Computing beyond Moore's Law

Article *in* Computer · December 2015

DOI: 10.1109/MC.2015.374

CITATIONS
50

READS
1,875

2 authors:

John Shalf
Lawrence Berkeley National Laboratory
262 PUBLICATIONS **8,513** CITATIONS

SEE PROFILE

Robert Leland
Sandia National Laboratories
58 PUBLICATIONS **2,319** CITATIONS

SEE PROFILE

SAND2015-8039J

Computing Beyond the End of Moore's Law: Is it really the end, and what are the alternatives?

John Shalf: *Department Head, Computer and Data Sciences, and CTO for National Energy Research Supercomputing Center, Lawrence Berkeley National Laboratory*

Robert Leland: *Vice President, Science & Technology, Chief Technology Officer (CTO), Sandia National Laboratories*

June 9, 2015

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

Abstract

The end of line for Moore's Law is now in sight as photolithography systems are on pace to reach atomic scale by the end of the next decade. Since 1965, the worldwide electronics industry has come to depend on the rapid, predictable, and cheap scaling of computing performance that was enabled by technology scaling. The end of conventional technology scaling impacts all manner of computing technologies that are dependent on improvements in cost and energy efficiency, and storage capacity — from large scale computing down to the smallest consumer electronic devices. This article describes limits in existing semiconductor microelectronic technology at the device level and their impact at the system level. It then examines alternatives for future computing technologies and proposes an approach to assessing their viability. Finally we offer in our conclusion a summary of major implications of the end of Moore's Law and a perspective on a research roadmap to navigate this challenge.

Introduction

In 1965, Gordon Moore famously observed that the number of components on an integrated circuit had been doubling every year on average since the introduction of this technology in 1959 [1] (although this eventually moderated to doubling approximately every 18 months today). He predicted this trend, driven by economic considerations of cost and yield, would continue for at least a decade. He also noted "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 observation that was formalized in nearly a decade later by Robert Dennard of IBM as Dennard Scaling. This mutually reinforcing scaling behavior (feature size, frequency, and power) meant that chip functionality would improve exponentially with time at roughly constant cost per generation, and Moore predicted this 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 hold true, and lasted 40 years longer than he initially predicted. However, these technology scaling rules are under challenges as Dennard scaling came to an end in 2004 – leading to a power efficiency crisis for CMOS logic. But traditional technology scaling will be faced with an even more fundamental challenge over the next decade.

Within that decade, the magical scaling process Moore described will come to an end as 2D lithography capability approaches atomic scale. This scaling limit will affect not only traditional digital computing technology, but all other conventional microelectronic technologies as well. Currently there is no obvious successor technology to the currently ubiquitous Complementary Metal Oxide Semiconductor (CMOS) logic. Three basic options for continued progress in computing occur to us: (1) new devices, (2) new architectures (with or without new devices), and (3) new paradigms of computation. Substantial exploration and innovation in each of these areas can be expected.

In the near term, emphasis will likely be on developing CMOS-based devices that extend into the third, or vertical, dimension and on improvements in materials technology. These will likely co-evolve with new architectural approaches that do a better job of tailoring computing capability to specific computing problems, driven principally by large economic forces associated with the $4T/year global IT market. In the longer term, a transition toward new device classes and the emergence of practical systems based on new approaches to computing are likely. To be effective at meeting societal needs and expectations in a broad context, these new devices and computing paradigms will need to be economically manufacturable at scale, and they will also need to provide an exponential improvement path. This may require a substantial technological shift analogous to that experienced in the transition from vacuum tube to semiconductor technology.

This transition will require effort on a decadal time scale, so independently of whether the semiconductor roadmap has 10 or 20 years of vitality left, it is important to be laying the strategic foundation for change now. It is crucial to prepare for this transition by supporting relevant research and development, and our intent is to contribute a strategic perspective regarding that need.

Is It Really the End?

Far from a law, Moore's observation was an economic theory driven by Technology Scaling – the constant improvement in photolithography processes for shrinking the size of components on the chips. Technology scaling is indeed the underlying driver, and there is much concern that one of the components of technology scaling (2D improvements in Silicon Lithography) is approaching fundamental limits by the end of next decade. Multiple assaults on conventional technology scaling for digital electronics over the 50 year history of Moore's Law have challenged the conventional approach to performance improvement.

For example, prior to the 1990's, most digital logic was based on bipolar transistor logic. In the 1980's scaling the power density of this technology became impractical. The inability to continue to scale bipolar (TTL, ECL, etc..) logic led to a wholesale transition to more efficient CMOS (Complementary Metal Oxide Semiconductor) logic technology, which enabled another two decades of technology scaling. More recently in 2004 Dennard scaling [2] (the ability to reduce device operating voltages and scale clock frequencies each generation at exponential rates) began to fail, and the computing industry was again in a power density crisis much like the bipolar crisis. With no performance growth per processing core, computer architects have responded by moving to exponentially scaling the number of cores per chip (multicore technology) to continue technology scaling for another decade. Today, the energy efficiency of logic gates continue to scale (albeit at a slower rate), the data carrying capacity and efficiency of wires is not improving substantially.

Whereas current programming systems are designed to conserve logic operations at the expense of data movement, the increased cost of data movement relative to logic operations portends a move from mostly compute-centric programming paradigms to more data-centric paradigms. The primary point is that despite the failure of numerous underlying physical mechanisms for technology over a 50-year history of Moore's law., the community has previously found new approaches to continue this scaling. One researcher famously quipped "I predict Moore's Law will never end -- that way, I will only be wrong once!"

Our view is that technology scaling is now at serious risk because the limits to 2D lithography scaling are fundamental and because there is no obvious successor technology. A silicon atom is approximately half a nanometer in diameter in semiconductor material. At the current rate of improvement, photolithography systems will be able to create transistor features on the scale of handfuls of atoms using 5nm technology in 2022–2024 [3]. This feature size corresponds to a dozen or fewer Si atoms across critical device features and the technology will therefore 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, e.g. tunneling of electrons through the gate oxide, which will increase energy losses due to leakage current. Although it is technologically feasible to reach 3–5nm by 2022 using Extreme Ultraviolet (EUV), the actual rate of adoption of smaller feature sizes is more driven by economics and return on investment (ROI) through performance improvements than by technological feasibility. Furthermore, the rapidly increasing cost of lithography methods may make their production economically infeasible.

At this point, further improvements in lithography will no longer result in smaller feature sizes or sufficiently better device performance to justify the increased costs. Hence the end of Moore's Law is really the end of useful two-dimensional scaling of photolithography. Constraints imposed by

fundamental device physics, and the increased manufacturing costs of producing smaller transistors (which are dominated by lithography) are looming limits. Which will become insurmountable first is unclear. Nevertheless, we conclude that further improvements in planar circuit density will become infeasible though one of these mechanisms. As a result, the computing industry will no longer be able to scale up computing capability using the basic approach that has worked so well since 1965.

The end of Moore's Law will affect all devices that depend on shrinking feature size to make progress. These include both processing devices and storage devices [4]. To increase circuit or storage density further, it will be necessary to build in the third dimension, which will require a technology that supports signal gain and reduces the energy consumed by data movement. The intrinsic resistance of the material used by "wires" (electrical connections of some form) will limit any solution involving electrons. The principal problem is that the energy consumed to transmit a bit is proportional to the distance it must travel due to the resistance and capacitance of the metal used to conduct the electrons representing the bit [11]. 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 out in the third dimension.

Society has come to expect and rely upon the benefits provided by Moore's Law. The end of Moore's Law will pose packaging and performance challenges for all manner of consumer electronic devices that are dependent on improvements in cost and energy efficiency to squeeze more functionality out of a device with a limited battery capacity. For example, the ability to make a smart phone "smarter" will be compromised if the industry is not able to pack more devices into a smaller space. The deeper issue presented by these changes is the threat to future economic growth of the U.S. computing industry. Moore's law scaling turned computing into a pervasive consumer technology. As computing technology became more powerful, it was used in more and more ways, and the market has grown exponentially as a result. We conclude that an end to Moore's law scaling threatens to cause stagnation in product innovation, which in turn could have significant negative economic impact.

What are the alternatives and how do we assess their viability?

A recent Intelligence Advanced Research Projects Activity (IARPA) report entitled *An Initial Look at Alternative Computing Technologies for the Intelligence Community* posits that, given these looming challenges, it may be necessary to consider a broader view of what constitutes computation. Further, it concludes that process should begin by considering four basic computational models:

1. **Classical Digital Computing (CDC):** CDC includes all forms of binary digital electronics that form the basis for the entire computing and consumer electronics industry.
2. **Analog Computing (AC):** Non-binary devices that implement computation through direct physical principles.
3. **Neuro-inspired Computing (NC):** Devices that are based upon the principles of operation of the brain and general neuronal computation.

4. **Quantum Computing (QC):** Quantum entanglement could in theory be used to solve some problems with combinatoric complexity through selection of a desired state from a superposition of all possible answers to a problem.

In an important insight, the report recommends that it is necessary to distinguish between new paradigms for computation and new technological implementations of existing paradigms. In particular the report makes the following observations;

- **Analog Computation:** AC can be simpler than some digital approximation, but does not lend itself to general purpose computing because the form of the device is specialized for computation. The computational precision is problematic to maintain and can be sensitive to its environment.

- **Neuromorphic/Bio-Inspired:** Digital computers are good at deterministic/algorithmic calculation, but are poor at simple tasks of reasoning and recognition. Neuro-inspired computing devices have been demonstrated to be inherently resilient and very good at problems that CDC is not. There are many unexplored opportunities for such computational models, but much is still not understood about how the brain actually computes.

- **Quantum information processing:** Quantum information processing could in theory enable efficient solution of some combinatorial and NP-Hard problems (problems that cannot be solved in polynomial time using digital computation). It is not a suitable replacement for CDC in areas where this approach excels.

We agree with the position in the iARPA white paper [12] that these technology options create the possibility of approaches that go well beyond what CMOS and digital electronics technologies have traditionally performed effectively and, therefore, are worthy of research investment. But they are not suitable as replacements for digital electronics for tasks on which digital computing already performs well. Therefore for the remainder of this article, we concentrate on new technological implementations of the CDC model because this is the model with the most immediate relevance to a broad set of societal concerns associated with the end of Moore's Law.

Metrics for evaluating a CMOS/CDC alternative

In the past, a competitor to CMOS/CDC would need to keep pace with a relentless schedule of improvement in which CMOS doubled its performance every 18 months or so and was able to leverage tremendous economies of scale. This combination of increased performance and economies of scale proved unbeatable except in specific, 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 we can apply metrics to assess the fitness of potential competitors. In a 2006 paper [5] Shekhar Borkar of Intel offered three properties required for a CMOS replacement to which John Shalf added an additional tenet for manufacturability:

1. **Gain**: The energy required to switch the device state from on to off must be less than the energy the device controls.
2. **Signal to noise immunity**: The signal used must be far enough above the background noise level to detect the signal.
3. **Scalability:** The technology must allow density increase and corresponding energy reductions as the technology improves.
4. **Scalable manufacturability**: The technology must be produced with a process capable of implementation at industrial scale.

Potential Post-CMOS Technology Options

This section lists potential post-CMOS technologies and describes the approach, benefits, and challenges to commercialization of each. Although we did not perform a detailed assessment of these technologies, the Borkar/Shalf and IARPA criteria are used to evaluate their potential.

Denser Packaging: Three Dimensional (3D) Integration and Packaging is being pursued vigorously and has been successful in mainstream devices to increase logic density and to reduce data movement distances. Most memory devices shipped today have some form of chip-stacking involved. Examples of approaches to increasing the density of future devices using 3D integration include:

- **Chip-Stacking in 3D Using Through-Silicon-Via (TSV):** stacking involves drilling holes through silicon chips to provide electrical connections between layers of stacked silicon chips. Chip stacks up to 8 layers deep are available that have lower engineering costs than adding layers lithographically or epitaxial deposition. TSV offers lower bandwidth and less efficient connectivity between chip layers than adding layers with photolithography processes, but far higher bandwidth and efficiency than using conventional chip packaging.
- **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 only between components on the 2D surface.
- **Epitaxial Deposition:** One issue with lithographic layering is that it results in only one layer of the semiconducting material (the bottom silicon layer), which is still 2D planar. To get more active transistors, more layers of semiconductor material need to be added. Epitaxial deposition involves a chemical or vapor deposition process to add layers of semiconducting materials on top of each other. Challenges remain in depositing high quality, single-crystal active layers, but there is substantial progress in studying approaches beyond standard silicon, for example approaches that use direct transfer of very thin layers of bulk crystalline material using processes other than epitaxial deposition.

The primary challenges to scaling 3D lithographic layering are improved defect tolerance and managing the thermal densities and intrinsic resistance. The stacking of cool technologies such as emerging nonvolatile memory cells (MRAM, Memristor, etc.) offer substantial opportunity to enable deeper lithographic layering (potentially a few orders of magnitude improvement). Although stacking in

3D will substantially reduce data movement requirements, which are a major contributor to thermal density, it is still unclear how much additional headroom that offers to enable deep stacking of logic layers. In the interim, 3D memory technologies will lead the way in 3D integration. Technologies that reduce operating voltages for digital circuits (which has been stalled since 2004), could provide further headroom in building out circuits in the 3rd dimension. Many 3D stacking approaches eventually fail to scale due to energy density limits of stacking energy-intensive logic layers atop one-another. Any future "electronic" system will need to have material that reduces switching power for the logic and resistive losses for information transfer within each constituent logic layer (e.g., exotic lower resistance wires, or switches that require far lower voltages).

Advanced Materials/Structures: Improving Transistor (Switch) Performance is one approach to improving device performance that underlies most digital electronic devices.

- **Tunneling Field-Effect Transistor (TFET)**: Conventional field-effect transistors (FETs) have a device performance limited by the voltage swing required to turn them completely on or off (this is gain in Borkar's criteria). A TFET uses a channel material that modulates the quantum tunneling effect, rather than the classical MOSFET modulation of thermionic emission, to create a switch that is more sensitive to gate voltage when turning on/off, and therefore can operate at a lower voltage. Since power dissipation of the device is proportional to voltage^2, the opportunity for energy efficiency improvements (and therefore future technology scaling) is substantial. Different materials systems are being investigated, but thermal sensitivity, speed, challenges in reliable manufacture (since more complicated materials systems are used than currently), and other scalability issues challenge current devices. Successful development of TFET devices may enable one or two additional generations of technology performance scaling improvements in the absence of lithography improvements, but it will take 10 years before such advances in the laboratory can be translated into mainstream mass production. Other technologies involve new gate designs to improve transistor sensitivity, such as Ferroelectric Gate FETs and other technologies. They all have similar challenges in manufacturing, and similar opportunities to extend technology energy efficiency (and hence performance) via lowering operating voltages.
- **Heterogeneous Semiconductors/Strained Silicon***:* Silicon has become the primary semiconductor material for integrated circuits due to its favorable chemical properties and physical robustness. For example, semiconductors formed from combinations of columns 3 and 5 of the periodic table (known as III-V materials) e.g. gallium arsenide offer much higher performance, but are more subject to cracking, form poor quality oxides, and require more challenging chemical processing steps. Relative to silicon, the manufacturability challenges have kept III-V materials on the margins of mainstream digital electronics. More recently, there have been dramatic improvements in the technology for integrating III-V materials (gallium arsenide, aluminum gallium arsenide, and germanium) as islands onto bulk silicon substrates so as to gain the manufacturing, chemical, and electrical benefits of silicon together with the device performance benefits of the embedded III-V materials. This is accomplished by straining the silicon substrate so that its atomic spacing aligns with that of the III-V material and doping 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. This approach, still in its infancy, remains an exotic processing technology. Gallium arsenide suffers from unbalanced P-type and N-type gate performance, which in turn affects its efficiency in CMOS devices, but combining with silicon using epitaxial deposition may overcome some of these challenges. This approach may offer an order of magnitude improvement in some *device* functions, but thus far many III-V materials do not offer an exact replacement for CMOS.

- **Carbon Nanotubes:** The band gap of carbon nanotubes is much smaller than that for silicon, so much less energy is required to operate devices based on carbon nanotubes. They also present lower resistance to electron movement, but this increases noise susceptibility. Carbon nanotube-based transistors have been demonstrated that deliver higher current levels than silicon-based devices, which in principle would enable them to operate at much higher switching rates and energy efficiency. Nanotube devices have been demonstrated with gain and noise rejection properties that are competitive with classical semiconductors for individual devices. Despite these favorable properties, mundane issues including contact resistance hold back progress in carbon nanotubes, and gate dielectric materials have yet to be fully engineered and optimized for nanotubes. Furthermore, nanotubes exist with a distribution of diameters and bandgaps. This leads to challenging device variation, so manufacturing high purity nanotubes with uniform diameter remains a major challenge. The primary challenge for nanotubes lies in finding a scalable manufacturing process, as current devices require precise placement of the individual tubes to form transistors and circuits. There have been dramatic advances recently in self-assembly processes for nanotube-based circuits [6], but there is still a long way to go before a competitive, commercially scalable process will be available.
- **Graphene** Unlike carbon nanotubes, graphene (a planar matrix of carbon atoms) does not have a bandgap, so it is not suitable for digital switches that turn off and have very low leakage current. The most promising solution for this is to fashion graphene into very narrow ribbons. In this case, they are similar to nanotubes. Nanotubes are in effect just graphene rolled into perfectly smooth tubes. The challenge here is to manufacture graphene nanoribbons with uniform width and atomically smooth edges. Graphene nanoribbons are less well developed than nanotubes, but breakthrough synthesis techniques for ribbons are emerging that might actually be better and cheaper at producing pure, uniform ribbon width than techniques for doing this for nanotubes.
- **Piezo Electric Transducer (PET):** PET devices make use of the piezo-electric effect in which an electric field induces a mechanical stress by changing the size of the material. The most common use of the PFET variants has been for micromechanical systems and force sensors, but if such piezo materials can be successfully miniaturized, the technology could be used to form an extremely fast (multi-gigahertz) micro-scale electronic relay [7]. This is one of many approaches involving micro-mechanical approaches to developing higher performance switches.

Computer Architecture: Technology advances described thus far focus on advanced materials or packaging. This section provides examples of advances in computer architecture and software that can offer a boost in performance.

- **Advanced Energy Management:** Current energy management technologies are ubiquitous and typically very coarse grained. Dynamic Voltage and Frequency Scaling (DVFS) and thermal throttling enables savings by lowering clock frequencies and lower voltages when computing demands do not require peak performance. Coarse-grained DVFS offers significant power savings for current consumer electronics devices (that are mostly idle), but are only marginally beneficial for devices operating near 100% utilization. There may be additional potential to recover energy with finer-grained power management and faster transitions between power states and through software direction of power state changes.
- **Advanced Circuit Design:** Approaches have been demonstrated that enable wires to operate at a lower voltage for long-haul connections and then be re-amplified efficiently at the end-points (there is a loss for re-amplification). Personal communication with Bill Dally indicated an opportunity for 2x–3x improvement using improved circuit design techniques based on current technologies.
- **Clockless/Domino Logic:** Clock distribution is known to consume a large fraction of system power, and consign circuit designs to operate at the speed of its slowest component. Practical and effective clockless designs have proven elusive, but recent examples have shown that this approach holds much promise to lower dynamic power consumption for both neurmorphic and digital applications.
- **System-on-Chip Specialization:** The core precept of System-on-Chip (SoC) technology is that chip cost is dominated by design and verification costs for individual circuit components. Therefore, it economically viable to tailor chips to include only the circuit components that are valuable to the application, rather than pursuing the current commodity design practice of designing one chip that serves the needs of a broad range of applications. This approach is in common practice for cell phone chips (such as that in the Apple iPhone, which uses commodity embedded processor cores combined in a specialized SoC design), but is only just beginning to be considered for the design of server and HPC chips.
- **Custom Logic:** Field-programmable gate arrays (FPGAs) and reconfigurable computing hold promise for improving performance by creating custom-tailored circuits for each problem, but are held back by lower efficiency in implementation. A typical FPGA implementation over-provisions wires (e.g., the majority of available reconfigurable wires on an FPGA remain unused [9]) in order to maximize use of Lookup Tables (LUTs). FPGAs offer the opportunity to improve performance over conventional computer processing unit instruction set architectures (ISAs), but, a custom application-specific integrated circuit (ASIC) design offers a tenfold improvement in performance over the FPGA design of the same circuit due to elimination of redundant wiring. In addition, circuit design is very expensive relative to software design, and most reconfigurable computing technology requires substantial expertise in hardware design to optimize performance. It is possible that the economic disincentive of designing and verifying custom circuits will be overcome by the alternative of having no performance scaling at all.
- **Dark Silicon:** The most extreme examples of customized logic propose a family of custom designs that are implemented in an ASIC and remain turned off ("dark") when not used. The concept is to expend more ASIC surface area for more efficient specialized circuits with the hope

that the cost of extra area will be offset by the performance benefits. This approach remains energy neutral by turning off the specialized circuits when they are not required. It is currently used to good effect for some specialized consumer electronics applications, but its usefulness for general-purpose computation has yet to be proven.

- **Near-Threshold Voltage Operation (NTV)**: Thus far, the mainstream computing community has shunned further reductions in device voltage because it reduces the signal-to-noise immunity of the transistors and would subject circuits to wider statistical variation on performance. Both effects present huge challenges to software and hardware development. From a software standpoint, the nondeterministic performance of individual circuits (and hence individual processor elements) would make conventional bulk-synchronous approaches to scaling parallel computing performance untenable – requiring a move towards entirely asynchronous software execution models and corresponding reformulations of algorithms and infrastructure. Applications and algorithm developers would need to substantially re-write software to accommodate this unpredictable performance heterogeneity. In hardware, the increased unreliability would require more pervasive detection of errors and corresponding software infrastructure to respond to such errors, and the cost of this detection is not known. The approach would also reduce clock frequencies substantially, putting more pressure on parallelism to gain performance improvements (an already daunting challenge for software). The opportunity offered by NTV circuit operation is the potential to reduce operating voltages and hence increase energy efficiency of devices (and hence usable performance and scalability) by an order of magnitude. It remains an area of active research [8] to determine whether the software challenges posed by reliability, performance heterogeneity, and increased parallelism will detract from the raw potential performance improvement offered.

Reducing Resistance is accomplished in most modern integrated circuits by using copper-based wires to interconnect circuits. Copper is a particularly good conductor, and at room temperature, there are few options that can offer lower electrical resistance, aside from:

- **Superconducting:** Superconducting may be a path forward to advancing HPC system performance, but will force a departure from the mainstream since it is unlikely that cooling will be practical for consumer devices. Even the cuprate-based high-temperature superconductors have cooling and magnetic shielding requirements impractical for consumer applications. The viability of cryogenically cooled electronics in mainstream phones or laptops is doubtful. There is a technological path to use cryo-cooled electronics to extend HPC performance, but it would entail a departure from the path of leveraging commodity component technology. This could have significant repercussions for U.S. domestic HPC competitiveness and the affordability of HPC systems, which is dependent on our ability to leverage the commodity electronics market.
- **Crystalline Metals:** Copper is widely used to interconnect layers of chip designs (through wires). Copper is a very good conductor, but in a typical polycrystalline configuration, electrons still scatter off of the boundaries between neighboring crystalline grains. The conductivity of metal layers could be improved by as much as a factor of 5 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 simply not being shared for proprietary reasons).

Change how Bits are Stored and Transformed: Other Advanced Materials Systems provide a critical path forward to improving electronic systems performance. The materials introduced above are able to improve the performance of devices using familiar digital computing architectures and computational models. The devices discussed below can also improve digital electronics, but require a departure from stock computing principles.

- **Spintronics:** Computation on and communication of information through manipulation of magnetic domains is lower in energy costs than moving electrons to such a degree that it is nearly inconsequential to overall power consumption. SEMATECH has stated that spin materials hold the promise for dual functionality (logic + memory) and new circuit (SRAM) concepts. 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 (full CMOS replacements), would require an adiabatic or reversible model of computation. Such models for computation can be highly restrictive and would fundamentally disrupt our current model for computation.
- **Topological Insulators:** Compared with conventional wires, 2D confined energy states can offer more efficient (higher noise margin) information transport and storage, but the proper approach to implementing logic is uncertain. For example, topological insulators offer a possible path to 2D image analysis algorithms using photo-galvanic effect to program initial state for qubits embedded in the topological insulator. According to SEMATECH new semiconductors with unique properties such as the 2Dsemiconductors are being considered carefully by the electronics industry.
- **Nanophotonics:** The key challenges in using subwavelength scale nanophotonics as a replacement for computing/transistor technologies is the low gain of available optical "transistors" and the large size of the optical wavelengths in comparison to to current realizable photolithographic scales. The more obvious benefit of photonic technology is for scalable communications. For communications (wire replacement), photonics has the benefit of having energy costs that are nearly independent of distance that data travels, whereas the standard electrical wires have a strong distance-dependent energy cost. Therefore, photonic technology overcomes the fundamental limitation of wire resistance. Unfortunately, the energy required to light up the laser to send information over a photonic connection is currently far higher than the cost for the wire, but it has steadily decreased over time[10]. Photonics will play an essential role in overcoming limits of wires and the disparity between on-chip and off-chip communications costs [11]. Development of an effective high-gain optical transistor would enable nanophotonics to also be competitive as a CMOS replacement for computation, but the technology for a high-performance optically controlled switch requires further development.
- **Chemical/Biological Computing:** The potential of biologically based devices is that it takes as its inspiration the most complex machines known on earth — animal brains. The principal challenges to biologically based computing devices include low gain, poor signal to noise, 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 exotic operating environments, current examples are difficult to scale.

Options for extending technology scaling of digital electronics beyond Moore's Law are summarized in Table 1. This review is by no means comprehensive, but provides an overview of some of the most commonly discussed options. Given that none of the options is clearly superior in all respects, it is likely that one or more of these options will find their way into mainstream use through integration with conventional silicon/CMOS platforms. Indeed, chip stacking is already enabling photonics technology to be stacked directly on conventional silicon logic and memory circuits.

Table 1: Summary of technology options for digital electronics.

Improvement Class	Technology	Timescale	Complexity	Risk	Opportunity
Improving Transistor Performance	Tunnel Field Effect Transistors (TFET)	Mid-Term	Low	Medium	High
	Heterogeneous Semiconductors/Strained Silicon	Mid-Term	Medium	Medium	Medium
	Piezo-Electronic Transistor	Far-Term	High	High	High
	Carbon Nanotubes and Graphene	Far-Term	High	High	High
Computer Architecture Advances	Advanced Energy Management	Near-Term	Medium	Low	Low
	Advanced Circuit Design	Near-Term	High	Low	Medium
	Near Threshold Voltage (NTV) Operation	Near-Term	Medium	High	High
	System on Chip Specialization	Near-Term	Low	Low	Medium
	Custom Accelerators/Dark Silicon	Mid-Term	High	High	High
3D Integration	Chip-Stacking in 3D using Thru-Silicon-Via (TSV)	Near-Term	Medium	Low	Medium
	Metal Layers	Mid-Term	Medium	Medium	Medium
	Epitaxial Deposition	Mid-Term	High	Medium	Medium
Reducing Resistance	Superconducting	Far-Term	High	Medium	High
	Crystalline Metals	Mid-term	unkown	Low	Medium
Other Advanced Materials Systems	Spintronics	Far-Term	Medium	High	High
	Topological Insulators	Far-Term	Medium	High	High
	Nanophotonics	Near/Far-Term	Medium	Medium	High
	Chemical/Biological Computing	Far-Term	High	High	High
	Diamond Films	Far-Term	High	High	Medium

The long-term challenge of post-Moore's Law technology requires investment in basic sciences, incuding materials science, to study candidate replacement materials and alternative device physics to enable continuation of technology scaling. Using the history of the silicon fin field-effect transistor (FinFET) as a guide, it takes about 10 years for an advance in basic device physics to make it in to mainstream use. Any new technology will require a long lead-time and sustained R&D on the order of 10-20 years. The winner of this technology race will influence not just chip technology – it will define the direction for the entire computing industry.

Conclusion

The coming end of Moore's Law presents major technological challenges for society. From now until the end of Moore's Law, the energy cost of data movement will become a dominant technical and economic factor because the energy cost of computational operations is improving at a faster rate than the energy cost of moving data to those operations. A near-term adaptation to this will be to increase the use of parallelism in software, which will require a huge commercial effort. Over time it will likely be necessary to go even further by moving from a computer- centric model of computation to a data-centric model.

In the near term emphasis is expected to be on developing CMOS-based devices that extend into the third, or vertical, dimension and on improvements in materials technology. These will likely co-evolve with new architectural approaches that do a better job of tailoring computing capability to specific computing problems. This evolution will be driven principally by large economic forces associated with the $4T/year global IT market, but well targeted R&D investments can play an important nurturing role.

In the longer term, we expect a transition toward new device classes and the emergence of practical systems based on new approaches to computing. To be effective at meeting societal needs and expectations in a broad context, these new devices and computing paradigms will need to be economically manufacturable at scale. They will also need to provide an exponential improvement path. This may require a substantial technological shift analogous to that experienced in the transition from vacuum tube to semiconductor technology. This transition is expected to occur on a decadal time scale, so whether the CMOS roadmap has in fact 10 years or 20 years of vitality left, it is important to be laying the strategic foundation for a major change now. The industry must prepare for this transition by sponsoring relevant research and development, and by articulating an over-arching strategic vision.

Bibliography

[1] Gordon E. Moore, "Cramming More Components onto Integrated Circuits," *Electronics* 38 (8): 114–117, April 19, 1965.

[2] Robert H. Dennard et al. "Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions," *IEEE Journal of Solid-State Circuits* SC-9 (5): 256–268, October 1974.

[3] Robert Colwell (DARPA) at HotChips 2013, http://www.hotchips.org/wp-content/uploads/hc_archives/hc25/HC25.15-keynote1-Chipdesign-epub/HC25.26.190-Keynote1-ChipDesignGame-Colwell-DARPA.pdf.

[4] Robert E. Fontana, S. R. Hetzler, and G. Decad, "Technology Roadmap Comparisons for TAPE, HDD, and NAND Flash: Implications for Data Storage Applications," *IEEE Transactions on Magnetics*48 (5): 1692,1696, May 2012 (doi: 10.1109/TMAG.2011.2171675).

[5] S. Borkar, "Electronics Beyond Nano-scale CMOS," *DAC 2006*, July 24–28, San Francisco, California, 2006.

[6] Park, Hongsik and Afzali, Ali and Han, Shu-Jen and Tulevski, George S. and Franklin, Aaron D. and Tersoff, Jerry and Hannon, James B. and Haensch, Wilfried, "***High-density integration of carbon nanotubes via chemical self-assembly,"*** Nature Nano, Vol 7, No 12, 2012, PP(787—791), http://www.nature.com/nnano/journal/v7/n12/full/nnano.2012.189.html

[7] T. N. Thies, "In Quest of a Fast, Low-Voltage Digital Switch," *ECS Transactions* 45 (6, 2012): 3–11.

[8] Open Community Runtime (OCR): https://01.org/open-community-runtime

[9] Andre DeHon, "Reconfigurable Architectures for General-Purpose Computing," AI Technical Report 1586, MIT Artificial Intelligence Laboratory, 545 Technology Sq., Cambridge, MA 02139, September 1996.

[10] Alan F. Benner, Michael Ignatowski, Jeffrey A. Kash, Daniel M. Kuchta, Mark B. Ritter: Exploitation of optical interconnects in future server architectures. *IBM Journal of Research and Development* 49(4–5, 2005): 755–776.

[11] D. A. B. Miller, "Device Requirements for Optical Interconnects to Silicon Chips," *Proceedings of the IEEE* 97 (2009): 1166– 1185.

[12] L. Joneckis, D. Koester, J. Alspector, "An Initial Look at Alternative Computing Technologies for the Intelligence Community," Institute for Defense Analysis (IDA), January 2014.