The future of computing beyond Moore's Law

John Shalf

Department of Computer Science, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA

JS, 0000-0002-0608-3690

Moore's Law is a techno-economic model that has enabled the information technology industry to double the performance and functionality of digital electronics roughly every 2 years within a fixed cost, power and area. Advances in silicon lithography have enabled this exponential miniaturization of electronics, but, as transistors reach atomic scale and fabrication costs continue to rise, the classical technological driver that has underpinned Moore's Law for 50 years is failing and is anticipated to flatten by 2025. This article provides an updated view of what a post-exascale system will look like and the challenges ahead, based on our most recent understanding of technology roadmaps. It also discusses the tapering of historical improvements, and how it affects options available to continue scaling of successors to the first exascale machine. Lastly, this article covers the many different opportunities and strategies available to continue computing performance improvements in the absence of historical technology drivers.

This article is part of a discussion meeting issue 'Numerical algorithms for high-performance computational science'.

1. Introduction

Society has come to depend on the rapid, predictable and affordable scaling of computing performance for consumer electronics, the rise of 'big data' and data centres (Google, Facebook), scientific discovery and national security. There are many other parts of the economy and economic development that are intimately linked with these dramatic improvements in information technology (IT) and computing, such as avionics systems for aircraft, the automotive industry (e.g. self-driving cars) and smart grid technologies. The approaching end of lithographic scaling threatens to hinder continued

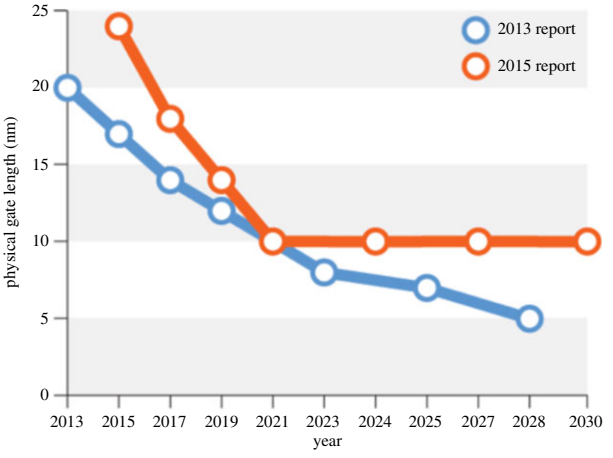


Figure 1. The ITRS most recent report predicts transistor scaling will end in 2021 (a decade sooner than was predicted in 2013). Figure from ITRS. (Online version in colour.)

health of the $4 trillion electronics industry, impacting many related fields that depend on computing and electronics.

Moore's Law [1] is a techno-economic model that has enabled the IT industry to double the performance and functionality of digital electronics roughly every 2 years within a fixed cost, power and area. This expectation has led to a relatively stable ecosystem (e.g. electronic design automation tools, compilers, simulators and emulators) built around general-purpose processor technologies, such as the ×86, ARM and Power instruction set architectures. However, within a decade, the technological underpinnings for the process that Gordon Moore described will come to an end, as lithography gets down to atomic scale. At that point, it will be feasible to create lithographically produced devices with dimensions nearing atomic scale, where a dozen or fewer silicon atoms are present across critical device features, and will therefore represent a practical limit for implementing logic gates for digital computing [2]. Indeed, the ITRS (International Technology Roadmap for Semiconductors), which has tracked the historical improvements over the past 30 years, has projected no improvements beyond 2021, as shown in figure 1, and subsequently disbanded, having no further purpose. The classical technological driver that has underpinned Moore's Law for the past 50 years is failing [3] and is anticipated to flatten by 2025, as shown in figure 2. Evolving technology in the absence of Moore's Law will require an investment *now* in computer architecture and the basic sciences (including materials science), to study candidate replacement materials and alternative device physics to foster continued technology scaling.

(a) Multiple paths forward

To address this daunting problem in both the intermediate and long term, a multi-pronged approach is required: evolutionary for the intermediate (10 year) term and revolutionary for the long (10–20 year) term strategy. Timing needs for the intermediate term will require an evolutionary approach based on achieving manufacturing technology advances allowing the continuation of Moore's Law with current complementary metal oxide semiconductor (CMOS) technology—relying on new computing architectures and advanced packaging technologies such as monolithic three-dimensional integration (building chips in the third dimension) and photonic co-packaging to mitigate data movement costs [4,5]. The long-term solution requires fundamental advances in our knowledge of materials and pathways to control and manipulate information elements at the limits of energy flow, ultimately achieving 1 attojoule/operation, which would

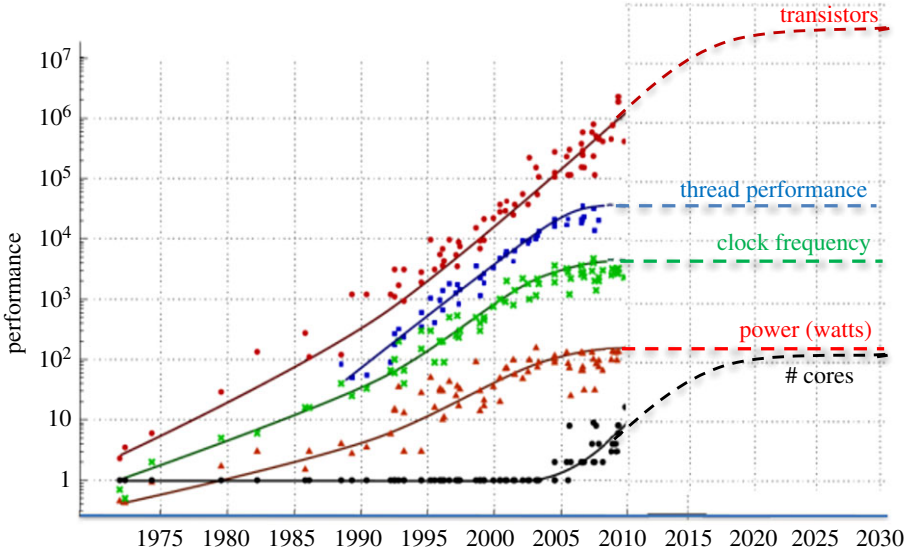


Figure 2. Sources of computing performance have been challenged by the end of Dennard scaling in 2004. All additional approaches to further performance improvements end in approximately 2025 due to the end of the roadmap for improvements to semiconductor lithography. Figure from Kunle Olukotun, Lance Hammond, Herb Sutter, Mark Horowitz and extended by John Shalf. (Online version in colour.)

be six orders of magnitude smaller than today's devices. As we approach the longer term, we will require ground-breaking advances in device technology going beyond CMOS (arising from fundamentally new knowledge of control pathways), system architecture and programming models to allow the energy benefits of scaling to be realized. Using the history of the silicon fin field-effect transistor (FinFET), it takes about 10 years for an advance in basic device physics to reach mainstream use. Therefore, 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 also will define a new direction for the entire computing industry and many other industries that have come to depend heavily on computing technology.

There are numerous paths forward to continue performance scaling in the absence of lithographic scaling, as shown in figure 3. These three axes represent different technology scaling paths that could be used to extract additional performance beyond the end of lithographic scaling. The near-term focus will be on development of ever more specialized architectures and advanced packaging technologies that arrange existing building blocks (the horizontal axis of figure 3). In the mid-term, emphasis will likely be on developing CMOS-based devices that extend into the third, or vertical, dimension and on improving materials and transistors that will enhance performance by creating more efficient underlying logic devices. The third axis represents opportunities to develop new models of computation such as neuro-inspired or quantum computing, which solve problems that are not well addressed by digital computing.

2. The complementary role of new models of computation

Despite the rapid influx of funding into these respective technologies, it is important to understand that they are not replacement technologies for digital electronics as we currently understand them. They certainly expand computing into areas where digital computing is deficient. Digital computing is well known for providing reproducible and explainable

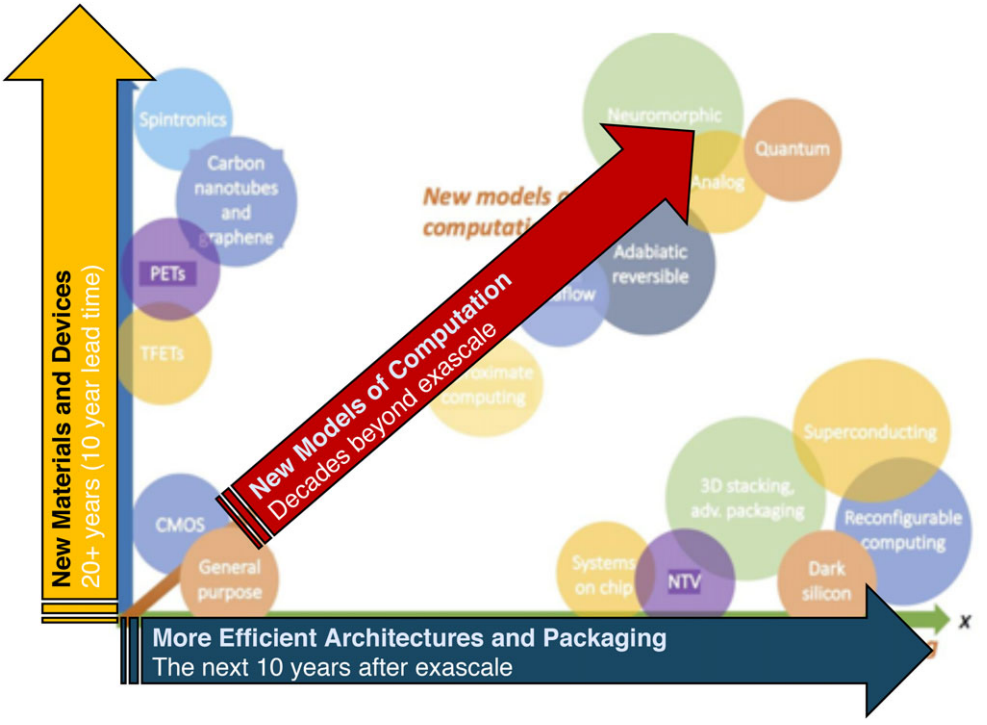


Figure 3. There are three potential paths forward to realize continued performance improvements for digital electronics technology. (Online version in colour.)

calculations that are accurate within the precision limit of the digital representation. Brain-inspired computational methods such as machine learning have substantially improved our ability to recognize patterns in 'big data' and automate data mining processes over traditional pattern recognition algorithms, but they are less reliable for handling operations that require precise response and reproducibility (even 'explainability' for that matter). Quantum computing will expand our ability to solve combinatorically complex problems in polynomial time, but they will not be much good for word processing or graphics rendering, for example. It is quite exciting and gratifying to see computing expand into new spaces, but equally important to know the complementary role that digital computing plays in our society that is not and cannot be replaced by these emerging modes of computation.

Quantum and brain-inspired technologies have garnered much attention recently due to their rapid pace of recent improvements. Much of advanced architecture development and new startup companies in the digital computing space are targeting the artificial intelligence/machine learning (AI/ML) market because of its explosive market growth rate. Growth markets are far more appealing business opportunities for companies and venture capital, as they offer a path to profit growth, whereas a large market that is static invites competition that slowly erodes profits over time. As a result, there is far more attention paid to technologies that are seeing a rapid rate of expansion, even in cases where the market is still comparatively small. So interest in quantum computing and AI/ML is currently superheated due to market opportunities, but it is still urgent to advance digital computing even as we pursue these new computing directions. Neither quantum nor brain-inspired architectures are replacement technologies for functionality that digital technologies are good at. Indeed, current AI/ML solutions are deeply dependent upon digital computing technology, and if there is any lesson to be learned from the diversity of AI/ML hardware solutions, it is that architecture specialization and custom hardware is *very* effective—the topic of the next section.

3. Architectural specialization

In the near term, the most practical path to continued performance growth will be architectural specialization in the form of many different kinds of accelerators. We believe this to be true because historically it has taken approximately 10 years for a new transistor concept demonstrated in the laboratory to become incorporated into a commercial fabrication process. Our US Office of Science and Technology Policy (OSTP) report with Robert Leland surveyed the landscape of potential CMOS-replacement technologies and found many potential candidates [4], but no obvious replacements demonstrated in the laboratory at this point. Therefore, we are already a decade too late to resolve this crisis by finding a scalable post-CMOS path forward. The only hardware option for the coming decade will be architectural specialization and advanced packaging for lack of a credible alternative. When competing against an exponentially improving general-purpose computing ecosystem, it was very difficult to compete using hardware specialization. In the past, the path of specialization has not been productive to pursue due to long lead-times and high development costs. However, as Thompson & Spanuth's [6] article on the evaluation of the economics of Moore's Law points out, the tapering of Moore's Law improvements makes architecture specialization a credible and economically viable alternative to fully general-purpose computing, but such a path will have a profound effect on algorithm development and on the programming environment.

Therefore, in the absence of any miraculous new transistor or other device to enable continued technology scaling, the only tool left to a computer architect for extracting continued performance improvements is to use transistors more efficiently by specializing the architecture to the target scientific problem(s), as projected. Overall, there is strong consensus that the tapering of Moore's Law will lead to a broader range of accelerators or specialization technologies than we have seen in the past three decades. Examples of this trend exist in smartphone technologies, which contain dozens of specialized accelerators co-located on the same chip; in hardware deployed in massive data centres, such as Google's Tensor Processing Unit (TPU), which accelerates the Tensorflow programming framework for ML tasks; in field-programmable gate arrays (FPGAs) in the Microsoft Cloud used for Bing search and other applications; and a vast array of other deep learning accelerators. The industry is already moving forward with production implementation of diverse acceleration in the AI and ML markets (e.g. Google TPU [7], Nervana's AI architecture [8], Facebook's *Big Sur* [9]) and other forms of compute-in-network acceleration for mega-data centres (e.g. Microsoft's FPGA Configurable Cloud and Project Catapult for FPGA-accelerated search [10]). Even before the explosive growth in the AI/ML market, system-on-chip (SoC) vendors for embedded, Internet of things (IoT) and smartphone applications were already pursuing specialization to good effect. Shao *et al.* [11] from Harvard University tracked the growth rate of specialized accelerators in iPhone chips, and found a steady growth rate for discrete hardware accelerator units, which grew from around 22 accelerators for Apple's 6th-generation iPhone SoC to well over 40 discrete accelerators in their 11th-generation chip. Companies engaged in this practice of developing such diverse heterogeneous accelerators because the strategy works!

There have also been demonstrated successes in creating *science-targeted* accelerators such as D.E. Shaw's Anton, which accelerates molecular dynamics (MD) simulations nearly 180× over contemporary high-performance computing (HPC) systems [12], and the GRAPE series of specialized accelerators for cosmology and MD [13]. A recent International Symposium on Computer Architecture workshop on the future of computing research beyond 2030 (http://arch2030.cs.washington.edu/) concluded that heterogeneity and diversity of architecture are nearly inevitable given current architecture trends. This trend toward co-packaging of diverse 'extremely heterogeneous' accelerators is already well under way, as shown in figure 4.

Therefore, specialization is the most promising technique for continuing to provide the year-on-year performance increases required by all users of scientific computing systems, but specialization needs to have a well-defined application target to specialize for. This creates a particular need for the sciences to focus on the unique aspects of scientific computing for both analysis and simulation. Recent communications with computing industry leaders

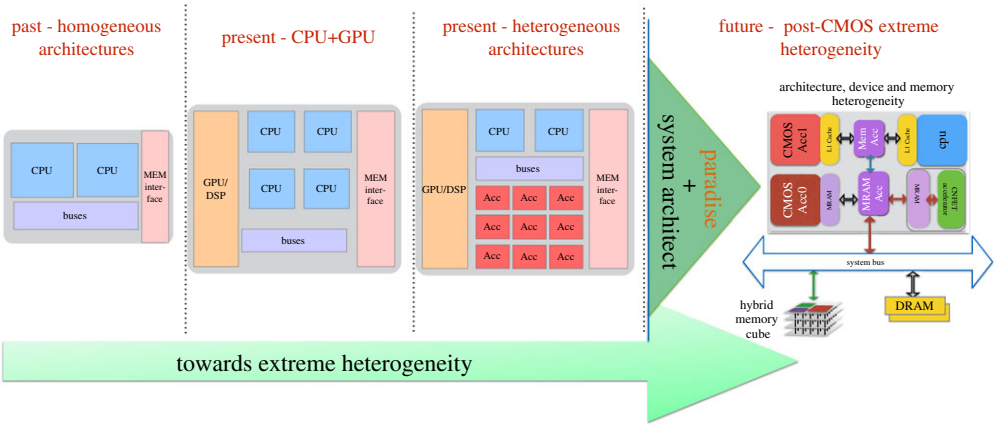


Figure 4. Architectural specialization and extreme heterogeneity are anticipated to be the near-term response to the end of classical technology scaling. Figure courtesy of Dilip Vasudevan from LBNL. (Online version in colour.)

suggest that post-exascale HPC platforms will become increasingly heterogeneous environments. Heterogeneous processor accelerators—whether they are commercial designs (evolutions of GPU or CPU technologies), emerging reconfigurable hardware or bespoke architectures that are customized for specific science applications—optimize hardware and software for particular tasks or algorithms and enable performance and/or energy efficiency gains that would not be realized using general-purpose approaches. These long-term trends in the underlying hardware technology (driven by the physics) are creating daunting challenges for maintaining the productivity and continued performance scaling of HPC codes on future systems.

As a means to organize the universe of options available, we subdivide the solution space into three different strategies:

(i) *Hardware-driven algorithm design*: where we evaluate emerging accelerators in the context of workload, and modify algorithms to take full advantage of new accelerators.
(ii) *Algorithm-driven hardware design*: where we design largely fixed-function accelerators based on algorithm or application requirements.
(iii) *Co-develop hardware and algorithms*: this represents a cooperative design with a selected industry partner or partnership to design algorithms and hardware together.

For hardware-driven algorithm design, we recognize that the industry will continue to produce accelerators that are targeted at other markets such as ML applications. In the near future, GPUs, accelerators (NVIDIA, AMD/ATI, Intel) and multi-core processors with wide-vector extensions (such as ARM SVE and Intel's AVX512) will continue to dominate. However, the boost in performance offered by the GPUs and wide-vector extensions to CPUs have offered a one-time jump in performance, but do not offer a new exponential growth path. There are a number of extensions emerging that are targeted at accelerating the burgeoning AI workloads, such as NVIDIA's tensor extensions in the V100 GPU. Such extensions are very specific tensor operations that operate at much lower (16-bit and 8-bit) precision, which may limit them unless algorithms are completely redesigned to exploit these features (where possible). While this puts the primary burden upon the algorithm and application developers, to some extent this is the strategy that has more or less been common practice since the 'attack of the killer micros' transformed the HPC landscape from purpose-built vector machines to clusters of commercial off-the-shelf (COTS) nodes nearly 3 decades ago.

Algorithm-driven hardware design would mark a return to past practices of designing purpose-built machines for targeted high-value workloads. As mentioned earlier, the rapid

growth and diversity in specialized AI architectures (Google TPU and others) as well as isolated examples in the sciences (D.E. Shaw's Anton, SPINNAKER, etc.) demonstrate that this approach can offer a path to performance growth. However, the development costs are high (tens to hundreds of millions of dollars per system in today's technology market), it requires long development lead times, and it risks having the application requirements shift so as to make the hardware obsolete. This concern has caused an increased interest in reconfigurable hardware such as FPGAs and coarse-grained reconfigurable arrays (CGRAs). These devices allow the logic and specializations within the chip to be reconfigured rather than having to build a new chip. The challenge with FPGAs is that the extreme flexibility to enable hardware reconfiguration comes at a cost of 5× slower clock rates (typical designs run at 200 MHz rather than at the gigahertz clock rates expected of custom logic) and a reduction of effective logic density (number of usable gates per chip) by a similar factor. CGRAs, such as Stanford's Plasticine [14], mitigate these problems by offering a coarser granularity of reconfiguration where the building blocks are full floating-point adders and multipliers rather than individual wires and gates offered by the FPGAs. The biggest challenge to making these devices useful is that the tools and programming models for programming these devices are extraordinarily difficult to use and it requires a lot of effort to get even simple algorithms to perform well. There is a lot of work going in to developing more agile hardware design methodologies such as higher-level hardware development languages (e.g. CHISEL, PyMTL), and more design automation to reduce human effort to make production of custom chips more affordable.

The third option of deeper co-design is less of a technological option than it is a new economic model for interacting with the industry that produces computer systems and the potential customers of said technologies. The era of general-purpose computing led to a more or less hands-off relationship between technology suppliers and their customers, as documented by Thompson & Spanuth [6], where a general-purpose processor could serve many different applications. In an era where specializing hardware to the application is the only means of performance improvement, the economic model for the design of future systems is going to need to change dramatically to lower design and verification costs for the development of new hardware. Otherwise, the future predicted by economists such as Thompson is one where high-value markets such as AI for Google and Facebook will be able to afford to create custom hardware (the fast lane) and the rest of the market will receive no such boosts (remaining in the slow lane). To prevent this kind of future from happening, the industry is adopting more agile hardware production methods such as using chiplets. Rather than have a single large piece of silicon that integrates together all of the diverse accelerators comprising the customized hardware, the chiplets break each piece of functionality into a very tiny *tile*. These chiplets/tiles are then stitched together into a mosaic by bonding them to a common silicon substrate. This enables manufacturers to rapidly piece together a mosaic of these chiplets to serve the diverse specialized applications at a much lower cost and much faster turn-around. However, this approach falls down if the desired functionality does not already exist in the available chiplets. Perhaps in the future the 'algorithm-driven hardware design' and this chiplets approach might be able to meet in the middle to bring forth a new economic model that can enable productive architecture specialization for small markets, such as Dr Sophia Shao's [11] vision for her Aladdin integrated hardware specialization/design environment.

(a) Programming system and software challenges

New software implementations, and in many cases new mathematical models and algorithmic approaches, are necessary to advance the science that can be done with new architectures. This trend will not only continue but also intensify; the transition from multi-core systems to hybrid systems has already caused many teams to re-factor and redesign their implementations. But the next step to systems that exploit not just one type of accelerator but a full range of heterogeneous architectures will require more fundamental and disruptive changes in algorithm and software approaches [15]. This applies to the broad range of algorithms used in simulation, data analysis

and learning. New programming models or low-level software constructs that hide the details of the architecture from the implementation can make future programming less time-consuming, but they will not eliminate nor in many cases even mitigate the need to redesign algorithms. Key elements of a path forward include:

— Understanding the impact of proposed architectures on current mathematical kernels and algorithms and using this knowledge to steer the HPC hardware deployment choices through feedback in an iterative co-design process.
— Redesigning current algorithms in response to proposed architectures; hardware choices should be based not only on current algorithms but also on the potential performance of new algorithms and even new science use cases.
— Developing advanced programming environments that ease the implementation of these new algorithms and numerical libraries and are able to generate code for these diverse, heterogeneous accelerators.

Applied mathematics is critical to our ability to co-design application- and science-relevant accelerators. There are two categories of applications that will need to be redesigned to run effectively in a heterogeneous accelerated environment. In the first type, a single computational motif or kernel is paramount, such as stencil computations with fixed spatial patterns. In this case, there is likely to be a single best choice of hardware design. Most of the success stories regarding specialized architectures fall into this category. The advances in numerical methods can be encapsulated in numerical libraries (such as SuperLU, GraphBLAS and STRUMPACK) and application frameworks (such as AMReX) to make these advances broadly available to the community. The second, more complex type is that in which solving the science problem requires fundamentally heterogeneous operations. The heterogeneous operations can be staggered, as one might envision in a data pipeline; as the data moves through the pipeline, different operations are performed on it. In this scenario, the data may also be moving physically in steps from source to destination, making the use of different architectures for different stages transparent and separable. Heterogeneous simulation algorithms place a different demand in that, unlike the data example, the flow is more fine-grained and tightly coupled. For example, in a simulation of a time-evolving state or any iterative solution procedure, each step may contain multiple heterogeneous substeps, with each step repeated multiple times, perhaps with different relative (i.e. dynamically changing) costs of the components. No single specialized architecture will be ideal for all stages, suggesting an architectural layout that allows a single code to exploit multiple specialized components. Existing hybrid CPU/GPU systems already allow this, and applications are being re-factored to use this capability; the current trend of offloading different algorithmic components to different specialized architectures will not only continue but become more important.

Performance portability is not an achievable goal if we attempt to do it using imperative languages like Fortran and C/C++. There is simply not enough flexibility built in to the specification of the algorithm for a compiler to do anything other than what the algorithm designer explicitly stated in their code. To make this future of diverse accelerators usable and accessible in the former case will require the co-design of new compiler technology and domain-specific languages (DSLs) designed around the requirements of the target computational motifs (the 13 motifs that extended Phil Colella's original Dwarfs of algorithmic methods [16]). The higher levels of abstraction and declarative semantics offered by DSLs enable more degrees of freedom to optimally map the algorithms onto diverse hardware than traditional imperative languages that over-prescribe the solution. Because this will drastically increase the complexity of the mapping problem, new mathematics for optimization will be developed, along with better performance introspection (both hardware and software mechanisms for online performance introspection) through extensions to the roofline model. Use of ML/AI technologies will be essential to enable analysis and automation of dynamic optimizations.

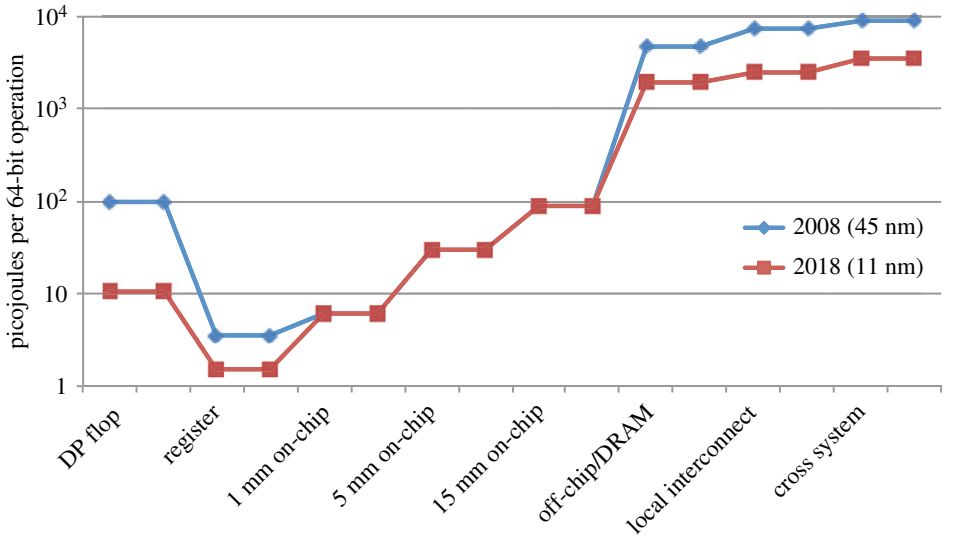


Figure 5. The energy consumption of compute and data movement operations at different levels of the compute hierarchy—from the arithmetic logic unit on the left to system-scale data movement across the interconnect on the right. As lithography has improved, the energy efficiency of wires has not improved as fast as the efficiency of transistors. Consequently, moving two operands just 2 mm across a silicon chip consumes more energy than the floating-point operation performed upon them. (Online version in colour.)

(b) Data movement challenges

Extracting more compute performance alone may not be sufficient to realize performance gains in future systems. A potential complication for future digital technologies is that the cost of data movement (and not necessarily compute) already dominates electrical losses, and could undermine any potential improvements in compute energy efficiency if not addressed. Since the loss of Dennard scaling in 2004, a new technology scaling regime has emerged. According to the laws of electrical resistance and capacitance, the intrinsic energy efficiency of a fixed-length wire does not improve appreciably as it shrinks in size with Moore's Law improvements in lithography, as elegantly described in Miller's articles [17,18]. By contrast, the power consumption of transistors continues to decrease as their gate size (and hence capacitance) decreases. Since the energy efficiency of transistors is improving as sizes shrink, and the energy efficiency of wires is not improving, we have come to a point where the energy needed to move data exceeds the energy used to perform the operation on those data, as shown in figure 5. This leads to extreme bottlenecks and heterogeneity in the cost of accessing data because the costs to move data are strongly distance-dependent. Furthermore, although computational performance has continued to increase, the number of pins per chip has not tended to improve at similar rates [19]. This leads to bandwidth contention, which leads to additional performance non-uniformity. The natural consequence of this technological limitation is an increased heterogeneity in data movement and non-uniform memory access (NUMA) effects so long as copper/electrical communication is used. Data locality and bandwidth constraints have long been concerns for application development on supercomputers, but recent architecture trends have exacerbated these challenges to the point that they can no longer be accommodated with existing methods such as loop blocking or compiler techniques. Future performance and energy efficiency improvements will require more fundamental changes to hardware architectures, advanced packaging approaches and new algorithm designs.

The most significant consequence of these assertions is the impact on scientific applications that run on current HPC systems, many of which codify years of scientific domain knowledge and refinements for contemporary computer systems. To adapt to computing architectures beyond

2025, developers must be able to reason about new hardware and determine what programming models and algorithms will provide the best blend of performance and energy efficiency into the future. Even our theory of complexity for numerical methods is based on counting the number of floating-point operations, which fails to account for the order of complexity of compulsory data movement required by the algorithm. Ultimately, our theories about algorithmic complexity are out of step with the underlying physics and cost model for modern computation. Future systems will express more levels of hierarchy than we are accustomed to in our existing programming models. Not only are there more levels of hierarchy, but it is also likely that the topology of communication will become important to optimize. Programmers are already facing NUMA performance challenges within the node, but future systems will see increasing NUMA effects between cores within an individual chip die in the future [15,20]. It will become important to optimize for the topology of communication; but current programming models do not express information needed for such optimizations, and current scheduling systems and runtimes are not well equipped to exploit such information were it available. Overall, our current programming methodologies are ill-equipped to accommodate changes to the underlying abstract machine model, which would break our current programming systems. There is a journal article by Unat *et al*. [21] from the PADAL (Programming Abstractions for Data Locality) workshop [22] that outlines the current state of the art in data locality management in modern programming systems and identifies numerous opportunities to greatly improve automation in these areas.

New algorithms favouring less data movement or higher arithmetic intensity, such as communication-avoiding and high-order operators, are already being developed, and data-centric programming abstractions must be built into new partitioned global address space (PGAS) programming systems in order to confer algorithmic information about data locality to the underlying software system. These capabilities are even more crucial for heterogeneous architectures where different accelerators have different memory/communication speeds. More complex algorithms increase the challenges of performance modelling, and tools such as the Roofline model need to be improved to take heterogeneity into account. Although applied mathematicians must lead the effort to re-factor core simulation and analysis algorithms, they should be working as part of collaborative teams containing algorithm, application, software, computer architecture and performance analysis expertise. Looking ahead, we expect to demonstrate algorithmic redesign of simulation algorithms that target multiple specialized architectures and refine the software prototypes to the point that they can transition to production release and adoption on leading-edge facilities.

(c) Photonics and rack disaggregation

Architectural specialization is creating new data centre requirements such as emerging accelerator technologies for ML workloads, and rack disaggregation strategies will push the limits of current interconnect technologies. While the latest high-throughput processor chips with many CPU/GPU cores are intrinsically capable of carrying out extremely demanding computing tasks, they do not have the necessary off-chip bandwidth for full and efficient utilization of their resources. In addressing this challenge, we must overcome packaging limitations—a challenge directly related to the limited bandwidth density limitations of current electrical packages. An alternative to this future is to explore co-integration of photonic technologies that do not suffer from these data movement distance constraints, such as photonic technologies. Photonic interconnect technologies have been proposed to address this critical data movement challenge because of their well-known bandwidth density and energy efficiency advantages, but system-wide energy efficiency and performance gains cannot be attained by simple photonic one-to-one replacement of existing links and switches. Observing that the in-package bandwidth densities due to the extremely high pin density enabled by copper pillar or solder microbump technologies is very well matched to photonic technologies, co-packaging of photonics as in-package devices for ‘photonic MCMs’ (multi-chip modules) has been offered as a potential approach. Whereas photonic technologies are often sold on the basis of higher bandwidth and energy efficiency

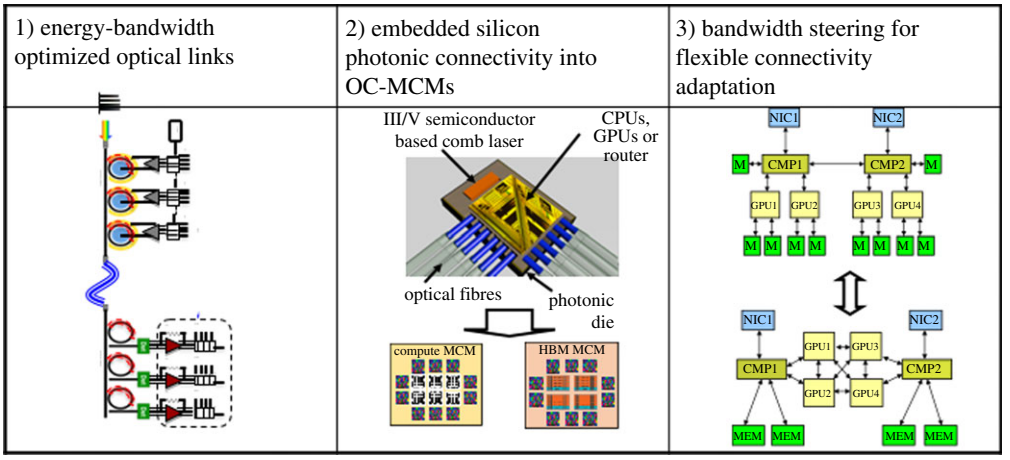


Figure 6. The three primary enabling optical technologies for system-wide disaggregation for data centres—efficient comb laser sources, photonic MCMs, and optical circuit switches for bandwidth steering to reconfigure the MCMs. System-wide resource disaggregation offers a path to co-integrating diverse technologies to support diverse workload requirements; this is being developed by industry/academic collaborations such as the ARPAe ENLITENED PINE efficient data centres project, led by Keren Bergman of Columbia University. (Online version in colour.)

(e.g. lower picojoules per bit), these emerging workloads and technology trends will shift the emphasis to other metrics such as bandwidth density (as opposed to bandwidth alone), reduced latency and performance consistency. For example, copper-based signalling technologies currently exhibit a maximum at 54 gigabits/second per wire and are struggling to double that figure—with the roadmap slipping by nearly 2 years at this point. By contrast, a single optical fibre can carry 1–10 terabits/second of bandwidth by carrying many non-interfering channels down the same path using different colours of light for each channel. This is a full 5 orders of magnitude improvement in carrying capacity for photonics in comparison to copper wires. However, such metrics cannot be accomplished with device improvements alone, but require a systems view of photonics in computing platforms.

Data centres support diverse workloads by purchasing from a limited menu of application-area-tailored node designs (e.g. big compute node, big DRAM node and big NVRAM node) and allocate resources based on instantaneous workload requirements. However, this can lead to marooned resources when the system runs out of one of those node types and is under-using other node types due to the ephemeral requirements of the workload. The 'disaggregated rack' involves purchasing the individual components and allocating the resources dynamically from these different node types on an as-needed basis across the rack [23,24]. Data centres are motivated to support this kind of disaggregation because it enables more flexible sharing of hardware resources. However, a conventional Ethernet fabric is a severe inhibitor to efficient resource sharing. Substantial increases in bandwidth density will be required.

Numerous projects have been working on using high-bandwidth-density photonics to enable this kind of system wide resource disaggregation by pumping up the off-package data bandwidths [25]. For example, PINE (Photonic Integrated Networked Energy efficient data centres) is an ARPAe ENLITENED project led by Keren Bergman of Columbia University and involving numerous industry and university partners, including NVIDIA, Microsoft, Cisco, University of California–Santa Barbara (UCSB), Lawrence Berkeley National Laboratory (LBNL) and Freedom Photonics [26,27]. The three principal elements of the project, shown in figure 6, are ultra-high-bandwidth-density (multiple terabits/second of bandwidth per fibre using a single comb laser source) links that are co-packaged with compute accelerators and memory in MCMs. This approach could revolutionize the use of resource disaggregation within the data centre to overcome the challenges of co-integrating extremely heterogeneous accelerators. 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 global IT market.

4. CMOS replacement: inventing the 'new transistor'

The development of new devices (e.g. a better transistor or digital logic technology) can greatly lower the energy consumed by logic operations. The development of the 'new transistor' will require fundamental breakthroughs in materials. The suitability for future computing devices must be evaluated in the context of circuits and full system architectures in order to determine how to make best use of those new devices and if efficiency improvements at device scale can translate into delivered improvements to applications at chip and system scale. An integral dimension of this challenge is combining these two primary paths with other promising avenues, such as three-dimensional integration and novel memory technologies, as well as packaging and integration challenges arising from new materials or technology improvements, taking information and metrology from those studies to guide the development of new post-CMOS transistor and logic technologies. A prior article written by myself and Robert Leland for the OSTP in 2013, and then re-released as an *IEEE Computer* article in 2015 [4], surveys the many different technology options that are currently available and scores those opportunities. However, Nikonov & Young [28] introduce us to the challenges of 'Boltzmann's tyranny' for electronic devices, and also illustrate quantitatively just how far these technologies are from being a clear candidate for completely replacing CMOS as we know it.

(a) Deep co-design to accelerate the pace of discovery

Typically, new electronic devices—such as new transistors or memory elements—are evaluated in isolation at a physical level, but this approach fails to capture the architectural-level impact of the device. It is essential to capture metrics that architects and system designers can use to reason about the impact of each to architectures, designs and their complex interactions with existing technologies. Existing hardware design tools do not account for the benefits, and limitations, of future devices. This creates an urgent and immediate need to efficiently and systematically explore the specialized architectural design space in combination with emerging device technologies to avoid stalling performance scaling while waiting for radical new technologies to mature. The ability to guide development of future devices requires evaluation of their performance based on ultimate outcomes for target applications. The value of new and novel materials or device technologies is not currently understood in a system context. Performance and behaviours in a system context are not currently understood in a device or materials context. True co-design to advance future systems containing novel devices and materials requires feedback that spans all layers, from atomic-scale materials to large-scale complex systems, to meet the needs of emerging scientific applications.

Only with co-design to cover this broad space and consideration of manufacturing challenges can we expect to make progress in all areas cohesively to bring about real change to the IT energy outlook. Further, the output of this work will provide a path to sustaining exponential growth in computing capabilities to enable new scientific discoveries and maintain economic vitality in all segments of the computing market (from IoT, to consumer electronics, to data centres, to supercomputing). LBNL is currently prototyping an integrated approach that spans from fundamental material discovery to architectures, circuits and full system architectures, as shown in figure 7, with the intent to dramatically accelerate the discovery process for future transistors. Our vision is to develop a co-design framework that integrates the physical layers, logical layers and control. We must propagate the quantitative information to guide development of better solutions. The co-design framework would enable us to develop unified materials/device/circuit/system electronic design automation simulation tools to ensure resilience to variability and reduce the development timeline for mission-critical science. The

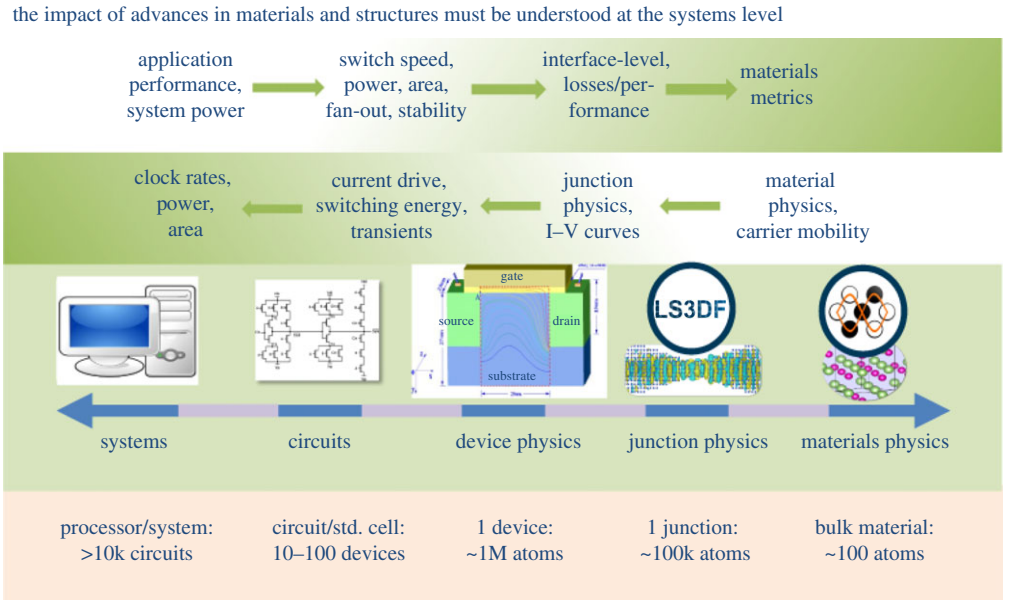


Figure 7. LBNL's prototype deep codesign framework to accelerate the discovery of CMOS replacement technologies. (Online version in colour.)

long-term solution requires fundamental advances in our knowledge of materials and pathways to control and manipulate information elements at the limits of energy flow. As we approach the longer term, we will require ground-breaking advances in device technology going beyond CMOS (arising from fundamentally new knowledge of control pathways), system architecture and programming models to allow the energy benefits of scaling to be realized. A complete workflow will be constructed, linking device models and materials to circuits and then evaluating these circuits through efficient generation of specialized hardware architectural models such that advances can be compared for their benefits to ultimate system performance. The architectural simulations that result from this work will yield better understanding of the performance impact of these emerging approaches on target applications and enable early exploration of new software systems that would make these new architectures useful and programmable.

In the longer term, we will expand the modelling framework to include non-traditional computing models and accelerators, such as neuro-inspired and quantum accelerators, as components in our simulation infrastructure. We will also develop the technology to automate aspects of the algorithm/architecture/software environment system co-design process so developers can evaluate their ideas early in future hardware. Ultimately, we will close the feedback loop from the software all the way down to the device to make software an integrated part of this infrastructure.

(b) Advanced manufacturing

To meet the goals of broad societal impact, we must ensure transition of basic research to high-volume manufacturing, and even more fundamentally reshape basic research from the start with an eye toward manufacturability. This will be achieved through the development of a new technology development capability that can evaluate and demonstrate the manufacturing and energy savings feasibility of next-generation technology options. Technologies will be rigorously evaluated for potential benefits on energy and implications on architecture and programming paradigms. The most promising technologies will be evaluated for issues around high-volume manufacturing followed by ramp-up demonstration and getting them to deliver on the energy promises. This phase will depend heavily on identifying specific manufacturing/device

materials where we will leverage the capabilities of advanced HPC capabilities to accelerate the development through modelling and 'virtual cycles of learning'. Manufacturing feasibility would also include demonstration of whatever patterning technology would be needed to support the various technologies and scaling of those technologies. Delivering on this vision will require the integration across layers of our R&D institutions and require close partnerships with industry to ensure success and economic impact.

5. Conclusion

Semiconductor technology has a pervasive role to play in future energy, economic and technology security. 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 a substantial technological shift analogous to the transition from vacuum tubes to semiconductors. This transition will require not years, but decades, so whether the semiconductor roadmap has 10 or 20 years of remaining vitality, researchers must begin now to lay a strategic foundation for change.

Data accessibility. This article has no additional data.

Competing interests. The author declares that he has no competing interests.

Funding. LBNL is supported by the Office of Advanced Scientific Computing Research in the Department of Energy Office of Science under contract no. DE-AC02-05CH11231.

Acknowledgements. I would like to acknowledge Ramamorthy Ramesh (LBNL/Berkeley), Dan Armbrust (former CEO of SEMATECH), Shekhar Borkar (Qualcomm), Bill Dally and Larry Dennison (NVIDIA Research) and Keren Bergman of Columbia University for productive discussions about the vision for the future of computing. I would also like to acknowledge the US Office of Science and Technology Policy (OSTP) and John Holdren for commissioning a report from me and Robert Leland (on loan to OSTP from Sandia National Labs) to research and write a report on 'Computing beyond Moore's Law' for them in 2013 that introduced me to many of the key technology challenges involved.

Disclaimer. The opinions expressed by the author are his own and not necessarily reflective of the official policy or opinions of the DOE or of LBNL.

References

1. Moore GE. 1965 Cramming more components onto integrated circuits. *Electronics* **38**, 33–35. (doi:10.1109/N-SSC.2006.4785860)
2. Mack C. 2015 The multiple lives of Moore's law. *IEEE Spectrum* **52**, 31–31. (doi:10.1109/MSPEC.2015.7065415)
3. Markov IL. 2014 Limits on fundamental limits to computation. *Nature* **512**, 147–154. (doi:10.1038/nature13570)
4. Shalf JM, Leland R. 2015 Computing beyond Moore's law. *IEEE Computer* **48**, 14–23. (doi:10.1109/MC.2015.374)
5. Law M, Colwell RC. 2013 The chip design game at the end of Moore's Law. *Hot Chips Symposium*. Keynote, pp. 1–16. See https://www.hotchips.org/wp-content/uploads/hc_archives/hc25/HC25.15-keynote1-Chipdesign-epub/HC25.26.190-Keynote1-ChipDesignGame-Colwell-DARPA.pdf.
6. Thompson N, Spanuth S. 2018 The decline of computers as a general purpose technology: why deep learning and the end of Moore's Law are fragmenting computing. *SSRN* abstract 3287769. (doi:10.2139/ssrn.3287769)
7. Jouppi NP *et al.* 2017 In-datacenter performance analysis of a tensor processing unit. In *Newslett. ACM SIGARCH Computer Architecture News – ISCA'17*, vol. 45 (2), May, pp. 1–12. New York, NY: ACM. (doi:10.1145/3079856.3080246)
8. Hsu J. 2016 Nervana systems: turning neural networks into a service. *IEEE Spectrum* **53**, 19. (doi:10.1109/MSPEC.2016.7473141)
9. Facebook Inc. 2017 *Introducing Big Basin: our next-generation AI hardware.* See https://code.facebook.com/posts/1835166200089399/introducing-big-basin-our-next-generation-ai-hardware/.

10. Caulfield A. 2016 A cloud-scale acceleration architecture. In *2016 49th Annu. IEEE/ACM Int. Symp. on Microarchitecture (MICRO-49), Taipei, Taiwan, 15–19 October*, 13pp. New York, NY: IEEE. (doi:10.1109/MICRO.2016.7783710)
11. Shao YS, Xi SL, Srinivisan V, Wei GY, Brooks D. 2016 Co-designing accelerators and SoC interfaces using gem5-Aladdin. In *2016 49th Annu. IEEE/ACM Int. Symp. on Microarchitecture (MICRO-49), Taipei, Taiwan, 15–19 October*, 12pp. New York, NY: IEEE. (doi:10.1109/MICRO.2016.7783751)
12. Shaw DE *et al.* 2014 Anton 2: raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer. In *SC'14: Proc. Int. Conf. for High Performance Computing, Networking, Storage and Analysis, New Orleans, LA, 16–21 November*, pp. 41–53. New York, NY: IEEE. (doi:10.1109/SC.2014.9)
13. Ohmura I, Morimoto G, Ohno Y, Hasegawa A, Taiji M. 2014 MDGRAPE-4: a special-purpose computer system for molecular dynamics simulations. *Phil. Trans. R. Soc. A* **372**, 20130387. (doi:10.1098/rsta.2013.0387)
14. Prabhakar R, Zhang Y, Koeplinger D, Feldman M, Zhao T, Hadjis S, Pedram A, Kozyrakis C, Olukotun K. 2018 Plasticine: a reconfigurable accelerator for parallel patterns. *IEEE Micro* **38**, 20–31. (doi:10.1109/MM.2018.032271058)
15. Johansen H *et al*. 2014 Software productivity for extreme-scale science. Report on DOE Workshop. See http://www.orau.gov/swproductivity2014/SoftwareProductivityWorkshopReport2014.pdf.
16. Asanovic K *et al.* 2006 The landscape of parallel computing research: a view from Berkeley. EECS Department, UC Berkeley. Technical Report No. UCB/EECS-2006-183. See http://www2.eecs.berkeley.edu/Pubs/TechRpts/2006/EECS-2006-183.pdf.
17. Miller DAB, Ozaktas HM. 1997 Limit to the bit-rate capacity of electrical interconnects from the aspect ratio of the system architecture. *J. Parallel Distrib. Comput.* **41**, 42–52. (doi:10.1006/jpdc.1996.1285)
18. Miller DAB. 2000 Rationale and challenges for optical interconnects to electronic chips. *Proc. IEEE* **88**, 728–749. (doi:10.1109/5.867687)
19. Horowitz M, Yang CKK, Sidiropoulos S. 1998 High-speed electrical signaling: overview and limitations. *IEEE Micro* **18**, 12–24. (doi:10.1109/40.653013)
20. Kogge P, Shalf J. 2013 Exascale computing trends: adjusting to the 'new normal' for computer architecture. *Comput. Sci. Eng.* **15**, 16–26. (doi:10.1109/MCSE.2013.95)
21. Unat D *et al.* 2017 Trends in data locality abstractions for HPC systems. *IEEE Trans. Parallel Distrib. Syst.* **28**, 3007–3020. (doi:10.1109/TPDS.2017.2703149)
22. Unat D, Shalf J, Hoefler T, Dubey A, Schulthess T. 2014 PADAL: Programming Abstractions for Data Locality Workshop Series. See http://www.padalworkshop.org/.
23. Meyer H, Sancho JC, Quiroga JV, Zyulkyarov F, Roca D, Nemirovsky M. 2017 Disaggregated computing. An evaluation of current trends for datacentres. *Procedia Comput. Sci.* **108**, 685–694. (doi:10.1016/j.procs.2017.05.129)
24. Taylor J. 2015 Facebook's data center infrastructure: open compute, disaggregated rack, and beyond. In *2015 Optical Fiber Communications Conf. and Exhibition (OFC), Los Angeles, CA, 22–26 March*, 1p. New York, NY: IEEE. See https://ieeexplore.ieee.org/abstract/document/7121902.
25. Tokunari M, Hsu HH, Toriyama K, Noma H, Nakagawa S. 2014 High-bandwidth density and low-power optical MCM using waveguide-integrated organic substrate. *J. Lightwave Technol.* **32**, 1207–1212. (doi:10.1109/JLT.2013.2292703)
26. Bergman K. 2018 Empowering flexible and scalable high performance architectures with embedded photonics. In *2018 IEEE Int. Parallel and Distributed Processing Symp. (IPDPS), Vancouver, BC, Canada, 21–25 May*, p. 378. New York, NY: IEEE. (doi:10.1109/IPDPS.2018.00047)
27. Michelogiannakis G, Wilke J, Teh MY, Glick M, Shalf J, Bergman K. 2019 Challenges and opportunities in system-level evaluation of photonics. In *Metro and Data Center Optical Networks and Short-Reach Links II, SPIE OPTO, San Francisco, CA, 2–7 February*, Proc. SPIE 10946. (doi:10.1117/12.2510443)
28. Nikonov DE, Young IA. 2016 Overview of beyond-CMOS devices and a uniform methodology for their benchmarking. *Proc. IEEE* **101**, 2498–2533. (doi:10.1109/JPROC.2013.2252317)