

Lawrence Berkeley National Laboratory

LBL Publications

Title

Computing beyond Moore's Law

Permalink

<https://escholarship.org/uc/item/9t3495n0>

Journal

Computer, 48(12)

ISSN

0018-9162

Authors

Shalf, John M
Leland, Robert

Publication Date

2015-12-01

DOI

10.1109/mc.2015.374

Peer reviewed



Computing beyond Moore's Law

John M. Shalf, Lawrence Berkeley National Laboratory

Robert Leland, Sandia National Laboratories

Photolithography systems are on pace to reach atomic scale by the mid-2020s, necessitating alternatives to continue realizing faster, more predictable, and cheaper computing performance. If the end of Moore's law is real, a research agenda is needed to assess the viability of novel semiconductor technologies and navigate the ensuing challenges.

In 1965, Gordon Moore famously observed that the number of components on an integrated circuit (IC) had doubled every year on average since the introduction of this technology in 1959.¹ He predicted that this trend, driven by economic considerations of cost and yield, would continue for at least a decade, although later the integration pace was moderated to doubling approximately every 18 months. He also noted that “shrinking the dimensions on an integrated structure makes it possible to operate the structure at higher speed for the same power per unit area”—an innovation that Robert Dennard of IBM formalized nearly a decade later as Dennard scaling, the ability to reduce device operating voltages and scale clock frequencies exponentially each generation.²

This mutually reinforcing scaling of feature size, frequency, and power meant that chip functionality would improve exponentially with time at a roughly constant

cost per generation, and Moore predicted this improvement, in turn, would lead to a cornucopia of societal benefits that would flow from semiconductor microelectronics technology. The serendipitous scaling effects Moore predicted did indeed persist, lasting 40 years longer than he predicted. However, Dennard scaling came to an end in 2004, which led to a power-efficiency crisis for CMOS logic and which poses an even more fundamental challenge for traditional technology scaling in the mid-2020s.

Within that decade, the magical growth process Moore described will come to an end as 2D lithography capability approaches the atomic realm. The end of conventional scaling will impact all computing technologies that depend on improvements in cost, energy efficiency, and storage capacity—from large-scale systems to the smallest consumer electronic devices.

The limits of existing semiconductor microelectronic technology at the device level and their impact at the system level demand a successor technology to the currently ubiquitous CMOS logic. There is not yet an obvious successor, but we see three basic paths to obtaining one, shown in Figure 1: create new devices, build new architectures with or without new devices, and develop new computational paradigms. We expect to see substantial exploration and innovation in each of these areas. New computation models will likely depart from digital computing and expand into new areas, where former technology paradigms are less suitable. New architectures and packaging will resourcefully arrange existing building blocks, improving performance irrespective of the underlying technology. Finally, new materials and transistors will enhance performance by creating more efficient underlying logic devices.

In the near term, emphasis will likely be on developing CMOS-based devices that extend into the third, or vertical, dimension and on improving materials technology. These efforts will likely coevolve with new architectural approaches that better tailor computing capability to specific problems, driven principally by large economic forces associated with the \$4 trillion-per-year global IT market.

In the longer term, we expect a transition toward new device classes and the emergence of practical systems based on novel computing approaches. To effectively meet societal needs and expectations in a broad context, these new devices and computing paradigms must be economically manufacturable at scale and provide an exponential improvement path. Such requirements could necessitate

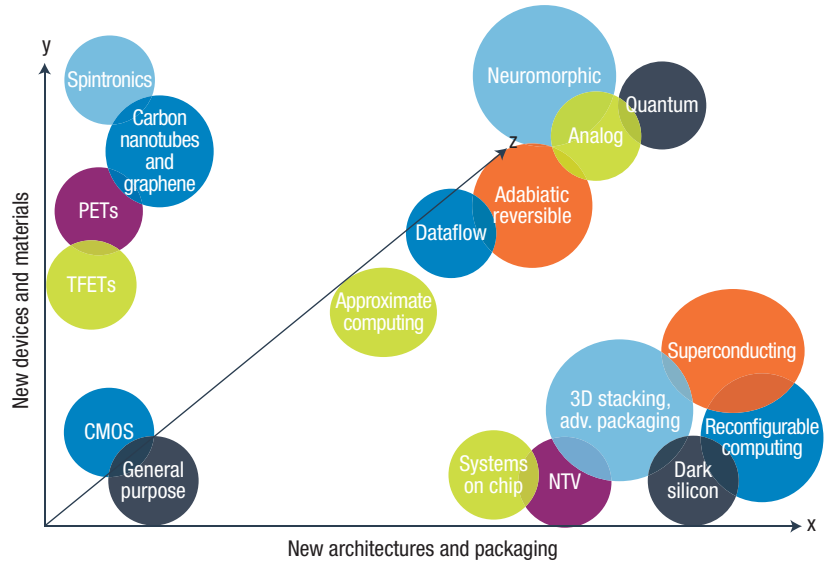


FIGURE 1. Technology scaling options along three dimensions. The graph’s origin represents current general-purpose CMOS technology, from which scaling must continue. All the dimensions, which are not mutually exclusive, aim to squeeze out more computing performance. PETs: piezo-electric transistors, TFETs: tunneling field-effect transistors; NTV: near-threshold voltage.

a substantial technological shift analogous to the transition from vacuum tubes to semiconductors.

This transition will require not years, but decades, of effort, so whether the semiconductor roadmap has 10 or 20 years of remaining vitality, researchers must begin now to lay a strategic foundation for change.

IS IT REALLY THE END?

Far from a physical law, Moore’s observation is an economic theory driven by technology scaling—constantly improving the photolithography processes that shrink on-chip component size. For the past 50 years (as of 2015), multiple assaults on conventional technology scaling for digital electronics have challenged Moore’s observations about performance improvement. As the sidebar “Moore’s Law Resilience” describes, despite the limitations of numerous underlying physical mechanisms, new approaches have materialized to continue Moore’s scaling. One researcher famously quipped, “I predict Moore’s law will never end—that way, I will only be wrong once!”

Why then should things be different this time?

Limits of 2D lithography

If technology scaling is indeed the underlying driver of progress, 2D silicon photolithography is central to that progress, and there is much concern that 2D scaling of photolithography will approach fundamental limits by the end of 2020. Moreover, there is no obvious successor technology. A silicon atom is approximately half a nanometer (nm) in diameter in semiconductor material. At the current rate of improvement, photolithography systems will be able to use 5-nm technology to create transistor features on the scale of handfuls of atoms by 2022 to 2024.³

This feature size corresponds to a dozen or fewer silicon atoms across critical device features, which means that the technology will be a practical limit for controlling charge in a classical sense. To go further would require engineering these devices in a regime in which quantum-mechanical effects will dominate, such as tunneling electrons through the gate oxide, which

MOORE'S LAW RESILIENCE

A brief history of threats to Moore's law and the computing community's responses gives a flavor of this theory's staying power. In the 1980s, scaling the power density of bipolar transistor logic (the basis for most digital logic) became impractical. The inability to continue to scale bipolar logic led to a wholesale transition to more efficient CMOS logic technology, which enabled another two decades of technology scaling.

In 2004, Dennard scaling began to fail, and the computing industry was again in a power density crisis much like the bipolar crisis. Computer architects responded to the lack of clock-frequency growth per processing core with multicore technology—exponentially scaling the number of per-chip cores—which enabled technology scaling into 2014.

In 2015, the energy efficiency of logic gates continues to scale, albeit more slowly, but the data-carrying capacity and efficiency of wire are not improving at the same pace. Moving from the current computation-centric paradigm to data-centric programming could forestall this trend, but could also be a shift that Moore's law might not withstand. On the other hand, Moore's law is not a physical law, but an economic theory that describes a powerful market force. If a path exists to continued computing technology improvements, Moore's law will be the motivation to find it.

risks current leakage and increased energy loss. Although it is feasible to reach 3 to 5 nm by 2022 with extreme ultraviolet (EUV) technology, the adoption rate of smaller feature sizes is driven more by economics and return on investment through performance improvements than by technological feasibility. The rapidly increasing cost of lithography methods could make the production of smaller feature sizes economically impractical.

In this sense, the end of Moore's law is really the end of useful 2D lithography scaling. Constraints imposed by fundamental device physics and the increased manufacturing costs of producing smaller transistors are looming limits. We are not certain which limit will ultimately end Moore's law, but clearly further improvements in planar circuit density will become implausible because of one of these,

and computing capability will no longer be able to scale through the basic approach that has worked so well for so long.

Data-movement energy cost

The end of Moore's law will affect all devices—both processing and storage—that depend on shrinking feature size to make progress.⁴ Increasing circuit or storage density will require a technology that supports signal gain and reduces the energy that data movement consumes. The intrinsic resistance of interconnect material will limit any solution involving electrons. The principal problem is that, because the metal used to conduct the electrons representing the bit has resistance and capacitance, the energy consumed to transmit a bit is proportional to the distance it must travel.⁵ Copper is as good a conductor as can

be expected for a common material at room temperature. Regardless, data movement will still dominate energy losses and restrict the ability to build the circuit out vertically, moving it from 2D to 3D.

NEW COMPUTING MODELS

The end of Moore's law will pose packaging and performance challenges for all manner of consumer electronic devices that depend on improvements in cost and energy efficiency to squeeze more functionality out of a device with a limited battery capacity or power supply. The ability to make a smartphone smarter will be compromised if the industry cannot pack more functionality into less space.

The deeper issue is the threat to the US computing industry's growth. Moore's law scaling turned computing into a pervasive consumer technology that has become increasingly more powerful within a market that has grown exponentially. An end to that scaling could slow the pace of product improvements, which could have a significant negative economic impact.

These challenges are prompting researchers to take a broader view of what constitutes computation. An *Initial Look at Alternative Computing Technologies for the Intelligence Community*, a recent report commissioned by the Intelligence Advanced Research Projects Activity (IARPA), proposes looking at four basic computational models:⁶

- › classical digital computing (CDC), which includes all the binary digital electronics that form the basis for the computing and consumer electronics industries;
- › analog computing (AC), which includes nonbinary devices

that implement computation through direct physical principles;

- › neuro-inspired computing (NC), which includes devices based on the principles of brain operation and general neuronal computation; and
- › quantum computing (QC), which could in theory be used to solve some problems with combinatorial complexity through the selection of a desired state from a superposition of all possible answers to a problem.

The authors also underline the importance of distinguishing between new paradigms for computation and new technological implementations of existing paradigms, making several key observations. One is that AC can be simpler than some digital approximation, but does not lend itself to general-purpose computing because the device is specialized for computation. The computational precision is problematic to maintain and can be sensitive to its environment.

Another observation is that digital computers are good at deterministic/algorithmic calculation, but poor at simple reasoning and recognition. NC devices have proven inherently resilient and very good at problems that CDC is not. Many unexplored opportunities exist for such computational models, but much is still not understood about how the brain actually computes.

Finally, the authors note that quantum information processing theoretically could enable the efficient solution of some combinatorial and NP-hard problems (problems not solvable in polynomial time using digital computation) or could be used to simulate the

electronic state of complex molecules, as Richard Feynman proposed. However, QC is not a suitable replacement for CDC in domains where CDC excels.

These technology options create the possibility of approaches that go well beyond what CMOS and digital electronics technologies have traditionally performed effectively. However, we do not believe that they are suitable as replacements for digital electronics in tasks that digital computing already performs well. For that reason, we choose to focus on new technological implementations of the CDC model because we view it as the most immediately relevant to a broad set of societal concerns associated with the end of Moore's law.

EVALUATING CDC CANDIDATES

In the past, a competitor to CMOS or CDC would need to keep pace with a relentless improvement schedule in which CMOS technology doubled its performance every 18 months or so and leveraged tremendous economies of scale. This combination proved unbeatable except in relatively narrow niches. With the end of CMOS technology scaling, these competitive conditions have changed. A come-from-behind competitor to CMOS is not yet apparent, but metrics are in place to assess the fitness of potential CMOS replacements. Shekhar Borkar of Intel has developed three metrics—gain, signal-to-noise immunity, and scalability⁷—to which we have added a fourth—scalable manufacturability:

- › *gain*—the energy required to switch the device state from on to off must be less than the energy the device controls;
- › *signal-to-noise immunity*—the

signal must be far enough above the background noise level to enable detection;

- › *scalability*—the technology must allow density increases and corresponding energy reductions as it improves; and
- › *scalable manufacturability*—the technology must be producible with a process capable of industrial-scale implementation.

Although we did not assess potential post-CMOS technologies in detail, we used these four metrics and IARPA's criteria of timescale, complexity, risk and opportunity to evaluate their merit. The results are shown in Table 1.

The list of options in the table is by no means comprehensive, but is meant as a glimpse of those most commonly debated in and outside the literature. No option is clearly superior in all respects, so we believe that one or more of them will reach mainstream use through integration with conventional silicon and CMOS platforms. Indeed, chip stacking is already enabling the stacking of photonics technology directly on conventional silicon logic and memory circuits.

Packaging and computer architecture do not require fundamentally new materials and underlying process technology, which can extend the same underlying silicon/CMOS technology. New devices, on the other hand, require fundamentally new materials and even new data and computational representations—a far deeper and less predictable revision of the digital computing paradigm.

Table 1. Summary of technology options for extending digital electronics.

TABLE 1. Summary of technology options for extending digital electronics.

Improvement Class	Technology	Timescale	Complexity	Risk	Opportunity
Architecture and software advances	Advanced energy management	Near-Term	Medium	Low	Low
	Advanced circuit design	Near-Term	High	Low	Medium
	System-on-chip specialization	Near-Term	Low	Low	Medium
	Logic specialization/dark silicon	Mid-Term	High	High	High
	Near threshold voltage (NTV) operation	Near-Term	Medium	High	High
3D integration and packaging	Chip stacking in 3D using thru-silicon vias (TSVs)	Near-Term	Medium	Low	Medium
	Metal layers	Mid-Term	Medium	Medium	Medium
	Active layers (epitaxial or other)	Mid-Term	High	Medium	High
Resistance reduction	Superconductors	Far-Term	High	Medium	High
	Crystalline metals	Far-Term	Unknown	Low	Medium
Millivolt switches (a better transistor)	Tunnel field-effect transistors (TFETs)	Mid-Term	Medium	Medium	High
	Heterogeneous semiconductors/strained silicon	Mid-Term	Medium	Medium	Medium
	Carbon nanotubes and graphene	Far-Term	High	High	High
	Piezo-electric transistors (PFETs)	Far-Term	High	High	High
Beyond transistors (new logic paradigms)	Spintronics	Far-Term	Medium	High	High
	Topological insulators	Far-Term	Medium	High	High
	Nanophotonics	Near/Far-Term	Medium	Medium	High
	Biological and chemical computing	Far-Term	High	High	High

ARCHITECTURE AND SOFTWARE ADVANCES

Architectural schemes to extend digital computing aim to better manage energy, decrease power consumption, lower overall chip cost, and improve error detection and response.

Energy management

Current energy-management technologies are ubiquitous and typically coarse grained. Dynamic voltage and frequency scaling (DVFS) and thermal throttling lower both clock frequencies and voltages when computing demands do not require peak performance. Coarse-grained DVFS can save significant power in current consumer electronics devices, which are mostly idle. However, it only marginally benefits devices that operate near 100 percent utilization. Finer-grained

power management might provide additional potential to recover energy, enabling faster transitions between power states by having the software direct state changes.

Circuit design

Studies have demonstrated approaches that enable wires to operate at a lower voltage for long-haul connections and then reamplify efficiently at the endpoints, although with some loss from reamplification. A recent NVIDIA paper estimated an opportunity for two to three times improvement using such advanced circuit design techniques with current technologies.⁸

A more aggressive path to performance enhancement is clockless (or domino logic) design. Clock distribution consumes a large fraction of system power, and constricts a circuit

to the operation speed of its slowest component. Practical and effective clockless designs have proven elusive, but recent examples show that this approach could be a viable way to lower dynamic power consumption for both neuromorphic and digital applications.⁹

System-on-chip (SoC) specialization

The core precept of SoC technology is that chip cost is dominated by component design and verification costs. Therefore, tailoring chips to include only the circuit components of value to the application is more economically efficient than designing one chip that serves a broad application range—the current commodity design practice. This tailoring strategy is common practice for cell-phone chips, such as

that in the Apple iPhone, which combines commodity embedded processor cores in a specialized SoC design, but is only just being applied in server and high-performance computing (HPC) chips.

Logic specialization

Field-programmable gate arrays (FPGAs) and reconfigurable computing hold promise for improving performance by creating tailored circuits for each problem, but they are not efficient to implement. In a typical FPGA implementation, most of the available reconfigurable wires remain unused to maximize the use of lookup tables.¹⁰ A custom application-specific integrated circuit (ASIC) design improves performance by 10 times over the FPGA design of the same circuit because the ASIC design eliminates redundant wiring.

Unfortunately, tailoring in either case requires substantial hardware design expertise, and circuit design in general is much more expensive than software design. Possibly, the economic disincentive of designing and verifying custom circuits will be overcome by the reality of having no performance scaling at all.

The most extreme proposals for customizing logic are intended for use in *dark silicon*—areas on the ASIC that remain turned off when not in use. The idea is to trade off increased ASIC surface area for more efficient specialized circuits with the aim of having performance gains offset the cost of extra area. By turning off specialized circuits that are not required, this approach is energy neutral (in the sense that it increases performance without increasing power consumption), and it is already being used in some specialized consumer electronics applications. However, its

utility in general-purpose computation is unproven.

Near-threshold voltage operation

The mainstream computing community has traditionally shunned further reductions in device voltage because such measures would reduce transistors' signal-to-noise immunity and subject circuits to wider statistical performance variation. Both effects would result in the unreliable performance of individual circuits and present daunting problems to software and hardware development.

From a software standpoint, conventional bulk-synchronous approaches to scaling parallel computing performance would become untenable, forcing a move to entirely asynchronous software-execution models and a corresponding reformulation of algorithms and infrastructure. Applications and algorithm developers would need to substantially rewrite software to accommodate this unpredictable performance heterogeneity.

In hardware, increased unreliability would require more pervasive error detection and corresponding software infrastructure to respond—the cost of this is unknown. Clock frequencies would be substantially lower, putting more pressure on parallelism to gain performance improvements, already a daunting software burden.

Near-threshold voltage (NTV) circuit operation provides the opportunity to reduce operating voltages and hence increase device energy efficiency (along with its usable performance and scalability) by an order of magnitude. NTV is still an active research focus, with efforts to determine whether the software challenges posed by reliability, performance heterogeneity, and increased parallelism

will detract from raw potential performance improvement.¹¹

3D INTEGRATION AND PACKAGING

3D integration and packaging has been used successfully in mainstream devices to increase logic density and reduce data-movement distances. Most memory devices involve some form of chip stacking, which will be critical in increasing the density of future devices.

The primary challenges to scaling 3D lithographic layering are improving defect tolerance and managing the thermal densities and intrinsic resistance. Stacking cool technologies such as emerging nonvolatile memory cells (magnetoresistive RAM, memristors, and so on) provides substantial opportunity for deeper lithographic layering and potentially a few orders of magnitude improvement in component density in terms of both increased functionality and memory capacity. Although 3D stacking will substantially reduce data-movement requirements—a major contributor to thermal density—it is unclear how much additional room it affords for deeply stacking logic layers.

3D memory technologies will pave the way to 3D integration. Technologies that reduce operating voltages for digital circuits (a development effort that has been stalled since 2004), could provide further room to build circuits out vertically.

Stacking with through-silicon vias

In this approach, holes are drilled through silicon chips to provide electrical connections between the stacked layers. Chip stacks of up to eight layers are already available, and engineering costs are lower relative to adding

layers lithographically or through epitaxial deposition. Through-silicon vias (TSVs) offer much higher bandwidth and energy efficiency than conventional chip packages, such as ball grid arrays and other pin packages. However, relative to adding chip layers with photolithography, TSVs do not offer as much bandwidth, efficiency, and connectivity between layers.

Metal layers

CMOS has traditionally been built out in 2D planar form with modest improvements in 3D. Modern chips have up to 11 metal layers. The number of metal layers could be improved, but these provide additional connectivity among components only on the 2D surface.

Epitaxial deposition

Lithographic layering yields only the bottom silicon layer, which is still 2D planar. More active transistors require adding layers of semiconductor material on top of each other. Epitaxial deposition meets that requirement through a chemical, molecular beam or vapor deposition process. Challenges remain in depositing high-quality, single-crystal active layers, but there is substantial progress in studying approaches that go beyond standard silicon, such as those that use processes other than epitaxial deposition to directly transfer very thin layers of bulk crystalline material.

RESISTANCE REDUCTION

Most ICs use a copper-based interconnect to reduce resistance because copper is a particularly good conductor, and at room temperature few options are capable of lower electrical resistance. Two alternatives are superconductors and crystalline metals.

Superconductors

Superconducting could be a way to advance HPC system performance, but it will force a departure from the mainstream because cooling is not likely to be practical for consumer devices. Even cuprate-based high-temperature superconductors have impractical cooling and magnetic shielding requirements for such devices. The viability of cryogenically cooled electronics in standard phones or laptops is doubtful.

Using cryogenically cooled electronics to extend HPC performance is technically feasible, but would entail a departure from the traditional leveraging of commodity component technology. There could be severe repercussions for the US in HPC competitiveness and system affordability, which critically depend on that leveraging ability.

Crystalline metals

Although copper is an excellent conductor, in a typical polycrystalline configuration, electrons still scatter off the boundaries between neighboring crystalline grains. Metal layers' conductivity could be improved by as much as five times by creating larger grain sizes. Techniques to create larger crystal grains in a scalable chip-manufacturing process are still not well understood or perhaps are not being shared because of proprietary concerns.

MILLIVOLT SWITCHES

Millivolt switches are essentially transistors that can operate at much lower voltages. Many 3D stacking approaches eventually fail to scale because stacking energy-intensive logic layers creates energy-density limits. Any future electronic system will

need material that reduces switching power for the logic and resistive losses from information transfer within each constituent logic layer. Examples of structures and materials that might improve device performance are tunneling field-effect transistors (TFETs), heterogeneous semiconductors, carbon nanotubes and graphene, and piezo-electric transducers (PETs).

Tunneling field-effect transistors

With conventional FETs, device performance is limited by the voltage swing required to turn them completely on or off (gain). A TFET uses a channel material that modulates the quantum tunneling effect, rather than the classical metal-oxide semiconductor (MOS) FET modulation of thermionic emission, which creates a switch that is more sensitive to gate voltage when turning on or off and can thus operate at a lower voltage. Because the device's power dissipation is proportional to voltage squared, there is substantial opportunity to improve energy efficiency.

Different materials systems are being investigated, but thermal sensitivity, speed, obstacles to reliable manufacturability, and other scalability issues challenge current devices. Without lithography improvements, successful development of TFET devices could enable one or two additional generations of improvement in technology performance scaling, but it will take a decade to translate laboratory advances to mainstream mass production.

Other technologies involve new gate designs to improve transistor sensitivity, such as ferroelectric gate FETs. All have similar challenges in manufacturing, and offer similar opportunities to extend technology energy efficiency (and hence performance) through lower operating voltages.

Heterogeneous semiconductors

Silicon has become the primary semiconductor material for ICs because of its favorable chemical properties and physical robustness. Semiconductors formed from III-V materials (so named for their source in columns 3 and 5 of the periodic table), such as gallium arsenide (GaAs), provide much higher performance, but are more susceptible to cracking from low-quality oxides and consequently require more involved chemical processing. These manufacturability problems have kept III-V materials on the margins of mainstream digital electronics.

Recent dramatic improvements in the ability to integrate islands of III-V materials into bulk silicon substrates address these problems by enabling a marriage of silicon's manufacturing, chemical, and electrical benefits to the performance benefits of embedded III-V materials. Heterogeneity is achieved through silicon straining, a process that alters the silicon substrate so that its atomic spacing aligns with that of the III-V material and dopes the III-V materials with additional impurities so that the atomic spacing aligns with that of silicon when it is vapor-deposited onto the silicon substrate. Alternatively, the silicon can be deposited on top of a substrate material with slightly larger lattice spacing, such as silicon germanium (SiGe). The atomic bonding between the layers stretches out the silicon lattice and can substantially improve charge-carrier mobility through the silicon.

Challenges in this area are due mostly to the cost and complexity of producing crystallographically perfect structures in a manner that integrates into a scaled up production lithography process. However, recent advances in ultra high vacuum

chemical vapor deposition (UHV-CVD), molecular beam epitaxy (MBE) and other epitaxial growth techniques have made reliable large-scale production more practical. Strained silicon is a novel processing approach that, although promising, is still finding its way into mainstream lithography processes, and heterogeneous semiconductors face obstacles stemming from the required materials. For example, gallium arsenide suffers from unbalanced P- and N-gate performance, which in turn affects its efficiency in CMOS devices. Combining silicon using epitaxial deposition could solve some of these problems and provide an order-of-magnitude improvement in some device functions. However, to date, many III-V materials are not candidates for an exact CMOS replacement.

Carbon nanotubes and graphene

The band gap of carbon nanotubes is much smaller than that of silicon, a characteristic that translates to less energy in operating carbon nanotube-based devices. These devices also present lower resistance to electron movement, which increases noise susceptibility. Experiments have shown that transistors based on carbon nanotubes can deliver higher current densities than silicon-based devices,¹² which in principle would enable them to operate at much higher switching rates and energy efficiency. Other studies show that nanotube devices have gain- (steeper subthreshold slope) and noise-rejection properties that can compete with those of classical semiconductors for individual devices.¹³

Despite these favorable properties, mundane issues like contact resistance are stalling progress in carbon nanotubes, and gate dielectric

materials have yet to be fully engineered and optimized for nanotubes. Furthermore, nanotube diameter and band-gap distribution leads to problematic device variation, making high-purity nanotubes with uniform diameter hard to manufacture. The primary challenge for nanotubes lies in finding a scalable manufacturing process, as current devices require precise tube placement to form transistors and circuits. Although recent advances in self-assembly processes for nanotube-based circuits have been significant,¹⁴ a competitive, commercially scalable process is still a long way off.

Graphene is a planar matrix of carbon atoms with no band gap, making it unsuitable for digital switches that turn off and have very low current leakage. One way around this is to fashion graphene into very narrow ribbons, as graphene rolled into a perfectly smooth tube is in effect a nanotube.

The challenge is to manufacture uniformly wide graphene nanoribbons with atomically smooth edges, and generally, graphene nanoribbons are less well developed than nanotubes. However, breakthrough synthesis techniques are emerging that might more efficiently and economically produce pure, uniform ribbon width than the techniques used to produce nanotubes. In addition to graphene, there has been rapid development of other 2D materials systems such as molybdenum disulfide (MoS₂) and phosphorene.¹⁵

Piezo-electric transistors

Piezo-electric transistors (PETs) use the piezo-electric effect, in which an electric field induces mechanical stress by changing material size. The

most common use of PETs has been in micromechanical systems and force sensors, but if such piezo materials can be successfully miniaturized, the technology could also be used to form an extremely fast (multi-gigahertz) microscale electronic relay.¹⁶ This strategy is one of many micromechanical approaches to developing high-performance switches.

BEYOND TRANSISTORS: NEW LOGIC PARADIGMS

The devices and techniques described so far aim to improve device performance with familiar digital computing architectures and computational models. However, technologies are being proposed that are more than just better transistors; they change how bits are stored and transformed. These radical performance-enhancement paths represent new logic paradigms, including spintronics, topological insulators, nanophotonics, and biological and chemical computing.

Spintronics

Computation on information and its communication through the manipulation of magnetic domains takes less energy than moving electrons to such a degree that it is nearly inconsequential to overall power consumption. Kevin Cummings of SEMATECH, a global consortium of semiconductor device, equipment and materials manufacturers, stated in an email to us that spin materials could benefit technologies aiming to provide dual functionality (logic and memory) as well as new circuit designs, such as static RAM.

For applications involving memory technologies, there is little impact on standard paradigms for computation, but broader use of spintronic devices

as general-purpose computing applications (fully replacing CMOS) would require an adiabatic or a reversible computing model. Such models can be highly restrictive and would fundamentally disrupt the current digital computing model.

Topological insulators

Topological insulators confine energy to a 2D space. Relative to conventional wires, these confined energy states can provide more efficient (higher noise margin) information transport and storage, but the proper approach to implementing logic is uncertain. One method is to apply 2D image-analysis algorithms that use a photogalvanic effect to program initial state for quantum bits (qubits) embedded in the topological insulator. According to ATECH's Cummings, the electronics industry is considering these 2D semiconductors as well as other new semiconductors with unique properties.

Nanophotonics

Photonic technology has obvious advantages for scalable communications, although using nanophotonics at subwavelength scale as a replacement for computing and transistor technologies is problematic because of scale incompatibilities: available optical transistors have a low gain, and optical wavelengths are large compared to current realizable photolithographic scales.

Unlike standard electrical wires, which have a strong distance-dependent energy cost, photonics' energy costs are nearly independent of data distance, allowing them to overcome the wire-resistance limitation. Unfortunately, although it has steadily decreased, the energy cost to activate

the laser to send information over a photonic connection is still far higher than the wire cost.¹⁷ Even so, photonics will be essential in overcoming wire limits and the disparity in on-chip and off-chip communications costs.⁵ An effective high-gain optical transistor would make nanophotonics a competitor in CMOS replacement, but the technology for a high-performance, optically controlled switch requires further development.

Biological and chemical computing

Computing devices based on the animal brain aim to emulate the most complex machine known. The principal challenges to biologically based computing devices include low gain, a poor signal-to-noise ratio, and exotic operating conditions.


The search continues for a chemical switching mechanism that offers sufficient gain and noise rejection to compete with silicon. Good candidates exist, but in addition to requiring complex operating environments, they are difficult to scale.

Society relies heavily on the benefits that Moore's law provides—cheap technology that continues to scale almost effortlessly. From this point, the energy cost of data movement will dominate both technical and economic issues because the energy cost to compute data is decreasing faster than the cost to move it to computing operations. Increasing the use of parallelism in software is a short-term fix that will require massive commercial effort. Longer term, the current computation-centric model might need to give way to a data-centric model.

ABOUT THE AUTHORS

ROBERT LELAND is vice president of science and technology and chief technical officer (CTO) of Sandia National Laboratories. His research interests include parallel algorithm development, sparse iterative methods, and applied graph theory. Leland received a PhD in parallel computing from Oxford University. Contact him at leland@sandia.gov.

JOHN M. SHALF is CTO of the National Energy Research Supercomputing Center and head of the Computer Science Department at Lawrence Berkeley National Laboratory. His research interests include parallel computing software and high-performance computing technology. Shalf received a MS in electrical and computer engineering from and Virginia Tech. He is a member of the American Association for the Advancement of Science, IEEE, and the Optical Society of America, and coauthor of the whitepaper “The Landscape of Parallel Computing Research: A View from Berkeley” (UC Berkeley, 2006). Contact him at jshalf@lbl.gov.

Evolving technology in the Moore’s law vacuum will require an investment now in basic sciences, including materials science, to study candidate replacement materials and alternative device physics to foster continued technology scaling. Using the history of the silicon FinFET, it takes about 10 years for an advance in basic device physics to reach mainstream use. Any new technology will require a long lead time and sustained R&D of one to two decades. Options abound, the race outcome is undecided, and the prize is invaluable. The winner not only will influence chip technology, but will define a new direction for the entire computing industry. 

ACKNOWLEDGMENTS

We thank Shekhar Borkar, Justin Rattner, Steve Pawlowski, and Al Gara of Intel; Catherine Jenkins and Jeff Bokor of Lawrence Berkeley National Laboratory/UC Berkeley; Erik Debenedictis of Sandia National Laboratories; Thomas Theis of SRI; and Kevin Cummings of SEMATECH for their helpful input.

REFERENCES

1. G.E. Moore, “Cramming More Components onto Integrated Circuits,” *Electronics*, vol. 38, no. 8, 1965, pp. 114–117.
2. R.H. Dennard et al., “Design of Ion-Implanted MOSFETs with Very Small Physical Dimensions,” *IEEE J. Solid-State Circuits*, vol. SC-9, no. 5, 1974, pp. 256–268.
3. R. Colwell, “The Chip Design Game at the End of Moore’s Law,” *Proc. IEEE/ACM Symp. High-Performance Chips (HC25)*, 2013; www.hotchips.org/wp-content/uploads/hc_archives/hc25/HC25.15-keynote1-Chipdesign-epub/HC25.26.190

- Keynote1-ChipDesignGame-Colwell-DARPA.pdf.
4. R.E. Fontana, S.R. Hetzler, and G. Decad, “Technology Roadmap Comparisons for TAPE, HDD, and NAND Flash: Implications for Data Storage Applications,” *IEEE Trans. Magnetics*, vol. 48, no. 5, 2012, pp. 1692–1696.
5. D.A.B. Miller, “Device Requirements for Optical Interconnects to Silicon Chips,” *Proc. IEEE*, vol. 97, no. 7, 2009, pp. 1166–1185.
6. L. Joneckis, D. Koester, and J. Alspector, *An Initial Look at Alternative Computing Technologies for the Intelligence Community*, tech. report, Inst. for Defense Analysis, Jan. 2014; <http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA610103>.
7. S. Borkar, “Electronics Beyond Nano-Scale CMOS,” *Proc. 43rd Ann. ACM Design Automation Conf. (DAC 06)*, 2006, pp. 807–808.
8. O. Villa et al., “Scaling the Power Wall: A Path to Exascale,” *Proc. IEEE Supercomputing Conf.*, 2014, pp. 830–841.
9. N. Imam et al., “Neural Spiking Dynamics in Asynchronous Digital Circuits,” *Proc. Int’l Joint Conf. Neural Networks (IJCNN 13)*, 2013, pp. 1–8.
10. A. DeHon, *Reconfigurable Architectures for General-Purpose Computing*, AI tech. report 1586, MIT Artificial Intelligence Lab., Sept. 1996; www.seas.upenn.edu/~andre/abstracts/dehon_phd.html.
11. H. Kaul et al., “Near-Threshold Voltage (NTV) Design: Opportunities and Challenges,” *Proc. 49th ACM/EDAC/IEEE Design Automation Conf. (DAC 12)*, 2012, pp.1149–1154.
12. J. Appenzeller, “Carbon Nanotubes for High-Performance Electronics—Progress and Prospect,” *Proc. IEEE*, vol.96, no.2, 2008, pp.201–211.
13. A.D. Franklin et al. “Sub-10 nm Carbon Nanotube Transistor,” *Nanotechnology Letters*, vol. 12, 2012, pp. 758–762.
14. H. Park et al., “High-Density Integration of Carbon Nanotubes via Chemical Self-Assembly,” *Nature Nano*, vol. 7, no. 12, 2012, pp. 787–791.
15. R.F. Service, “Beyond Graphene,” *Science*, vol. 348, no. 6234, 2015, pp. 490–492.
16. T.N. Thies, “In Quest of a Fast, Low-Voltage Digital Switch,” *ECS Trans.*, vol. 45, no. 6, 2012, pp. 3–11.
17. A.F. Benner et al., “Exploitation of Optical Interconnects in Future Server Architectures,” *IBM J. Research and Development*, vol. 49, nos. 4-5, 2005, pp. 755–776.