's 1965 paper championed the importance of integrated circuits at a time when many still felt that the job of semiconductor manufacturers was to deliver discrete diodes and transistors so that designers could build their own circuits.[4] The 1965 paper also described how integration complexity can be traded for manufacturing yield to minimize cost per component, and on that basis, predicted continued exponential improvements in cost and complexity with foreseeable improvements in manufacturing processes. Moore's 1975 paper presented another important insight: advances in integration complexity come from three distinct factors—increase in silicon die size, reduction in feature size, and "device and circuit cleverness." These insights from 1965 and 1975 have broadly guided investment in semiconductor technology development ever since. On top of that, Moore's famous 1975 prediction that circuit complexity would double every two years[2] proved remarkably prescient and astonishingly durable. But all exponential trends must come to an end.

 Copublished by the IEEE CS and the AIP 1521-9615/16/$33.00 © 2016 IEEE 

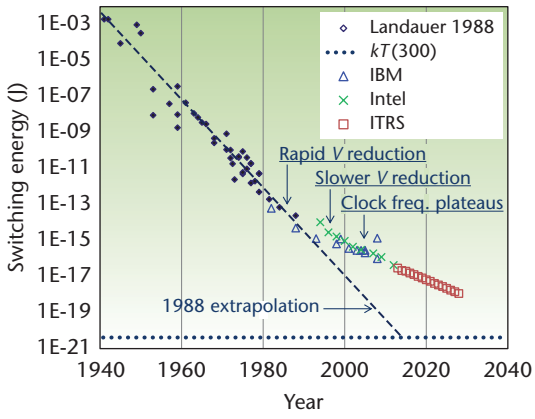


**Figure 1.** Minimum switching energy dissipation in logic devices used in computing systems as a function of time. Black diamonds replicate data from Rolf Landauer,[5] and the dashed line is Landauer's 1988 extrapolation of the historic trend toward $kT$ (evaluated at $T = 300$ K), indicated by the dotted line. Triangles and Xs are published values from IBM and Intel, respectively, compiled by Chi-Shuen Lee and Jieying Luo at Stanford University during their PhD thesis research with one of the authors (Wong). Open squares are values from the 2013 International Technology Roadmap for Semiconductors (ITRS). The data is available at https://purl.stanford.edu/gc095kp2609. Current, up-to-date data can be accessed at https://nano.stanford.edu/cmos-technology-scaling-trend.

Figure 1 shows a very important exponential trend in information technology that already shows a sharp slowing of progress. In 1988, Rolf Landauer[5] published some remarkable data on energy dissipation in computing that had been collected over many years by his IBM colleague Robert Keyes. From the 1940s through the 1980s, a span that includes the replacement of vacuum tubes by bipolar transistors, the invention of the integrated circuit, and the early stages of the replacement of bipolar transistors by field-effect transistors (FETs), the typical energy dissipated in a digital switching event dropped exponentially by over 10 orders of magnitude. We replotted that data in Figure 1 along with Landauer's extrapolation of the trend toward switching energies on the order of the thermal fluctuation energy, $kT$, evaluated at $T = 300$ K. Landauer was well aware that the switching energy would not approach $kT$ around 2015, not with the established complementary metal-oxide-semiconductor (CMOS) device and circuit technology. His extrapolation was a way of highlighting the possibility of, and perhaps the need for, a new way of computing.

Some have mistaken the $kT$ per switching event as a fundamental lower bound on the energy consumption of digital computation. Landauer knew that it isn't. His 1988 publication reviewed fundamental research demonstrating the possibility of an energy-conserving form of computation. As in today's circuits, the devices in energy-conserving circuits would store enough energy—many times $kT$—to reliably distinguish the digital state from the inevitable thermal noise. For good engineering reasons, today's circuits dissipate that stored energy every time a device is switched. In contrast, energy-conserving circuits would dissipate only a small fraction of the stored energy in each switching event. In such circuits, there's no fundamental lower bound on the energy efficiency of digital computation.

Although no commercially viable energy-conserving computing systems emerged in the 1990s or later, digital quantum computing, still in its infancy, exemplifies the energy-conserving approach. To show what did happen in the commercial sector after 1988, we added data to Figure 1 showing switching energies for minimum channel width CMOS FET technologies based on technical publications from IBM and Intel. For a while, switching energy continued to drop rapidly, as IBM led the industry in rapidly reducing operating voltage. Roughly following the elegant scaling rules laid out by Robert Dennard and colleagues,[6] each successive generation of smaller, lower voltage, lower power devices was also faster. Increasingly potent CMOS technology extended the long run of exponential increases in microprocessor clock frequency[7]—a key measure of computing performance—that had begun with the Intel 4004 in 1972. And the ever smaller, ever cheaper transistors enabled rapid elaboration of computer architecture. For example, the introduction in the 1990s of sophisticated approaches to instruction-level parallelism (superscalar architectures) further multiplied the system-level performance gains from increasing clock speed. By the late 1990s, CMOS had displaced the more power-hungry bipolar transistor from its last remaining applications in high-performance computing. Despite these triumphs, the historic rate of reduction in switching energy couldn't be maintained through the 1990s as the FET approached some fundamental constraints to its further development.

## Physical Constraints on Continued Miniaturization

Note that this slowing of a highly desirable exponential trend in information technology—the break in slope so evident in Figure 1—has nothing to do with the approach of switching energy toward the thermal fluctuation energy. Even in

It's no surprise that the replacement cycle for computing equipment of all sorts has lengthened.

these minimum channel width devices from IBM and Intel, the stored energy that distinguishes digital state is orders of magnitude greater than $kT$. In practical logic circuits, with wider channel devices and associated wiring, the stored energy is roughly 100 times greater than the minimum switching energies shown in Figure 1. Thus, in today's smallest commercial devices, thermal upsets of digital state are extremely unlikely and therefore irrelevant to device and circuit design.

So what *did* change in the 1990s? Miniaturization of devices continued at pace but along lines that increasingly deviated from Dennard's scaling rules. In particular, the gate insulator thickness and the operating voltage could no longer be simply reduced along with other device dimensions. Further reduction in insulator thickness would have resulted in unacceptable (and exponential) increases in gate leakage current through direct quantum tunneling. Further reduction in operating voltage swing would have resulted in either unacceptably low channel current in the "on" state (unacceptable decreases in switching speed) or increased leakage current in the "off" state (unacceptable increases in passive power). The physics that limits further voltage reduction is well known and relatively straightforward.[8] Suffice it to say that with operating voltages now on the order of 1 V, the FET is close to its voltage scaling limit for operation at room temperature and above. Future voltage reductions will be limited.

Detailed and extensive analyses of these and other scaling issues can be found in the literature,[9–11] but a simple scaling model provides insight regarding their impact on the industry.[8] The model assumes that continued rapid innovation in materials and device structures[12,13] will sustain the Moore's law trends in integration density and device switching speed. It further assumes that operating voltage is fixed from technology generation to technology generation. Under these simplifying assumptions, the areal power density grows exponentially from generation to generation unless clock frequency is fixed and processor cores are brought onto the die at a rate significantly less than what would be possible based on lithographic ground rules.[8]

This simple model thus gives a surprisingly good account of some broad developments in microelectronics in the past decade. Clock frequencies plateaued between 2003 and 2005 and have been stagnant ever since. The performance of today's systems is increasingly power constrained. The devices and circuits could be clocked at higher frequencies but only at unacceptable levels of power and heat generation that would compromise critical attributes such as battery life in consumer products and the cost of cooling and powering servers in large datacenters. And the introduction of multiple cores has been slower than the rate that could be supported by advances in lithography and integration density. Perhaps this reflects the failure of software developers to find ways to fully exploit core-level parallelism for many important applications of computing, but the model suggests that the physics of CMOS transistor switching and the resulting heat generation has been an additional, if not dominant, limiting factor. The outlook for the future is more of the same, as suggested by the forward-looking data in Figure 1 from the 2013 International Technology Roadmap for Semiconductors. The modest reductions in switching energy are the projected result of continued aggressive reduction of device size and very modest reductions in operating voltage. These advances would support a glacially slow 4 percent compounded annual growth in clock frequency through 2028.

The break in slope in Figure 1 thus marks the beginning of a still-ongoing transition to an era of constant voltage transistor scaling. This transition is having a big effect on the microelectronics industry, as fewer and fewer manufacturers strive to maintain the Moore's law development cadence. Many observers, including Gordon Moore himself,[4] have pointed to capital costs associated with manufacturing, which are rising much faster than industry revenue, as the eventual limiting factor. To this factor, we must add the impact of a declining return on investment for developing each new generation of smaller devices and denser circuits. For the three decades prior to 2005, clock frequency, integration density, and cost per device all improved exponentially with each technology generation, while active and passive power were contained within economically acceptable bounds. Since 2005, integration density and cost per device have continued to improve, and manufacturers have emphasized the increasing number of processors (cores) and the amount of memory they can place on a single die. However, with clock frequencies stagnant, the resulting

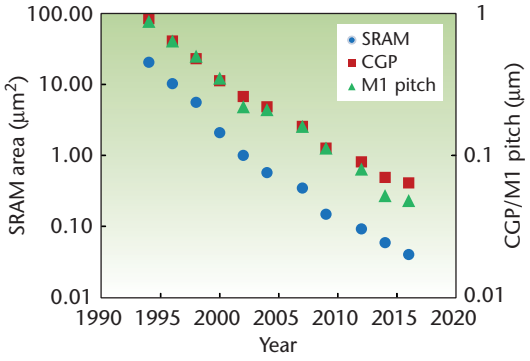


**Figure 2.** Three key measures of integration density as a function of time. Blue dots show static random access memory (SRAM) density. Green triangles show M1 pitch, the minimum wire-to-wire spacing in the first wiring layer. Red squares show CGP or contacted gate pitch, the minimum spacing between transistors. Progress is still rapid, but there's evidence of slowing in recent years. Data compiled from published literature by Chi-Shuen Lee at Stanford University during his PhD thesis research with one of the authors (Wong). The data is available at https://purl.stanford.edu/gc095kp2609. Current, up-to-date data can be accessed at https://nano.stanford.edu/cmos-technology-scaling-trend.

performance gains have been muted. Furthermore, the more straightforward elaborations of the standard von Neumann computer architecture have already been implemented,[7] and prospects for significant performance gains from further increases in parallelism appear limited even at the multicore level.[14,15] It's no surprise then that the replacement cycle for computing equipment of all sorts has lengthened. An increasing number of semiconductor manufacturers are finding profits by investing in development of product attributes, such as architectures for improved memory access, that have little to do with smaller feature size.

Figure 2 shows progress in three key indicators of the achievable integration density for complex digital logic circuits since 1994. Progress is still rapid, but taken together, these indicators show a significant slowing in the last decade. The Moore's law development cadence—the regular doubling of integration density—appears to be slowing.

While end users still see improvements in performance and energy efficiency as greater numbers of memory devices are brought closer to processor cores, these gains cannot be long sustained with current memory devices and architectures. Static random access memory (SRAM), the fast memory that now occupies over half the surface area of a microprocessor chip, uses six FETs to store a bit of information. Dynamic random access memory (DRAM), the slower but denser and therefore less expensive memory that usually resides on memory chips peripheral to the processor, uses one FET and one capacitor to store a bit. Flash memory, the very dense but rather slow memory that stores data when the power is off, uses one FET with a specially designed gate structure to store one bit (or more recently, several bits). Thus each of these dominant devices in today's memory hierarchy is subject to scaling constraints similar to those for the FETs used in logic, plus additional constraints unique to each device.

Despite these daunting problems, we're very optimistic about the prospects for further dramatic advances in computing technology. After decades of progress centered on miniaturization of the CMOS transistor, we see a growing potential for advances based on the discovery and implementation of truly new devices, integration processes, and architectures for computing. By truly new devices, we mean devices that operate by physical principles that are fundamentally different from the operating principle of the FET and are therefore not subject to its fundamental limits, particularly the voltage scaling limit. By truly new integration technologies, we mean monolithic integration in three dimensions in a fine-grained manner that immerses memory within computational units. And by truly new architectures, we mean circuit- and higher-level architectures that are much more energy efficient than the von Neumann architecture, particularly for the important algorithms and applications of the coming decades. We now touch briefly on some of the emerging research concepts that fuel our optimism.

## New Devices for Logic

As we write, several distinct physical principles are known by which a voltage-gated switch (that is, a transistor-like device) might avoid the voltage scaling limit of the conventional FET.[8,16] For example, some of these operating principles invoke a physical mechanism that breaks the direct link between externally applied operating voltage and the internal potential that gates the flow of current. Of course, just changing the device physics to embody one of these operating principles doesn't guarantee a more energy-efficient low-voltage digital switch. For each proposed device concept, the switching characteristics and other important device attributes will depend critically on the achievable properties of materials and the details of the device structure.

All currently proposed low-voltage device concepts are still in the early research stage. Laboratory prototypes do not yet exhibit characteristics that would justify focused commercial product development. However, many of these device concepts are evolving rapidly as researchers discover and understand the problems and invent solutions. It therefore seems likely that additional low-voltage devices will be invented. If a high-performance low-voltage device does emerge in coming years, it could greatly loosen the power and heat generation constraints that currently limit computing.

## New Devices for Memory

Several forces drive the exploration of new devices for memory. We already mentioned the increasing difficulty in further miniaturization of the established memory devices—namely, SRAM, DRAM, and flash. New memory devices might be more easily scaled to smaller sizes. A second driving force is the changing computing workload. The memory hierarchy employed today has been optimized for applications with data locality, yet a growing fraction of common workloads involve loosely structured data and require frequent memory access across a large address space. Such data movement is expensive in latency as well as energy consumption, especially when data need to come from off-chip memory through a data bus with limited bandwidth.[17] Off-chip memory access can account for as much as 90 percent of energy and commensurate execution time in today's computing systems running data-intensive algorithms.[17] Finally, and perhaps most importantly, some emerging memory devices may allow more cost-effective integration of large amounts of memory with logic. Of today's dominant memory technologies, only SRAM can be readily integrated with high-performance CMOS logic. Integrating DRAM and flash on the same chip with processor cores is difficult and often not cost-effective.

Thus, many new memory device options are being explored,[18,19] including spin transfer torque magnetic RAM (STT-MRAM), ferroelectric RAM (FERAM), conductive bridge RAM (CBRAM), resistive RAM (RRAM), and phase change memory (PCM). All of these memories share some highly desirable attributes: they're nonvolatile, each cell of the memory array can be randomly read without destroying the information stored and written without first erasing the stored bit, they cover a broad range of read/write characteristics that span the entire memory hierarchy, and they're fabricated using temperatures below those used for fabricating the interconnections (the wires connecting transistors).

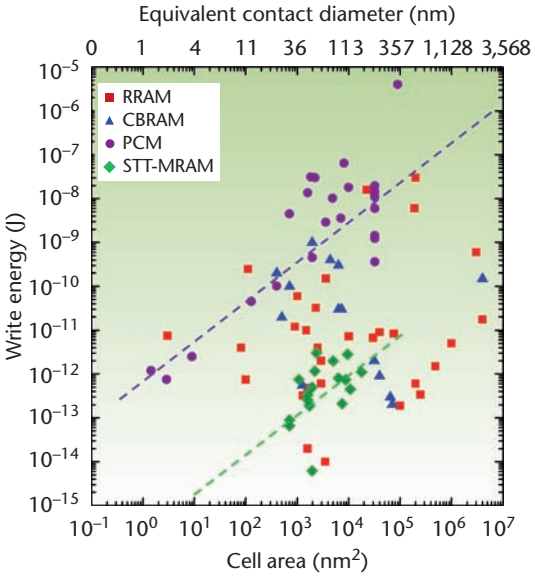


**Figure 3.** Programming energy versus memory cell area from published data for leading emerging nonvolatile memory technologies: spin transfer torque magnetic random access memory (STT-MRAM), phase change memory (PCM), conductive bridge random access memory (CBRAM), and resistive random access memory (RRAM). Both STT-MRAM and PCM require a critical current density to switch the memory state, thus the programming energy is proportional (purple and green dashed lines) to the memory cell area. Conduction in CBRAM and RRAM is filamentary, therefore the programming energy is independent of the memory cell area. Data from https://nano.stanford.edu/stanford-memory-trends.

The first three attributes open vast opportunities for rethinking the design of the memory hierarchy to optimize energy dissipation and performance for various application workloads. The last attribute enables incorporation of memory devices above blocks of CMOS logic and thus allows seamless, fine-grained integration of memory and logic.

While sharing desirable common attributes, each of these emerging technologies differs from and complements the others on the basis of key attributes: read/write speed, read/write power and energy consumption, retention and endurance properties, and device density (area of the memory cell). For example, Figure 3 shows published data[20] for write energy plotted against memory cell area. The tradeoffs vary greatly among the various types of memory. The physics of magnetic switching makes STT-MRAM singularly fast and thus well suited to be placed close to processor cores. However, one of its present limitations is the high write energy and resulting energy consumption. The write energy for STT-MRAM (and also for PCM) is proportional to the memory cell area because a

Much of the interest in nanomagnetic logic lies in the promise of combining the functions of logic and memory in a single device.

certain current density is required to switch cell resistance. In contrast, RRAM and CBRAM write energy shows no dependency on the memory cell area because of the filamentary nature and very small cross-sectional area of the conduction path. Continued miniaturization should reduce the energy consumption of STT-MRAM and PCM to acceptable levels. Continued advances in materials and device physics could further reduce the energy consumption of these memories.

## New Devices Combining the Functions of Memory and Logic

Magnetism has long been the basis for information storage devices such as the hard-disk drive and STT-MRAM. As a magnet is made smaller, less energy is required to switch its polarization. Furthermore, physicists and materials scientists have in recent years discovered new, energy-efficient switching mechanisms. Researchers are therefore beginning to explore and exploit the new physics of nanomagnetism in devices for digital logic.

Early device concepts for magnetic logic suffered from the drawback that there was no simple and direct way for the magnetic state of one device to switch the magnetic state of another device. (A circuit designer would point out that the devices don't concatenate.) Thus All-Spin Logic, a nanomagnetic device and circuit family that solves this problem,[21] generated significant interest when it was proposed. Another exciting research direction is voltage-controlled magnetism.[22] Compared to a current-controlled switching mechanism such as that employed in STT-MRAM, voltage-switched magnetic devices should be faster and more energy efficient.

Much of the interest in nanomagnetic logic lies in the promise of combining the functions of logic and memory in a single device. Such devices could eliminate the need to save the state of a computation to memory before the power is turned off. This capability would be of immediate value in power-starved systems dependent on intermittent power sources, and in the longer term, it could profoundly change computer architecture.

## New Integration Processes

The realization of monolithically integrated multilayer logic and memory would be a revolution, and that revolution could already be brewing. Multilayered flash memory with 48 layers is already in production and represents one of the earliest device technologies to truly embrace monolithic 3D integration as a route for technology advancement.

The mixing of logic and memory in monolithic 3D stacks is much easier if the high temperatures often required for synthesis of successive layers of electronic materials can be avoided in the low-temperature device integration process. In addition, device layers should be very thin so that the holes for electrical connections between layers can have a low aspect ratio. Thus, it's desirable to synthesize high-quality material for a device layer at high temperature, thin the layer, and subsequently transfer it to the stack. For silicon, this layer transfer concept has a long history[23,24] with a proven track record as a manufacturing process.[25] Emerging electronic materials such as carbon nanotubes[26] and 2D atomic crystals[27] are promising as future FET channel materials because their natural crystal structures are atomically thin due to the special bonding configuration of the constituent atoms, and their carrier transport isn't affected by the imperfections of the surfaces. These materials, which are synthesized and then transferred to a target substrate for 3D integration, are making rapid progress toward meeting future performance targets. Many of the new memories described above already use low-temperature deposited materials and are therefore commensurate with monolithic 3D integration technologies. While each additional layer incurs some yield loss, a conventional 2D chip with an equivalent area would incur similar yield loss without the benefit of much shorter 3D interconnections and without the ability to optimize fabrication processes for each device layer.

Thus, not too many years from now, a computing system could have register files and SRAM as fast first-level cache. Second-level cache could utilize high-endurance, fast-access STT-MRAM or a variant. Slower, nonvolatile, very high-density memory further out in the memory hierarchy might utilize PCM, RRAM, or CBRAM, monolithically integrated with the processor cores. RRAM and CBRAM have already demonstrated the ability to read/write at about 1 V at 10 ns speed, with more than a billion endurance cycles and good energy efficiency.

We can easily envision future energy-efficient systems consisting of a great many accelerators executing specific operations or algorithms, their interactions orchestrated to perform larger tasks, and turned on and off as needed.

Device architectures have already been demonstrated for a 3D RRAM that uses a cost-effective approach that doesn't require a lithography step for every additional layer. With a future pattern feature size of 5 nm half-pitch and 128 layers of 3D structure, 64 Tbits of relatively fast nonvolatile memory could reside on a single microprocessor chip.

## New Architectures for Computing

Today's standard "von Neumann" computer architecture was originally developed to allow a simple sequential programming model. (Language, and therefore conscious logical thought, is inherently linear and sequential. Humans aren't good at multitasking.) Focusing on the instruction set as an abstraction independent of the hardware has allowed hardware and software designers to work independently, promoted backward compatibility of code, and delivered many other benefits. However, these benefits came at a cost in computational efficiency. This is evident from the one- to three-orders-of-magnitude improvements in computational performance and energy efficiency that are routinely obtained on specific algorithms implemented on application-specific integrated circuits (ASICs), digital signal processors, and field-programmable gate arrays (FPGAs).

At a fundamental level, this is because the associated design and programming models for ASICs and FPGAs result in better mapping of the operations of algorithms to a set of physical resources used to perform those operations. A typical smartphone now contains a dozen or more specialized circuits or "accelerators" to relieve the central processing units of repetitive time- and energy-consuming tasks such as video processing. Designers of larger systems such as data servers are also incorporating such specialized "accelerators." The broad adoption of GPUs and their generalization as general-purpose GPUs (GPGPUs) has also pointed to the merits of tailoring the computing architecture for specific applications, algorithms, and dataflow. In increasingly power-constrained systems, this is a good use of the ever cheaper transistors delivered by the Moore's law development paradigm.

We can easily envision future energy-efficient systems consisting of a great many accelerators executing specific operations or algorithms, their interactions orchestrated to perform larger tasks, and turned on and off as needed. System architecture could be moving in this direction, but there's risk of complicating and encumbering the programming model. Perhaps more important, the number of different useful algorithms is large, and there seems to be little commonality of optimum hardware resources even among algorithms seen from a software perspective as closely related.[28] Because transistors will never be completely free (a daring prediction!), system architects must strive to identify the most important and broadly applicable operations and algorithms for acceleration.

One area ripe for innovation is the architecture of memory access. Computing workloads are changing. Although single-thread performance is still important, many applications are increasingly memory bound. The memory hierarchy has been optimized for applications with data locality, yet many of today's applications involve large, loosely structured datasets, requiring frequent memory access across a large address space. The data movement increases latency as well as energy consumption, particularly when data must be brought to the processor through a bus with limited bandwidth.[17] Integrating ever larger amounts of memory on chip with the processor cores will therefore continue to be a priority, and progress is likely to be sustained and accelerated by the emergence of memory devices such as STT-MRAM, PCM, and RRAM. At the cost of some increase in architectural complexity, concepts such as logic in memory (moving some computation to the cache) and memory in logic (moving some cache inside the CPU) further reduce latency and energy dissipation by reducing data movement. In the literature for many years, these concepts are now the focus of serious development efforts. Progress will be sustained and accelerated by the development of device technologies that enable monolithic integration of multiple layers of logic and memory connected by nanometer-scale, ultra-dense interlayer connections, or even putting logic and memory on the same layer.[17,18]

These developments in memory access can be seen as relatively straightforward elaborations of the conventional von Neumann architecture. The recent

emergence of deep learning algorithms[29] for important commercial applications could herald more radical changes in computer architecture. A type of neural network algorithm, inspired by ideas about the architecture and function of the brain, deep learning algorithms have, in recent years, surpassed the performance of all other known algorithms for image classification, speech recognition, and other important pattern recognition tasks.[29] However, training these algorithms on conventional computers is energy- and time-intensive. Researchers are therefore exploring the potential of GPGPUs and FPGAs to accelerate these tasks and improve energy efficiency.[30]

The commercial success of deep learning algorithms is also motivating the broader exploration of neuromorphic or other biologically- inspired architectures for computing. To improve performance and energy efficiency, some research groups are designing and testing dedicated hardware to realize various neural network models. IBM's TrueNorth chip[31] illustrates some of the design tradeoffs. TrueNorth uses digital CMOS technology to implement a particular model known as a spiking neural network. The choice of a fully digital implementation gives the designers considerable freedom to balance the goals of improved performance and energy efficiency against the need for programmability in configuring the hardware for variants of the ever evolving algorithms. However, much of the circuit area in the resulting chip is dedicated to the SRAM memory, which stores synaptic weights (the strength of the connections between computational nodes), and much of the power goes to accessing this memory.

In general, the number of synaptic weights that typically must be stored in today's neural networks ranges from tens of millions to a billion. The ability to implement larger networks would translate to better algorithmic performance on larger problems but would push the limits of available general-purpose or dedicated hardware. The continuing Moore's law trend in integration density will gradually ease this constraint, but the research community is exploring more daring approaches.

A simple two-terminal nonvolatile analog memory integrated on chip with the neural network nodes would increase the achievable integration density by at least a factor of 20 over what's possible with SRAM. This would greatly reduce the area and power required to implement a network of a given complexity, or equivalently, enable a large increase in the complexity and capabilities of the networks that can be implemented.[32] The research community is therefore exploring the use of emerging memory devices and materials to store the connection strength as an analog value in a single memory device located near each computational node of the network.[33] Memory technologies such as PCM, RRAM, and CBRAM appear to be well suited to this tight integration of memory with logic. Some still highly exploratory nanomagnetic devices also hold promise for the efficient implementation of neural network circuit architectures.[34] RRAM could also facilitate the development of neuromorphic systems, neural network architectures that continuously optimize synaptic weights (learn) while solving problems. (This approach is distinct from the "train and then solve" approach of the community focused on execution of deep learning algorithms.) This ongoing research illustrates the broad possibilities and potential rewards for co-development and co-optimization of new devices and new architectures for computing.

## Supporting Research on New Devices and Architectures

The Executive Order establishing the National Strategic Computing Initiative (NSCI) lists five strategic objectives, the first being to accelerate delivery of a capable exascale computing system (https://www.whitehouse.gov/the-press-office/2015/07/29/executive-order-creating-national-strategic-computing-initiative). This objective can be achieved based on the continued evolution of CMOS transistor technology and newer, rapidly commercializing technologies such as silicon photonics and 3D integration. However, another strategic objective of NSCI is to establish, over the next 15 years, a viable path forward for future high-performance computing systems even after the limits of current semiconductor technology are reached (the "post-Moore's law era"). In other words, longer-term research on new devices and architectures is needed now if we want to take computing beyond the exascale.

Thus the National Science Foundation (NSF), a lead NSCI agency, and the Semiconductor Research Corporation, sponsored by companies in the microelectronics industry, recently announced a daring new research program, Energy Efficient Computing: from Devices to Architectures (E2CDA; https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505212). Most of the initial research awards went to highly multidisciplinary teams, each of which is challenged with the demanding task of simultaneously exploring a new architecture for computing and developing a new hardware platform to implement that architecture. For example, one project will explore networks of coupled optical oscillators as a hardware platform

to emulate certain dynamical properties of biological neural networks. These properties are believed to be relevant to the still poorly understood ability of the brain to continuously learn while solving problems.

Exciting new application areas from self-driving cars to the Internet of Things to health informatics will demand computation with energy efficiency orders of magnitude higher than the state of the art. Even though CMOS technology advancement as measured by clock speed has stopped, and progress in planar device density appears to be slowing, new applications[35] will continue to drive innovation in computing (https://www.src.org/newsroom/rebooting-the-it-revolution.pdf). What will replace Moore's law as the metronome of technology progress? Where should the new technology investments be placed?

Based on ongoing trends in research and development, we see opportunities for dramatic advances based on the co-development and coordinated introduction of new devices, new 3D integration processes, and new architectures for computing. These opportunities are largely complementary and some are broadly multiplicative. For example, if one or more emerging low-voltage devices reduce total system power, including the power to access memory, by a factor of 20, and architectural innovations in memory access reduce total system power by a factor of 5, implementing both could conceivably reduce active power by a factor of 100. The foreseeable future will be less about shrinking the FET and more about the sequential introduction of increasingly diverse device technologies integrated in increasingly heterogeneous computer architectures optimized for performance and energy efficiency.

Longer term, a vast landscape of research opportunities remains to be explored, and the future of information technology still appears unbounded. As just one example, consider again the horizontal dotted line in Figure 1 indicating $kT$ at room temperature, and consider the current explosive growth of research in quantum computing. Quantum computing requires quantum coherent circuits, in other words, circuits that are as isolated from the rest of the world as possible and therefore as close to perfectly energy conserving as possible. Quantum computing is still far from any substantive commercial impact, and early systems will require lots of power to run the refrigeration and the necessary error correction infrastructure that protects the fragile quantum states. Still, current fundamental research provides a glimpse of the more distant possibilities. These include atomic-scale computational "devices" arranged in circuits that can maintain quantum coherence that's "good enough" even at room temperature and above. Engineers could someday think with amusement about the days when computing below the "$kT$ limit" was just a theoretical possibility. We don't know if that particular dream will become reality, but we're certain that the approaching end of the Moore's law era will mark a new beginning for information technology. ■


## Acknowledgments

HSPW is supported in part by member companies of the Stanford Non-Volatile Memory Technology Research Initiative (NMTRI), member companies of the Stanford SystemX Alliance, the National Science Foundation (CISE [Awards #1059020, #0726791, #0702343], E3S S&T Center [Award #0939514], Expeditions in Computing [Award #1317560], E2CDA [Award #1640060]), and also in part by Systems On Nanoscale Information fabriCs (SONIC) and Function Accelerated nanoMaterial Engineering (FAME), two of the six SRC STARnet Centers, sponsored by MARCO and DARPA.


## References


1. G.E. Moore, "Cramming More Components onto Integrated Circuits," *Electronics*, vol. 38, no. 8, 1965, pp. 1–14.
2. G.E. Moore, "Progress in Digital Integrated Electronics," *Proc. Technical Digest Int'l Electron Devices Meeting*, vol. 21, pp. 11–13, 1975.
3. H.D. Gilbert, ed., *Miniaturization*, Reinhold Publishing, 1961.
4. G.E. Moore, "Lithography and the Future of Moore's Law," *SPIE*, vol. 2438, 1995; http://proceedings.spiedigitallibrary.org/data/Conferences/SPIEP/54397/2_1.pdf.
5. R. Landauer, "Dissipation and Noise Immunity in Computation and Communication," *Nature*, vol. 335, Oct. 1988, pp. 779–784.
6. 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.
7. H. Sutter, "The Free Lunch Is Over: A Fundamental Turn toward Concurrency in Software," *Dr. Dobbs J.*, vol. 30, no. 3, 2005; www.gotw.ca/publications/concurrency-ddj.htm.
8. T.N. Theis and P.M. Solomon, "In Quest of the 'Next Switch': Prospects for Greatly Reduced Power Dissipation in a Successor to the Silicon Field Effect Transistor," *Proc. IEEE*, vol. 98, no. 12, 2010, pp. 2005–2014.

9. Y. Taur et al., "CMOS Scaling into the nm Regime," *Proc. IEEE*, vol. 85, no. 4, 1997, pp. 486–504.
10. D.J. Frank et al., "Device Scaling Limits of Si MOSFETs and Their Application Dependencies," *Proc. IEEE*, vol. 89, no. 3, 2001, pp. 259–288.
11. W. Haensch et al., "Silicon CMOS Devices beyond Scaling," *IBM J. Research and Development*, vol. 50, July/Sept. 2006, pp. 339–361.
12. H.-S.P. Wong, "Beyond the Conventional Transistor," *IBM J. Research and Development*, vol. 46, Mar./May 2002, pp. 133–168.
13. K.J. Kuhn, "Considerations for Ultimate CMOS Scaling," *IEEE Trans. Electronic Devices*, vol. 59, no. 7, 2012, pp. 1813–1828.
14. H. Esmaeilzadeh et al., "Dark Silicon and the End of Multicore Scaling," *IEEE Micro*, May/June 2012, pp. 122–134.
15. S. Borkar and A. A. Chien, "The Future of Microprocessors," *Comm. ACM*, vol. 54, no. 5, 2011, pp. 67–77.
16. T.-J. King Liu and K. Kuhn, eds., *CMOS and Beyond: Logic Switches for Terascale Integrated Circuits*, Cambridge Univ. Press, 2015.
17. M.M.S. Aly et al., "Energy-Efficient Abundant-Data Computing: The N3XT 1,000X," *Computer*, Dec. 2015, pp. 24–33.
18. H.-S.P. Wong and S. Salahuddin, "Memory Leads the Way to Better Computing," *Nature Nanotechnology*, vol. 10, no. 3, 2015, pp. 191–194.
19. S. Mueller et al., "Incipient Ferroelectricity in Al-Doped HfO2 Thin Films," *Advanced Functional Materials*, vol. 22, no. 11, 2012, pp. 2412–2417.
20. H.-S.P. Wong et al., *Stanford Memory Trends*, tech. report, Stanford Univ., 2016; https://nano.stanford.edu/stanford-memory-trends.
21. B. Behin-Aein et al., "Proposal for an All-Spin Logic Device with Built-In Memory," *Nature Nanotechnology*, vol. 5, 2010, pp. 266–270.
22. S. Fusil et al., "Magnetoelectric Devices for Spintronics," *Ann. Rev. Materials Research*, vol. 44, July 2014, pp. 91–116.
23. H.-Y. Chen et al., "HfOx Based Vertical RRAM for Cost-Effective 3D Cross-Point Architecture without Cell Selector," *Proc. IEEE Int'l Electron Devices Meeting*, 2012, pp. 497–500.
24. M. Bruel, "The History, Physics, and Applications of the Smart-Cut Process," *MRS Bulletin*, vol. 23, no. 12, 1998, pp. 35–39.
25. L. Clavelier et al., "Engineered Substrates for Future More Moore and More than Moore Integrated Devices," *Proc. Int'l Electron Devices Meeting*, 2010, pp. 2–6.
26. M. Shulaker et al., "Carbon Nanotube Computer," *Nature*, vol. 501, Sept. 2013, pp. 256–530.
27. K.S. Novoselov et al., "Two-Dimensional Atomic Crystals," *Proc. Nat'l Academy of Sciences*, vol. 102, no. 30, 2005, pp. 10451–10453.
28. V.C. Cabezas and P. Stanley-Marbell, "Parallelism and Data Movement Characterization of Contemporary Application Classes," *Proc. 23rd Ann. ACM Symp. Parallelism in Algorithms and Architectures*, 2011, pp. 95–104.
29. Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, May 2016, pp. 436–444.
30. K. Ovtcharov et al., "Accelerating Deep Convolutional Neural Networks Using Specialized Hardware," Microsoft, 2015; www.microsoft.com/en-us/research/publication/accelerating-deep-convolutional-neural-networks-using-specialized-hardware.
31. P.A. Merolla et al., "A Million Spiking-Neuron Integrated Circuit with a Scalable Communication Network and Interface," *Science*, vol. 345, no. 6197, 2014, pp. 668–673.
32. S.B. Eryilmaz et al., "Device and System Level Design Considerations for Analog-Non-Volatile-Memory Based Neuromorphic Architectures," *Proc. IEEE Int'l Electron Devices Meeting*, 2015, pp. 64–67.
33. D. Kuzum, S. Yu, and H.-S.P. Wong, "Synaptic Electronics: Materials, Devices and Applications," *Nanotechnology*, vol. 24, no. 38, 2013, p. 38.
34. D. Fan et al., "Hierarchical Temporal Memory Based on Spin-Neurons and Resistive Memory for Energy-Efficient Brain-Inspired Computing," *IEEE Trans. Neural Networks and Learning Systems*, Aug. 2015; doi:10.1109/TNNLS.2015.2462731.
35. L. Whitman, R. Bryant, and T. Kalil, "A Nanotechnology-Inspired Grand Challenge for Future Computing," blog, 25 Oct. 2015; https://www.whitehouse.gov/blog/2015/10/15/nanotechnology-inspired-grand-challenge-future-computing.



**Thomas Theis** is a professor of electrical engineering at Columbia University and executive director of the Columbia Nano Initiative. His research interests include emerging devices and computer architectures for more energy-efficient computing. Theis received a PhD in physics from Brown University. He's an IEEE Fellow and a Fellow of the American Physical Society. Contact him at tnt2122@columbia.edu.

**H.-S. Philip Wong** is the Willard R. and Inez Kerr Bell Professor in the School of Engineering at Stanford University. His research aims at translating discoveries in science into practical technologies. His works have contributed to advancements in nanoscale science and technology, semiconductor technology, solid state devices, and electronic imaging. Wong received a PhD in electrical engineering from Lehigh University. He's an IEEE Fellow. Contact him at hspwong@stanford.edu.