Future High Performance Computing Capabilities

Summary Report of the Advanced Scientific Computing Advisory Committee (ASCAC) Subcommittee

March 20, 2019

Contents

List of Figures

List of Tables

Chapter 1

Executive Summary

The ASCAC Subcommittee on Future High Performance Computing (HPC) Capabilities has reviewed opportunities and challenges related to the most promising technologies that are currently planned for the post-exascale (2020's) and post-Moore (2030's and beyond) timeframes. We briefly summarize below the key findings and recommendations from this review, from the perspective of planning and research directions that need to be given priority to prepare for the very significant challenges that await us in the post-Moore computing era. An overarching concern that emerged from the subcommittee's deliberations is that DOE has lost considerable momentum in funding and sustaining a research pipeline in the applied math and computer science areas that should have been the seed corn for preparing for these future challenges, and it is therefore critical to correct this gap as soon as possible. While the subcommittee understands the paramount importance of DOE's commitment to deliver exascale capabilities, we believe that it is essential for DOE ASCR to fund research and development that looks beyond the Exascale Computing Project (ECP) time horizon so as to ensure our nation's continued leadership in HPC.

Finding 1: Need for clarity in future HPC roadmap for science applications. The challenges associated with post-exascale and post-Moore computing are receiving significant attention from multiple government agencies and initiatives including DARPA, DOE, IARPA, NSF and NSCI. The subcommittee believes that Science will need to be prepared for a period of uncertainty and exploration in future HPC technologies and computing paradigms, and that, because of this uncertainty, there is a need to focus on strategy and planning activities so as to better anticipate and update, on an ongoing basis, what the future HPC roadmap possibilities will be for science applications.

Finding 2: Extreme heterogeneity with new computing paradigms will be a common theme in future HPC technologies. As discussed in the report, there is a great diversity in the technologies that are expected in the post-exascale and post-Moore eras, which has been appropriately labeled as "extreme heterogeneity" in the ASCR workshop held in January 2018 [2] and related discussions. The subcommittee believes that there is value in focusing on extreme heterogeneity as a common theme in future HPC technologies, so as to enable a broader view of post-Moore computing rather than focusing solely on point solutions.

Finding 3: Need to prepare applications and system software for extreme heterogeneity. As discussed in the report, different applications have responded to past technology transitions (e.g., from vector to MPP, terascale to petascale, petascale to exascale) in different ways. We are rapidly approaching a period of significant redesign and reimplementation of applications that is expected to surpass the disruption experienced by the HPC community when transitioning from vector to MPP platforms. As a result, scientific teams will need to prepare for a phase when

they are simultaneously using their old codes to obtain science results while also developing new application frameworks based on the results of new applied math and computer science research investments. High-quality design and implementation of these new frameworks will be crucial to the future success of DOE computational science.

Finding 4: Need for early testbeds for future HPC technologies. Given the wide diversity of technologies expected in the post-Moore era, accompanied by radically new computing paradigms in many cases, there is a need for building and supporting early testbeds for future HPC technologies that are broadly accessible to the DOE community, so as to enable exploration of these technologies through new implementations of science (mini-)applications.

Finding 5: Open hardware is a growing trend in future platforms With extreme heterogeneity, there is a growing trend towards building hardware with open interfaces so as to integrate components from different hardware providers. There is also a growing interest in building "open source" hardware components through recent movements such as the RISC-V foundation. For the purpose of this report, the term "open hardware" encompasses both open interfaces for proprietary components as well as open source hardware components. The presence of open interfaces and open source hardware components focuses, rather than restricts, the role of proprietary hardware innovation.

Finding 6: Synergies between HPC and mainstream computing Though this report has focused on future high performance computing requirements from the perspective of science applications, there are notable synergies between future HPC and mainstream computing requirements. One application area where these synergies are already being leveraged, and will undoubtedly grow in the future, is in the area of data-intensive applications and data analytics, which includes the current explosive growth in hardware accelerators for deep learning.

Recommendation 1: Office of Science's Role in Future HPC Technologies. The findings in this study have identified the urgency of developing a strategy, roadmap and plan for high performance computing research and development in the post-exascale and post-Moore eras, so as to ensure continued advancement of Science in the future. Though there are multiple government agencies that are stakeholders in post-Moore computing, the subcommittee recommends that the DOE Office of Science play a leadership role in developing a post-Moore strategy/roadmap/plan for advancing high performance computing in the service of Science.

Recommendation 2: Investing in Readiness of Science Applications for post-Moore era. The findings in this study have identified the challenges involved in preparing applications for past technology disruptions, and the fact that future disruptions will require exploration of new computing paradigms as we move to extreme heterogeneity in the post-exascale and post-Moore computing eras. The subcommittee recommends that the Office of Science work with other offices of DOE to ensure that sufficient investment is made with adequate lead time to prepare science applications for the post-Moore era. While the adaptations that ECP application teams are starting to make for supporting current and emerging heterogeneous execution environments is good preparation for some of the anticipated post-exascale technologies, additional investments will be needed to explore the newer computing paradigms that will emerge in the post-exascale and post-Moore timeframes. In addition, we recommend that R&D in best practices for design and development of scientific software be given high priority to best assure that new scientific application frameworks benefit from the state of the art in software best practices.

Recommendation 3: Investing in Research related to Platforms with Open Hardware interfaces and components. The findings in this study have identified a growing trend in the use of open hardware interfaces and components in the post-exascale and post-Moore eras, relative to current and past approaches for hardware acquisition. In the interest of future Science needs, the subcommittee recommends that the Office of Science foster this ecosystem by investing

in research related to open hardware platforms, i.e., platforms built using open interfaces that support high-performance and reliable integration of open hardware components with proprietary components from different hardware providers.

Recommendation 4: Investing in Research related to System Software. The findings in this study have identified the need for advancing system software to meet the requirements of post-Moore computing. The DOE should support active and sustained efforts to contribute to relevant software projects to ensure that HPC concerns such as performance isolation, low latency communication, and diverse wide area workflows are addressed in the design and adoption of system software for future HPC platforms.

Recommendation 5: Early Testbeds in DOE Computing Facilities. The findings in this study have identified the need for providing users of DOE computing facilities early access to testbeds and small-scale systems that are exemplars of systems expected in the post-Moore computing roadmap. The subcommittee recommends that the Office of Science's computing facilities address this need by acquiring such testbeds and small-scale systems, and providing and supporting access to these systems by current HPC users. The investments in Recommendations 2, 3, 4 will help create a community of researchers that can assist computing facilities staff in training activities related to these early testbeds.

Recommendation 6: Recruiting, Growing and Retaining Talent for the post-Moore era. The findings in this study have identified the need for significant innovation in support of the enablement of science applications on post-Moore hardware. The subcommittee recommends that DOE national laboratories prioritize the recruiting and nurturing of top talent in all aspects of mapping applications onto emerging post-Moore hardware, including skills and talent related to development of science applications, applied mathematics research, system software research, and hardware research for future platforms.

Chapter 2

Background

2.1 Moore's Law and Current Technology Roadmaps

Moore's Law [3,4] has been the bedrock for growth in the capabilities of all computing systems, including high performance computing (HPC) systems. Simply stated, Moore's Law is the prediction that the number of transistors (components) in an integrated circuit would double approximately every two years. The significance of Moore's Law is that the semiconductor industry has strived to maintain this exponential growth for over five decades, resulting in unsurpassed benefits in cost and performance for all semiconductor consumers. The cost implication of Moore's Law is that if the cost of an integrated circuit remains approximately constant, then the cost per transistor decreases exponentially with time. The performance implication of Moore's Law was historically tied to Dennard Scaling [5], which stated that, as transistors become smaller, their power density remains constant, i.e., the power consumed by an integrated circuit remains proportional to the area of the circuit rather than the number of transistors in the circuit. An underlying assumption in the Dennard Scaling prediction is that the power consumed by an integrated circuit is dominated by its dynamic (switching) power, which in turn is proportional to the clock frequency and the square of the operating voltage. As a result, when Dennard Scaling holds, the power per transistor decreases exponentially with time, which in turn made it possible to increase clock frequencies from generation to generation of a semiconductor technology without increasing the total power consumed by the integrated circuit.

One of the major challenges recently faced by the computing industry is the fact that Dennard Scaling ended over a decade ago, as shown in Figure 2.1, which includes trend data for microprocessors built during the last 40 years. (Note that the y-axis numbers are plotted on a logarithmic scale.) The first observation from the figure is that Moore's Law has remained robust during this period, since the number of transistors in a microprocessor continued to increase at an exponential rate until the present time. However, the clock frequencies flattened in the 1 GHz ($= 10^3$ MHz) range since around 2005, thereby signalling the end of Dennard Scaling. The two main reasons for this end were that the operating voltage for the transistors could not be lowered any further, and that the leakage power started becoming a significant component of the power consumed by transistors, as the transistor sizes decreased. Past 2005, any attempt to increase clock frequency became impractical because doing so would cause the chip to overheat. Instead, 2005 marked the start of the "multicore era" in which the additional transistors predicted by Moore's Law are being used to increase the number of processor cores in a single integrated circuit, without increasing their clock frequencies.

If Moore's Law were to continue indefinitely, we could continue getting more performance from

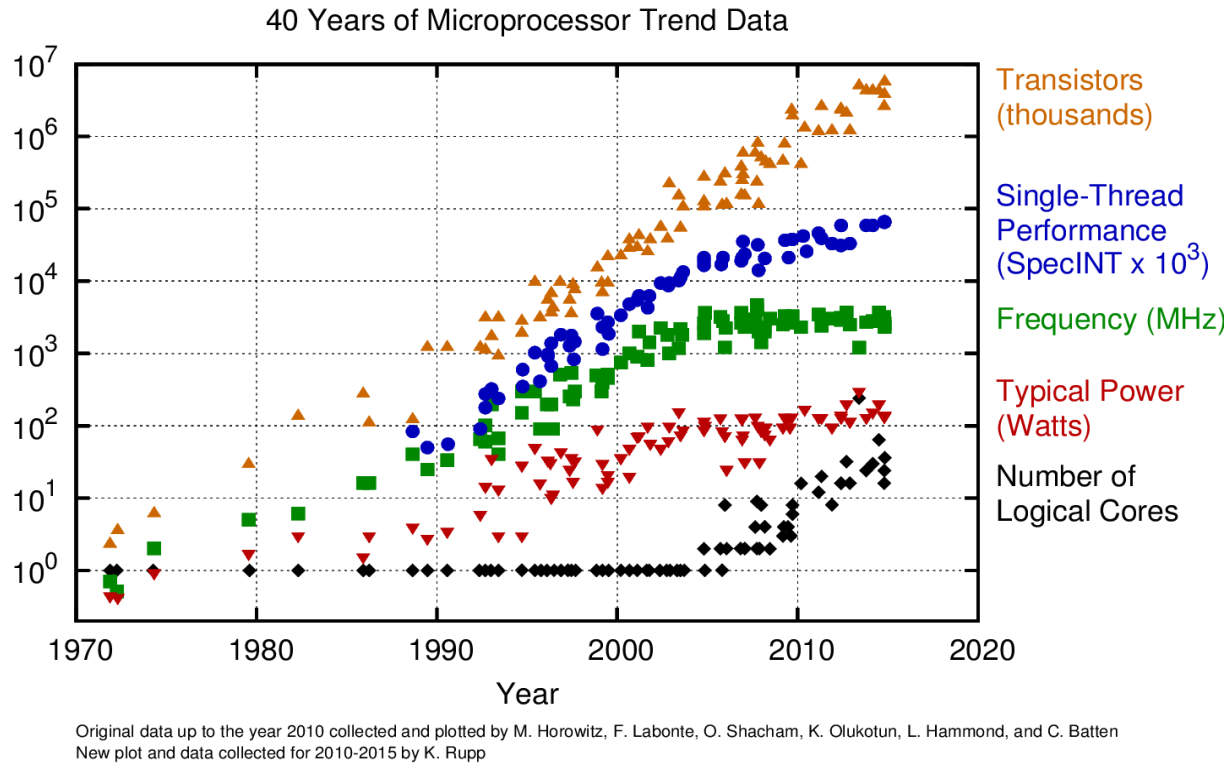


Figure 2.1: 40 years of Microprocessor Trend Data for 1) Number of Transistors, 2) Single Thread Performance, 3) Frequency, 4) Power, 5) Number of Cores.

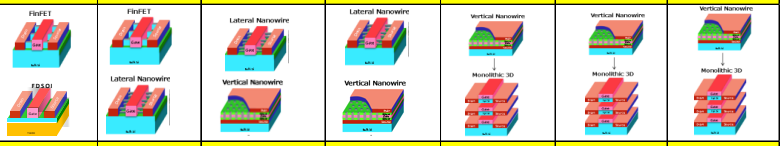
Table MM01 - More Moore - Logic Core Device Technology Roadmap							
YEAR OF PRODUCTION	***2017***	***2019***	***2021***	***2024***	***2027***	***2030***	***2033***
	P54M36	**P48M28**	**P42M24**	**P36M21**	**P28M14G1**	**P26M14G2**	**P24M14G3**
Logic industry "Node Range" Labeling (nm)	**"10"**	**"7"**	**"5"**	**"3"**	**"2.1"**	**"1.5"**	**"1.0"**
IDM-Foundry node labeling	**i10-f7**	**i7-f5**	**i5-f3**	**i3-f2.1**	**i2.1-f1.5**	**i1.5-f1.0**	**i1.0-f0.7**
Logic device structure options	**finFET FDSOI**	**finFET LGAA**	**LGAA VGAA**	**LGAA VGAA**	**VGAA M3D**	**VGAA M3D**	**VGAA M3D**
Logic device mainstream device	**finFET**	**finFET**	**LGAA**	**LGAA**	**VGAA**	**VGAA**	**VGAA**
Logic device technology naming							
Patterning technology inflection for Mx interconnect	**193i**	**193i, EUV**	**193i, EUV**	**193i, EUV**	**193i, EUV**	**193i, EUV**	**193i, EUV**
Channel material technology inflection	**Si**	**SiGe25%**	**SiGe50%**	**Ge, IIIV (TFET)**	**Ge, IIIV (TFET)**	**Ge, IIIV (TFET)**	**Ge, IIIV (TFET)**
Process technogy inflection	**Conformal deposition**	**Conformal Doping, Contact**	**Channel, RMG**	**CFET**	**Seq. 3D**	**Seq. 3D**	**Seq. 3D**
Stacking generation	**2D**	**2D**	**2D 3D: W2W or D2W**	**3D: P-over-N**	**3D: SRAM-on-Logic**	**3D: Logic-on-Logic, Hetero**	**3D: Logic-on-Logic, Hetero**
Design-technology scaling factor for standard cell	**-**	**1.11**	**2.00**	**1.13**	**0.53**	**1.00**	**1.00**
Design-technology scaling factor for SRAM (111) bitcell	**1.00**	**1.00**	**1.00**	**1.00**	**1.25**	**1.00**	**1.00**
Number of stacked devices in one tier	**1**	**1**	**3**	**4**	**1**	**1**	**1**
Tier stacking scaling factor for SoC	**1.00**	**1.00**	**1.00**	**1.00**	**1.80**	**1.80**	**1.80**
Vdd (V)	**0.75**	**0.70**	**0.65**	**0.60**	**0.50**	**0.45**	**0.40**
Physical gate length for HP Logic (nm)	**20.00**	**18.00**	**14.00**	**12.00**	**10.00**	**10.00**	**10.00**
SoC footprint scaling node-to-node - 50% digital, 35% SRAM, 15% analog+IO	**-**	**64.9%**	**51.3%**	**64.3%**	**64.2%**	**50.9%**	**50.7%**

Table 2.1: Projections for the continuation, and end, of Moore's Law during the next 15 years (Source: IEEE IRDS 2017 Edition).

successive generations of semiconductor technology by doubling the number of processor cores in an integrated circuit rather than by increasing the clock frequency. However, it stands to reason that Moore's Law must come to an end due to basic physical limitations, including the fact that the size of the atoms used in silicon chip fabrication is around 0.2nm. Table 2.1 shows the projected transistor size ("node range") decreasing from 10nm in 2017 to 1.0nm in 2033, at which point a single transistor would shrink to the size of five Silicon atoms. Further, achieving the reductions shown in Table 2.1 will require major technology advances, including monolithic 3D transistors expected from 2024 onwards. It is therefore clear that alternate computing technologies and paradigms urgently need to be explored for future HPC, to ensure the continued and sustained performance gains to which HPC users and customers are accustomed. Given this context, we will refer to the 2020's decade as "post-exascale" and the 2030's decade and beyond as "post-Moore" in this report.

2.2 Levels of Disruption in Post-Moore era

The IEEE Rebooting Computing Initiative [1] has characterized a range of possible approaches to address the end of Moore's law. As shown in Figure 2.2, these approaches can be classified in terms of the amount of disruption to the computing stack they would require [1].

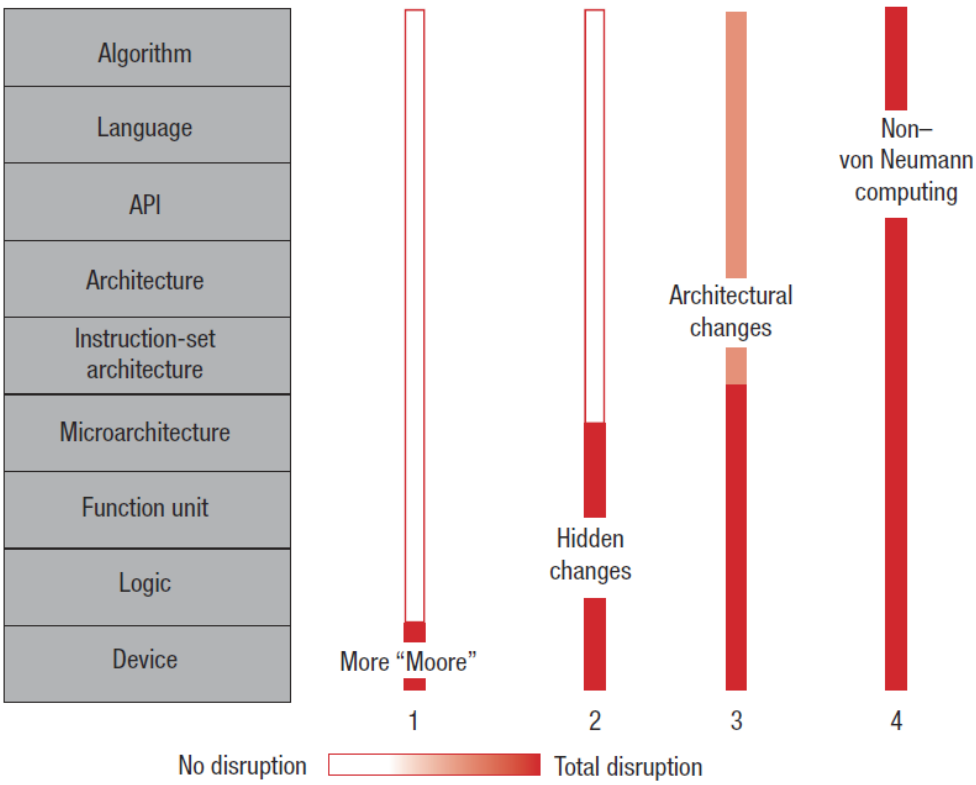


Figure 2.2: Levels of disruption in the computing stack, from [1].

The least disruptive approach in Figure 2.2 is for the industry to find a drop-in replacement for the CMOS switch. Existing transistor technologies cannot be both power efficient and operate reliably at the scales at the end of the roadmap. Thus, this approach is to create a new transistor technology. Although this is the least disruptive approach to the computing stack, it is exceedingly

challenging. The IRDS roadmap shows that Moore's Law will run out even with these new transistor types by 2033 [6].

The next least disruptive approach is to use novel ways to construct computer microarchitectures while still maintaining software compatibility to the existing software base. These include microarchitectures implemented using techniques such as some Silicon Photonics approaches (Section 4.3). Other approaches not discussed in this report include adiabatic/reversible logic and cryogenic/superconducting logic.

The next disruptive approach involves making architectural changes that are "programmer visible." Where these approaches will require new programming systems., they generally do not abandon the von Neumann computing paradigm. These approaches include Reconfigurable Logic (Section 4.1), Memory-Centric Processing (Section 4.2), and some approaches that employ Silicon Photonics (Section 4.3), all of which are promising approaches for the post-exascale era.

The most radical (Level 4) approaches rethink computing paradigms from the ground up, and will require new algorithms, programming systems, system software, and hardware. Examples of this include Neuromorphic Computing (Section 4.4), Quantum Computing (Section 4.5) and Analog/Thermodynamic Computing (Section 4.6). All of these represent potential candidates for the post-Moore era.

2.3 National Landscape for Post-Moore Computing

Leadership in HPC is critical to the success of many federal agencies, as well as that of many commercial enterprises; all these players are concerned about what the future portends beyond the end of Moore's Law. Many are investing, or planning to do so, and there is an opportunity for DOE to coordinate its efforts with them, so as to maximize the benefit to all. Where serious sustained investments are being made, DOE need not duplicate them but can instead leverage their synergies. IARPA is investing in both specialized analog quantum systems (QEO) and the foundations of general purpose devices (LOGIQ). A recent NSF Expeditions project, EPiQC, is focused on advancing algorithms, software, and physical machines for quantum computing. In general, quantum computing is receiving significant new attention, in part due to USA's National Quantum Initiative Act which became public law in December 2018.

DARPA MTO kicked off the Electronics Resurgence Initiative (ERI) in 2018; many of the programs in this initiative are focused on embedded computing and data analytics, which are areas where there may be synergistic benefits with DOE's needs for advancing science applications and analysis of experimental data. MTO is also investing in HPC related technologies such as hybrid analog and digital systems (ACCESS), design automation (CRAFT), IP reuse (CHIPS), integrated photonics (POEM), and energy efficiency (PERFECT). IARPA is also exploring superconducting logic as a basis for classical computing (C3). And, of course, many commercial enterprises are investing in the development of special-purpose accelerators for deep learning and related AI algorithms and applications. Accelerating and advancing AI applications is also a major focus of the recent (February 2019) Executive Order on Maintaining American Leadership in Artificial Intelligence.

A key point underlying all the activities under way in other agencies and commercial entities is that, while they may not be directly working on advancing HPC for science applications, they are investing in technologies that could be highly relevant to DOE's future HPC roadmap for science. It is also worth noting that the NSCI has designated DOE as playing the leadership role for HPC. Therefore, DOE has a unique opportunity to not only explore the future of HPC for scientific leadership, but to also determine if the broader HPC technology investments in the US government

are adequate to enhance and sustain the economy and security of the US as has been done by past investments in computing technologies.

2.4 International Landscape for Post-Moore Computing

In late 2013, IEEE launched the international IEEE Rebooting Computing Initiative (IEEE RCI) to begin to look at potential post-Moore computing possibilities [1]. Since that time, IEEE RCI has held four invitational summits of thought leaders across multiple fields. The IEEE RCI sponsors the annual International Conference on Rebooting Computing (ICRC), starting in 2016 as its inaugural year. ICRC attracts researchers from around the globe to share their latest research on post-Moore computing directions.

In 2016, the Semiconductor Industry Association pulled its sponsorship for the venerable International Technology Roadmap for Semiconductors (ITRS). IEEE moved swiftly to become the new sponsor of the roadmap. To give the roadmap a post-Moore viewpoint, two new focus teams were added, one to track application performance and one to track architectural ideas. The IEEE renamed the roadmap the *International Roadmap for Devices and Systems* (IRDS) to stress the changing nature of the industry towards post-Moore technology considerations [1]. IRDS partner organizations include the Japan Physics Society's *Systems and Devices Roadmap for Japan* (SDRJ) and the EU's *NanoElectronics Roadmap for Europe: Identification and Dissemination* (NEREID). IRDS produced a roadmap at the end of 2017 and will continue the ITRS' historic cadence of a new roadmap every two years, with a roadmap update in the intervening years [6].

Finally, subcommittee members are aware of recent announcements from China, Europe and Japan related to Quantum Computing and Neuromorphic Computing that foretell a high level of international competitiveness in the post-Moore Computing era.

2.5 Interpretation of Charge

The subcommittee appreciated the timeliness of the charge, a copy of which is included in Appendix A. At the same time, we acknowledge that a single study cannot provide a comprehensive answer to identifying research opportunities and challenges for future HPC capabilities in the post-exascale and post-Moore timeframes, which span multiple decades. We trust that there will be follow-on studies to elaborate further on these challenges and opportunities as details of emerging HPC technologies become clearer in the coming years. To focus our efforts in this study, we made the following two assumptions when interpreting the charge:

- There are multiple Federal government initiatives and programs in the early stages of addressing the challenges of post-Moore computing. The subcommittee explicitly restricted the scope of this study to considerations pertinent to the use of computing for the advancement of Science, thereby focusing on the Office of Science's mission needs, while still identifying synergies with strategic needs of other government agencies and commercial endeavors.

- The charge did not specify a timeframe to be assumed for our recommendations, though it was clear that the charge refers to timeframes that follow the accomplishment of exascale capability in the DOE. The subcommittee concluded that it was appropriate to focus on different timeframes for different technologies, based on their anticipated levels of readiness. These timeframes include the post-exascale (2020's) and post-Moore (2030's and beyond) eras mentioned earlier.

Chapter 3

Application lessons learned from past HPC Technology Transitions

3.1 Background

All HPC technology transitions have focused on new algorithm and application designs that expose concurrency and locality at different levels. The advent of vector supercomputers such as the Control Data Cyber 205 and Cray 1 were notable early examples. Application developers organized data and computations to expose unit stride memory accesses and conflict-free writes that could be written as Q8 function calls on the Cyber 205, or converted to Cray vector instructions by the compiler. Clock speed improvements and improved functional parallelism (simultaneous execution of instruction streams) were important for performance improvements from one generation of machines to another, and had the advantage of not forcing substantial application refactoring to realize those benefits.

Disruptive transitions occurred when the fundamental strategy for organizing data and computations changed. Vector supercomputing applications represented the first large body of optimized applications where the data and computation strategies were specialized to match a particular parallel computing model. Multiprocessing vector computations were also important, but few codes were explicitly organized to exploit multiple vector processors, relying instead on shared memory fork-join models that required minimal code modifications. The first Gordon Bell Prize was given for an auto-tasked, vectorized version of a multifrontal, super-nodal sparse direct solution on an 8-processor Cray Y-MP, but, practically speaking, the best use of multiple Cray vector processors was to improve job throughput of single processor vector codes.

3.2 Vector-MPP Transition

The large body of vector HPC applications developed in the 1980s and early 1990s represented a valuable collection of HPC capabilities. Cray systems were available long enough to allow the HPC community an opportunity to create a large number of highly optimized codes for defense, engineering, weather, chemistry, oil & gas applications, and more. Many of these codes were large and full-featured. The arrival of Massively Parallel Processing (MPP) computers, which relied on a very different data and computation organization, represented a challenge to developers of vector applications. There was no incremental transition path from a shared memory vector design to a distributed memory MPP design.

Many vector codes did not make the transition to MPP. For those that did, the most successful

transitions started by first designing a new application framework specifically for distributed memory. Typically the framework partitioned logically global objects such as grids for PDE calculations into distributed subgrids with halos, and then provided halo exchange functions that would update halo values when called. The framework also provided reduction operations such as distributed dot products.

Given such frameworks, most of the computations that were part of the vector code could be migrated to the new paradigm with minimal changes. Assuming the halo exchange operation had been called to exchange remote values, halos enabled most computations to work with local data, just as before. Local reductions just needed a single new step to compute the global reduction. It is also worth noting that vectorization was not important for early MPPs. Maintaining vectorizable code is difficult because its presence is ubiquitous across loop nests, and often requires special design considerations. Without regular testing, vectorization impediments were introduced as new features are added to the code. In most MPP codes, vectorization features were not maintained, and eventually removed, especially as cached data access became more important.

The transition from vector computing to MPP was challenging because constructing the new MPP framework took substantial time (months or years), during which the previous vector code had to remain the production platform, and the development team was split across two codes. Many vector codes were eventually retired as new MPP codes emerged.

It is worth noting that, in the transition from vector to MPP, we took advantage of the inherent disruption to introduce advancements in modeling. MPPs offered greater computing and memory capacities that in-turn enable higher fidelity modeling and simulation. We see the same dynamics occurring now. For example, many ECP application efforts are focused on improved multi-scale, multi-physics or ensemble computations that are qualitatively different from current capabilities.

3.3 Terascale-Petascale Transition

The Terascale to Petascale transition has been less disruptive overall. For most applications this transition was incremental in the sense that the MPP framework continued to be applicable. Certainly, the framework had to be refined and scalability bottlenecks removed, as the number of distributed processors and the partitioning of data increased. But there was no disruptive ramp-up phase as was the case in the vector to MPP transition.

The path to Petascale included the introduction of intra-process parallelism, e.g., use of OpenMP threading, use of GPU accelerators, and a renewed focus on exposing vectorizable code to compilers. But these features did not force a complete redesign for most codes. Instead, application developers had to incrementally refactor the most important computational kernels to run well and could leave much of the remaining code untouched. One notable exception was the disruptions incurred for migrating applications to the petascale RoadRunner computer, which were more extensive than for other (later) petascale systems. However, it can also be argued that the application changes needed for the RoadRunner system may have served as good preparation for the multi-GPU on-node parallelism (as an example) that needs to be exploited on exascale systems.

The approach used for the terascale to petascale transition continues to be very effective, even as we go beyond petascale. It was the primary strategy used to port applications to the Sunway TaihuLight, the fastest LINPACK machine in 2017. This system has thousands of distributed memory nodes that can be used as a large Linux cluster by mapping execution to just the Management Processing Elements (MPEs). Porting any MPP code to the MPE processors of the TaihuLight platform is very straightforward if the code is designed to run on scalable Linux clusters. The performance of the initial port can be very poor, since the MPEs represent a tiny fraction of the system

performance. But once the code is working on the MPEs, incremental porting of functionality to the CPEs (8x8 processor mesh) is possible, and is very similar to porting strategies use for GPU offloading. Certainly, very substantial data structure and execution strategy changes are required, but again an incremental approach is possible.

3.4 Petascale-Exascale Transition

The petascale to exascale transition is currently under way. So far, the terascale to petascale approach is working well as a starting point for the petascale to exascale transition. At the same time, the applications that have been successful using this approach are typically highly structured and compute-intensive, but have still not achieved uniformly high performance across all the problem formulations that they are designed to handle. Furthermore, they are not prepared for simultaneous heterogeneous execution, where subproblem sizes must vary to tune for optimal performance on different processor types, nor is there sufficient on-node control of data partitioning and mapping, or concurrent execution of heterogeneous tasks.

Another concern is resilience. With each new factor of 1000× performance improvement (tera, peta, and now exa) seems to come increased concern about the ability of computer system designers to preserve the illusion for application developers that they are using a "reliable digital machine." This same concern arose as we started preparations for exascale, but as we approach the arrival of exascale platforms, the general belief is that application developers need not worry about additional reliability concerns in exascale, relative to petascale approaches. Even so, we continue to monitor system reliability and believe that post-exascale computing plans should include efforts for application-level resilience, and the software stack R&D needed to support applications in this effort. As it becomes increasing expensive in funds, time and effort to create reliable leadership platforms, investing in application-level resilience could very well contribute to new cost-effective ways to continue scientific advancement with the latest computing technologies.

In order to bring a full portfolio of applications to the exascale threshold, and to bring all applications forward beyond exascale, we face another disruptive phase. The growth of on-node concurrency, the need to execute concurrently on multiple heterogeneous nodes, and the increasing penalty for having any sequential execution regions in our codes indicate that we are on the front end of a new transition. While there is much research required, early indications are that we need to introduce new control layers and system software support (e.g., pervasive support of asynchronous tasking and data movement), that will enable us to better handle simultaneous heterogeneous execution, support task-enabled functional parallelism and latency hiding, and move toward an effective strategy for implementing application-level resilience capabilities.

3.5 Lessons Learned

A summary of some of the key lessons learned from the three transitions summarized above is as follows:

- Vector-MPP: Investing in new application frameworks, built using results from related Applied Math and Computer Science research, was critical for success in this transition.
- Terascale-Petascale: Leveraging incremental approaches to application migration can be extremely valuable, when possible to do so.

- Petascale-Exascale: Investing in new control layers and system software support (e.g., for asynchronous heterogeneous tasking and data movement) is helpful for addressing the disruption of large on-node heterogenous parallelism.

The HPC community has been gaining experience with increasingly diverse computing architectures. Heterogeneous architectures, first broadly encountered with attached GPUs, and now present on the Summit, Sierra, and Trinity platforms have exposed application developers to the demands that we must address. In particular, our application designs and base implementations must lend themselves to rapid adaptation to new node architectures and flexible execution models. Use of discrete devices has also taught us important lessons of shipping computation to data and managing remote resources.

In addition, code teams are migrating to new languages as opportunities arise. For example, several Exascale Computing Project codes that were formerly Fortran or C based, e.g., NWChemEx and SLATE, have moved to C++. Teams report that C++ enables more rapid code development and improved adaptability; many programming model research projects now offer C++ library interfaces as a primary parallel programming interface for scientific application developers.

Even so, we have much to learn about software design. Porting existing codes to new platforms can require a monumental effort, or can be designed into the code. An example of the former is the recent Gordon Bell finalist paper on porting the DOE climate CAM-SE dynamical core code to TaihuLight [7]. The authors reported that the effort required modification of 152,336 of the original 754,129 lines of code (20%), and the addition of 57,709 new lines (8% increase). While this porting effort was incremental, it is still very expensive. In contrast, the Uintah application [8] is coded using C++ with template meta-programming techniques that enable compile time mixing of platform-specific adaptations to general parallel pattern expressions. This approach enables support of many node types from the same source, including simultaneous heterogenous execution on more than one type.

3.6 Assessing Application Readiness

The lessons learned from past technology transitions confirm that mapping applications to new platforms can be costly and risky. Most computational scientists are focused primarily on the new scientific insights that can be achieved through computation. Combined with the competition to produce new scientific results on a regular cadence, few computational scientists are prepared to take on the risk of migrating applications to new computing paradigms, unless absolutely necessary.

We briefly present an exemplar scorecard framework to illustrate how application readiness can be assessed for new computing platforms and paradigms. Table 3.1 lists attributes can be used to assess and prioritize scientific problems that would be good early targets for different kinds of future HPC systems. A high rating in all areas indicates strong likelihood of success as an early adopter. The contents of the table include a simple illustration using sparse linear solvers as a target problem.

3.7 Next Steps

We believe that recent experiences with preparing applications for emerging heterogeneity will also help with preparations for some of the post-exascale technologies in Chapter 4, though new challenges remain for post-Moore technologies. A good resource for any software refactoring effort is the book entitled “Working Effectively with Legacy Code” by Michael Feathers [9]. This book

Problem: Large sparse linear systems on von Neumann (vN) + accelerators/ interconnect/memory-centric. (Non von-Neumann notes)		Score
Potential	Opportunities for R&D are numerous for all vN+accel, interconnect and memory centric. (Non-vN options are possible, but appear to have lower potential.)	High
Readiness	Current algorithms, with adaptations that are underway already, are suitable for vN, interconnect and memory centric. (Fundamentally new approaches are needed for non-vN.)	High
Novelty	Many known approaches that can be explored first. (There are potential algorithms for non-vN architectures. Solution of real valued systems can be recast in the complex field for use with at least one known quantum algorithm. ML-based approaches could be a suitable replacement for a linear solver, at least to a coarse level approximation.)	Medium
Demand	Linear solvers remain an important enabling capability for many scientific problems. On vN, interconnect and memory centric, funding for new algorithms (which will typically be incremental) is important.	High
Feasible	Adaptations to all vN technologies are feasible with adequate resourcing.	High
Total Rating	Overall possibility that this is a high priority research direction.	High

Table 3.1: This table shows a simple illustration using sparse linear solvers as the target problem. For all von Neumann technologies, this is a good target problem. For non-von Neumann architectures, linear solvers do not have a clear mapping. In fact, alternative algorithms are most likely required, or the need to solve a linear system may be bypassed completely.

provides a practical step-by-step approach to planning and executing changes in an existing code. Fundamental to the effort is covering the code that will be refactored with adequate regression testing. The scope of change should be incremental when possible, making sure that one change set is fully integrated and tested before starting the next.

Of course, the disruptive transition required to introduce a tasking control layer and supporting system software between the current MPI and low-level threading and vectorization layers cannot be easily partitioned for incremental changes. Even so, Feathers' basic strategy can guide part of the approach. In addition to Feathers' recommendations, we need to use the same basic approach that succeeded when moving from vector to MPP codes. We need to first construct a new framework that includes only a minimal representative subset of the application's functionality. Then we construct the new framework to include the MPI (SPMD) and threading/vectorization layers of the old application, and a new task control layer in between the two. Proper design and implementation of these new frameworks is essential, and will impact scientific developer productivity and software sustainability. Adequate investment in R&D of best practices for scientific software is essential, and should be on an equal footing with R&D in other Office of Science research areas.

Despite some promise from initial efforts to introduce tasking, there are many research questions that must be addressed. Examples include what new mathematical formulations expose better computation intensity, how we can realize the potential of asynchronous execution in the presence of deep memory hierarchies that further penalize remote data accesses, how to effectively schedule fine grain dynamic workloads with locality considerations, and how to write software that is easily adapted to a variety of heterogeneous processors. Furthermore, the disruptive change that this effort requires (similar to the vector-to-MPP transition in the 1990s) will be experienced across

the entire DOE application portfolio. Over time, asynchronous tasking (for computation and data movement tasks) may become a replacement for message passing. A task-based model can provide a more expressive and flexible environment for parallel execution, especially for applications that have rapidly changing dynamic workloads.

DOE has a very large parallel scientific software base. Transforming this base to exploit post-exascale and post-Moore systems will be disruptive and require a signifant investment. Applied math and computer science research will inform when and how to proceed. Better software design and practices will enable productivity and sustainability improvements; improved modeling, simulation and scientific insight will be the reward. The migration path, and when to embark on it, will vary for each application area, and is best executed as a collaborative effort among computational scientists, computer scientists and applied mathematicians, informed by modern software design and development practices.

Chapter 4

Future HPC Technologies: Opportunities and Challenges

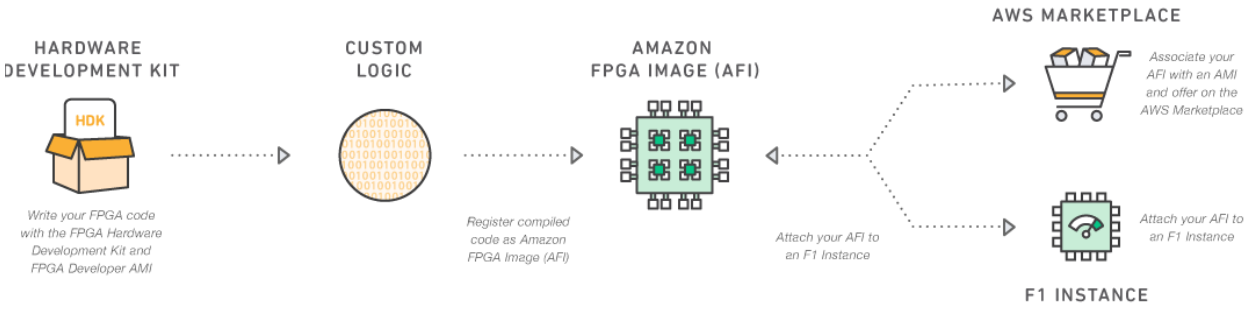
In this chapter, we provide a summary of six major technologies (Chapters 4.1–4.6) that the subcommittee felt were most representative of the trends expected in future HPC systems, based on our current knowledge. While there are some natural omissions in this list (e.g., application-specific computers like Anton 2 [10], 3D chips or 3D stacks of chips, or computing with carbon nanotube transistors [11]), our belief is that the general findings and recommendations that were derived from studying these six technologies will apply to other future HPC technologies as well. We conclude the chapter with a discussion of application challenges related to the new technologies, as well as opportunities arising from the growing trend towards building systems with open hardware interfaces and open hardware components.

4.1 Reconfigurable Logic

Application-specific acceleration hardware mapped onto Field Programmable Gate Arrays (FPGAs) offers a low-power, high performance option for exascale and post-exascale computing. Though the primary use of these devices was general purpose glue logic between ASICs, reconfigurable computing with FPGAs has been pursued for almost three decades [12], [13]. Over this period of time, FPGA architectures have evolved to complex systems on chip, including embedded processors, on-chip reconfigurable memory, network interfaces, DSP arithmetic blocks, and millions of system gates to hold arbitrary application-specific logic. For some application kernels, FPGAs can offer two orders of magnitude performance improvement over general purpose processors.

Research into reconfigurable computing was supported in part by the DARPA Adaptive Computing Systems program, which led to the design of coarse grained reconfigurable architectures such as PipeRench [14] from CMU, RAW [15] from MIT, and MorphoSys [16] from UC Irvine. Coarse grained architectures have primarily 8-16 bit data paths and function units in contrast to fine grained FPGAs with bit level resources. RAW was commercialized as the Tilera chip. Other commercial coarse grained reconfigurable architectures that have come and gone included MathStar [17] and Ambric [18]. The Tensor Processing Unit [19] from Google is a recent example of a coarse grained reconfigurable architecture specialized for neural network processing. While general purpose coarse grained architectures have not been stable in the marketplace, FPGAs remain highly successful commercial offerings with architectures suitable for a wide range of applications, including, for some large FPGAs, high performance computing.

Despite successful demonstration of many applications on FPGAs, interest in reconfigurable



DEVELOP

Develop custom Amazon FPGA Images (AFI) using the Hardware Development Kit (HDK) and full set of design tools and simulators.

DEPLOY

Deploy your AFI directly on F1 instances and take advantage of all the scalability, agility, and security benefits of EC2.

OFFER

Offer AFIs you design on the AWS Marketplace for other customers.

PURCHASE

Purchase AFIs built and listed on AWS Marketplace to quickly implement common hardware accelerations.

Figure 4.1: Growing an ecosystem for Amazon EC2 F1 FPGA instances (image source: `https://aws.amazon.com/ec2/instance-types/f1`)

computing for HPC declined in the last decade with the advent of GPGPUs, which were capable of many factors of performance improvement over CPU at a fraction of the cost of high end FPGAs, and a considerably easier application development cycle [20]. Recently however, the drivers of improved performance per watt and better memory bandwidth utilization has resulted in a renewed interest in reconfigurable computing elements in exascale and post-exascale architectures.

Applications exploiting FPGAs can be found in bioinformatics (sequence alignment such as Smith Waterman or Needleman-Wunsch), signal processing, image processing, and network packet processing [21] domains. Of these, signal and image processing continue, especially in deployed platforms, and network packet processing has grown. The latter has been adopted in the finance sector [22] to enable microsecond turnaround by processing the packet payload on the network interface without having to make a round trip through the CPU. Database acceleration, data analytics for search engine applications and genomics have also been pursued, often in the context of hardware appliances. In scientific computing, recent algorithmic studies investigating the impact of reduced precision arithmetic on numerical stability are particularly relevant to reconfigurable logic that can support custom floating point formats [23].

The slow adoption of FPGAs for general purpose application acceleration has been principally due to the difficulty of mapping algorithms to hardware. For maximum performance, key kernels are written in Hardware Description Language (HDL), which requires hardware design expertise and has a much longer development cycle than software. High Level Synthesis (HLS) of C, C++, or OpenCL [24] continues to improve in quality of generated hardware and synthesizable subset of the language. However, performance gain may diminish considerably when HLS is employed. Additionally, the compile cycle (synthesis, map, place, and route) can take hours to days for large FPGAs and complex designs. Recent investments in the DARPA ERI Software-Defined Hardware program may pay off with new algorithms and techniques to speed up HLS for ASICs, FPGAs,

and coarse grained reconfigurable architectures.

Factors that improve the prospects for reconfigurable computing with FPGAs in the exascale to post-exascale timeframe include:

- increased urgency to reduce power while increasing compute capability;
- improvements in design tools and access to design tools through the Amazon "free" design tool model (see below);
- increases in availability of open source hardware Intellectual Property (IP) libraries;
- federal research investments in design tools;
- cloud-based application kernels and libraries from third party sources;
- integration of data analysis with simulation; and,
- workflows that can exploit in-transit data processing.

Technology Readiness Timeframe: FPGAs are available today and with the Intel acquisition of Altera, it is anticipated that the integration of CPU with reconfigurable logic will grow even closer in the next 2-5 years. Early adoption in the data analysis and in-transit processing areas are most promising: for example, using reconfigurable logic to compress, clean, filter data streams generated by instruments [25].

Recently, FPGAs have become available in cloud computing servers, as illustrated by Amazon's F1 FPGA option for compute nodes (Figure 4.1). In the Amazon business model, application developers can create FPGA applications for the F1 in the Amazon cloud. Developers can offer those applications for customers to use. Customers pay for each use of the F1 configured to run the application in the same way they pay for any other cloud resource. This model enables more people to create FPGA applications since the cost of the CAD tools, FPGA board, and associated software are provided by Amazon. This model may ease the considerable burden of developing the reconfigurable computing hardware blocks for many commercial use cases, and may eventually lead to creation of an ecosystem that could increasingly support HPC needs.

4.2 Memory-Centric Processing

When we think of the effects of Moore's Law, we think of a continued increase in the compute performance of conventional processor chips. While true, this ignores what is needed from memory chips to balance this performance increase. To get a sense of proportion, as pictured in Figure 4.2, from the year 2000 to now the peak bandwidth per commodity "DDR"-style DRAM chip has risen by about 10×, whereas peak floating point performance per commodity processor chip has risen by over 200×. "GDDR" chips as used in earlier GPU accelerators, have higher bandwidths, but lower capacity and higher power, and still have not climbed at the same rate as GPU chip floating performance has. Chip architects have responded to this disconnect by adding more memory ports (limited by available chip pins), and by switching to 3D stacks of memory chips that have more exotic interfaces ("HBM" and "HMC"), but that still have not kept up with peak processor chip performance, and have driven up power and complexity.

Until now, this has not been a show-stopping issue, as the focus on dense linear algebra as a performance metric has meant that increasing on-chip caches could overcome almost any deficiency in memory bandwidth. This is no longer true as applications (both scientific and non-numeric)

become more sparse and irregular in their access patterns, and are significantly less cache-friendly. An example is HPCG (High Performance Conjugate Gradient)[1] that also solves large sets of linear equations, but where the matrices are very sparse. Whereas LINPACK can efficiently utilize 90% of the floating point performance that Moore's Law has brought us, HPCG typically is capable of using only 1-4%. In fact, analysis [26] has shown that HPCG is almost totally dominated by memory bandwidth; floating point capability or cache size is irrelevant.

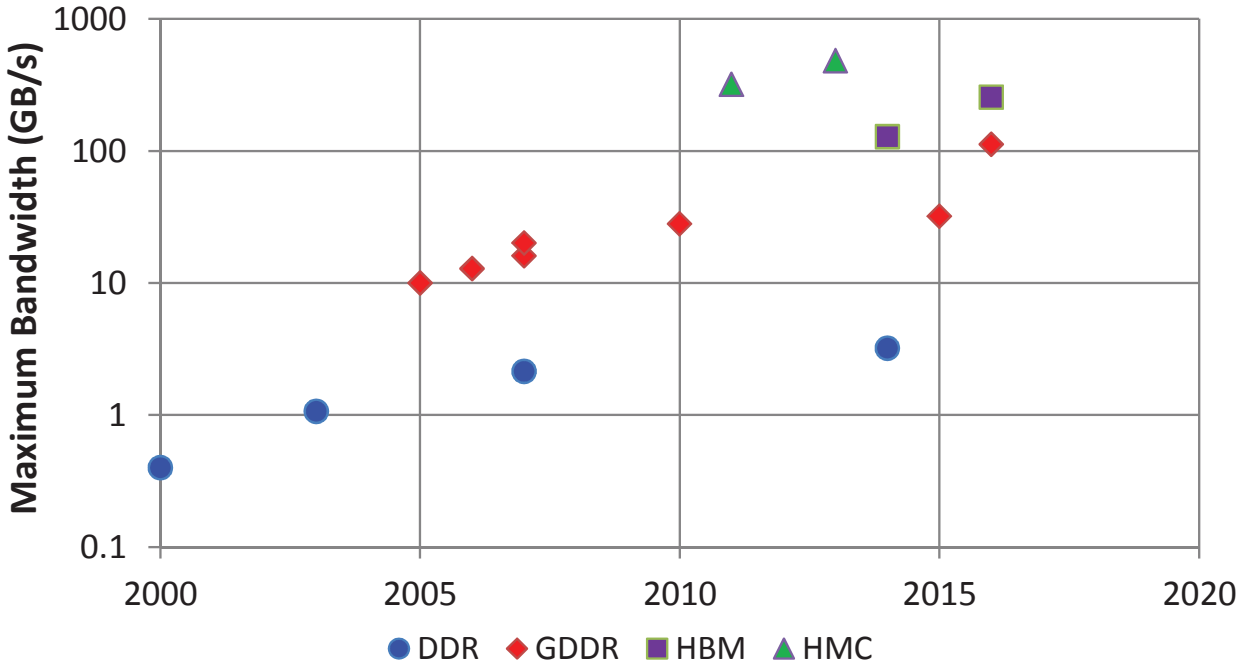


Figure 4.2: Growth in Memory Chip Bandwidth

Memory-centric processing is a technique that attempts to break this interface problem by moving processing much closer to memory than a conventional core. As shown in Figure 4.3, there is a definite taxonomy for where such memory-centric processing may be positioned, which includes:

- **In Cell**: within the bit cell storing the data.
- **At the Sense Amps**: at the bottom of the block of memory cells, at the first point where the data is converted to a digital level, and where it has access to literally hundreds to thousands of bits from a complete "row".
- **In-Situ**: a bit further down the digital chain but still within a memory bank, typically just after a "column" multiplexer that is driven from the output of the sense amps.
- **On Memory**: on the memory die itself, typically with access to all the independent memory banks on the die.
- **In Memory**: on a die between a memory, or stack of memory die, and the processor.
- **Near memory**: near the memory controller that may be on the memory die, but typically on a processor die.

[1] http://www.hpcg-benchmark.org/

Such architectures have several potential advantages:

- Finer grain control over the amount of data accessed is often possible, meaning that less data access is wasted.
- The energy costs of moving data across chips and chip boundaries may be significantly reduced.
- Latency of access is significantly reduced, meaning that less logic is needed to track multiple outstanding memory requests, and processing logic does not lie idle as long waiting for data.
- Such memory-centric logic is typically "outside" the normal cache hierarchy, including outside the coherency mechanism for multi-core architectures. This greatly reduces energy spent in managing copies of data that may be used only once.
- Being close to memory makes the ability to make atomic operations more efficient.
- Since many memories have significantly more internal "channels" than are presented to a conventional processor, there is an opportunity to have many more near-memory cores in action at the same time, greatly increasing concurrency.

In summary, virtually all of these advantages reduce energy, which is perhaps the biggest obstacle to exascale performance and beyond. Architecturally, the key research challenges include how to maintain some level of coherency with copies of the same data further down the cache hierarchy, how to spawn such remote computations, how to maintain a global address space, how to recognize completion of such operations, and how to handle cases where data from several separated memories need to be combined.

Table 4.1 illustrates several performance characteristics for these different levels of memory-centric processing. The columns are as follows:

- **Bits Reachable**: The number of different bits that might be accessible by a core at the specified location generating an address. For example, for "In-Situ" a core would have access to all the data in the memory block, whereas for "On Memory" it may have access to any of the memories on the die.
- **Bits per Access**: On each access, how many bits are possibly returned to the core. For example, for "In-Situ" it may be the width of a memory bank row.
- **Accesses per Sec (M/s)**: From a core in the specified position, how many different memory accesses could be made per second. For example, a 3200 MT/s DDR4 DIMM with a burst depth of 8 can make up to 400M accesses/s.
- **Bandwidth**: The product of the two above terms, bits per access and access rate.
- **Movement on Chip**: How far across a die must data be moved to get to either the processing core or the off-chip interface that leads to the core. This can be a significant source of energy overhead.
- **Chip Crossings per Access**: How many times must a chip edge be crossed. This can also be a significant source of energy overhead.
- **Functionality**: What kind of processing is reasonable.

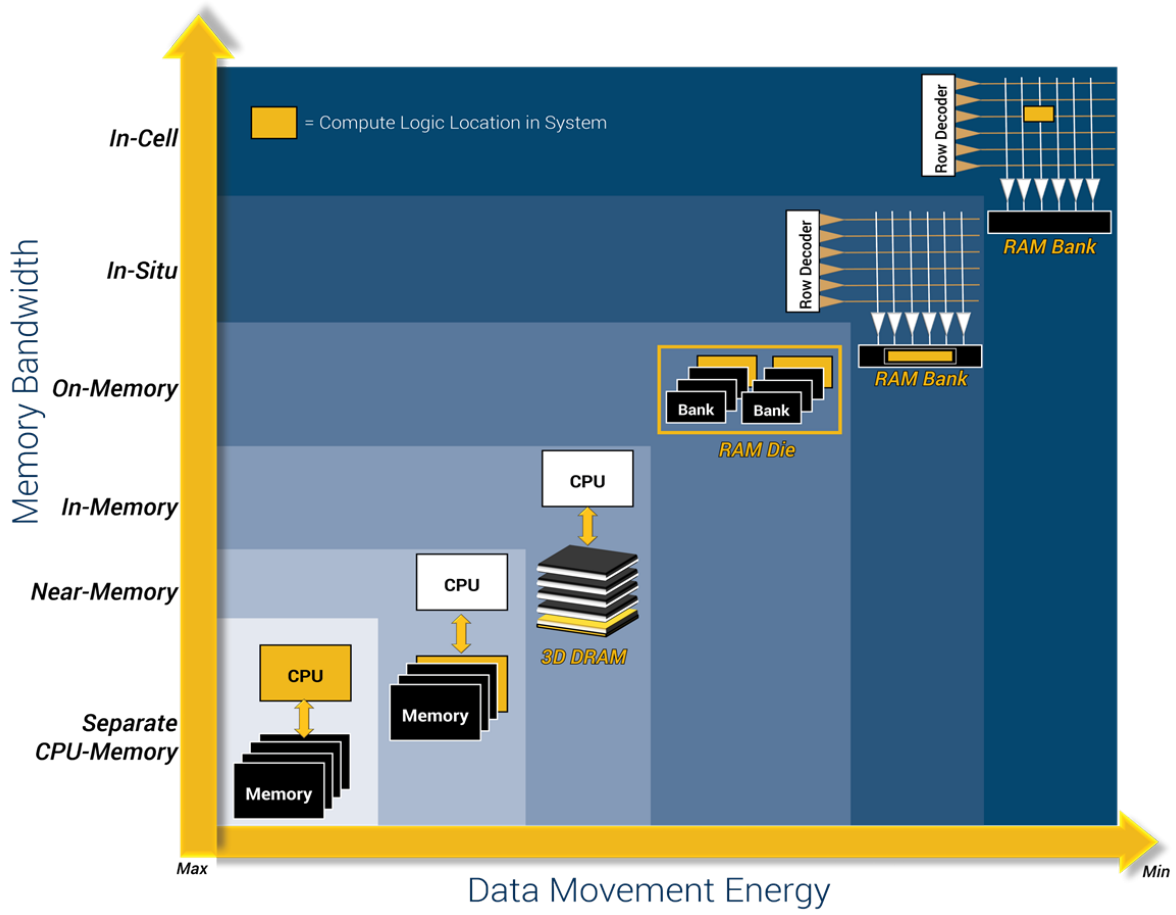


Figure 4.3: Different levels of memory-centric processing

- **ECC Possible?**: Is it feasible to include ECC - both the extra bits and the logic.

What is important to recognize about these numbers is that they are on a much finer scale than conventional memory, where one memory channel may be built from 36-72 DRAM chips. In contrast, most of the rows of Table 4.1 refer to capabilities that may be present in multiple instances on each of these chips.

Technology Readiness Timeframe: Looking forward, while examples exist of all these techniques today, the Near-Memory capability is perhaps of most interest because of its applicability to 3D stacks of chips, where the bottom chip of the stack has logic and network routing. This is likely a few years away, with no real technological hurdle. Also DARPA's "chiplet" program may very well develop processors that can be combined with a variety of memory technologies, as will possibly SRC's recently awarded JUMP programs. Candidates for "killer apps" for near-memory processing include memory-centric streaming operations such as encryption/decryption, search, big data, big graphs, and possibly deep learning.

Also, given the range of options demonstrated in Figure 4.3, it is clear that this technology will further contribute to the extreme heterogeneity anticipated in Future HPC systems.

	Bits Reachable	Bits per Access	Accesses per Sec (M/s)	Bandwidth (GB/s)	Movement on Chip	Functionality	ECC Possible
In-Cell	1	1	50	0.006	0	Bit-level SIMD	No
At Sense Amps	1Mb	2Kb	50	12+	Down Column	SIMD + Full core (Up to Vector)	Yes
In-Situ	1Gb	2Kb	50	12+	Down Column	SIMD + Full core (Up to Vector)	Yes
On-Memory	8Gb	64b	400	3.2	Down Bank	SIMD + Full core (Up to Vector)	Yes
In-Memory	4-8GB	1Kb	800	100	Across Chip	Full Core	Yes
Near-Memory	64+GB	64B	400	3.2	Across Chip	Full Core	Yes

Table 4.1: Performance characteristics for different levels of memory-centric processing.

4.3 Silicon Photonics

Among the technologies emerging toward creating a fundamentally energy efficient interconnect, photonics is perhaps the most promising to enable a transition to a new generation of scaled extreme performance computing systems [27]. Optical technologies can directly impact the critical communications challenges within computing systems through their remarkable capabilities to generate, transmit, and receive ultra-high bandwidth densities with fundamentally superior power efficiencies and with inherent immunity to noise and degradation. Unlike prior generations of photonic technologies, recent breakthroughs in silicon photonics offer the possibility of creating highly-integrated platforms with dimensions and fabrication processes compatible with electronic logic and memory devices. During the past decade, a series of major breakthroughs in silicon photonic devices has demonstrated that all the components that are necessary to build chip-scale photonic interconnect components (e.g. modulators, filters, switches, detectors) can be fabricated using common CMOS processes.

4.3.1 Current Photonic Interconnect Technologies

Most optical links in today's supercomputers are based on multi-mode optical fibers and Vertical Cavity Surface Emitting Lasers (VCSELs). They are also generally built around a one "channel per fiber" format. Signals received from the electrical side are directly used to drive the laser diode, without format conversion or adaptation of any kind. Based on recommendations issued by standardization bodies such as IEEE, transceivers receive electrical signals at 10, 14, 28 Gb/s on one to ten lanes, each being coupled into its separate fiber. Transceivers with electrical signals at 56 Gb/s (QSFP56 format) will arrive soon in the market. Standards for electrical signaling at 112 Gb/s are in preparation. Traditional non-return-to- zero (NRZ) signaling will be kept for 56G but

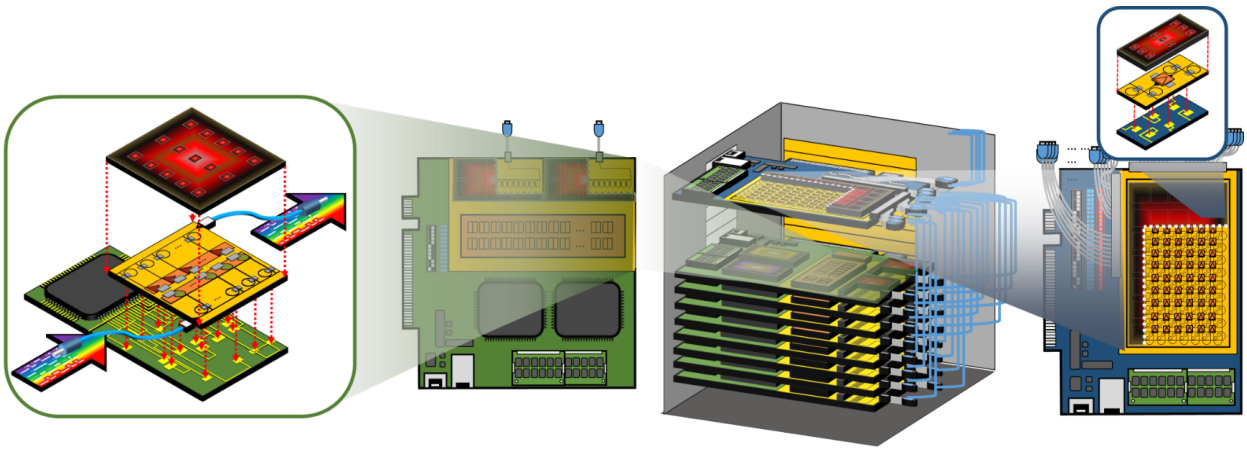


Figure 4.4: Current photonic interconnect technologies

most likely PAM4 signaling will be adopted for higher speeds. Directly modulated VCSELs have been shown capable of supporting extreme bit-rates provided that the adequate driving circuitry realizing pre-emphasis is provisioned alongside [28]. Products with 50 Gb/s or more per lane are only about to emerge, but VCSEL based systems have been scaled beyond the 50 Gb/s already, by means of fiber parallel systems. Multi fiber array connectors (MPO) with up to 24 fibers have been standardized (TIA 604-5-D) and standards with 72 fibers are in preparation. Such fiber ribbons and multi fiber connectors are, for instance, used in commercial products, such as Cisco's CPAK 100GBASE-SR10 module. The CDFP standard is based on cables made of 32 fibers, but including multiple fibers in parallel has an impact on cable management and cable cost. In addition, connectors involving many fibers are susceptible to show a higher loss. For this reason, VCSEL based multi-wavelength links (coarse WDM) have been proposed. The acronym SWDM, standing for Shortwave Wavelength Division Multiplexing, has been recently introduced to distinguish this technology. To realize the multiplexing and demultiplexing operations, solutions based on thin-film filters are among the most mature. Each thin film transmits a wavelength and reflects the others, at low loss in both cases. Such filters are cascaded to progressively isolate all wavelengths. Solutions to efficiently couple signals emitted by an array of VCSELs into optical fibers have also been investigated. Short-reach VCSEL based transceivers are expected to scale to ≈1 Tb/s bandwidth by means of highly fiber-parallel cables and/or WDM, in conjunction with high-speed signaling at or beyond 50 Gb/s. VCSELs have the important property to authorize testing at the wafer level, whereas other laser sources must generally be tested after dicing. They also show an emission aperture about three times larger, which greatly facilitates packaging. Altogether, these advantages allow VCSEL based links to show cost figures of a few dollars per Gb/s. This metric will be further scaled down by means of higher signaling speeds, increased wavelength and/or fiber parallelism, and as a result of further simplified packages and test procedures. Increase in manufacturing volumes will contribute to further cost reductions.

4.3.2 Emerging Silicon Photonics Interconnect Technologies

Silicon photonics (SiP) emerged in the last decade as a promising optical interconnect technology. SiP takes advantage of the high index contrast between silicon (3.476 at 1550 nm) and silica (1.444 at 1550 nm) to enable micro-meter scale optical guiding structures such as add-drop filters

and switches. For modulation, free-carrier dispersion effect is the only mechanism in silicon fast enough to enable purely silicon-based high-speed electro-optic modulation (10 Gb/s and beyond). Combined with the resonant nature of ring resonators, compact wavelength-selective electro-optic modulators with very small footprint can be realized in SiP platforms [29]. An array of such modulators can provide WDM transmission with aggregate rates in the excess of 100 Gb/s. Modulation can also be realized in silicon alone by means of Mach-Zehnder Interferometers (MZI). MZIs are less sensitive to thermal fluctuation than ring resonators, but are not wavelength selective, obliging each wavelength to be independently modulated before being multiplexed. Another modulation approach consists of selectively growing SiGe waveguides on top of a silicon wafer to form an electro-absorption modulator.

WDM operation can provide unprecedented interconnect bandwidths that fall well within the requirements of supercomputers in the near future. This concept was demonstrated by using a single quantum dot comb laser and an array of SiP ring modulators with 10 Gb/s per laser line. Based on this capability recent work on SiP-based DWDM interconnects showed the possibility of 1.56 Tb/s bandwidth at 25 Gb/s signaling rate and overall 7.5 pJ/bit consumption (assuming full link utilization) [30]. More recently, updated work showed a maximum aggregation of 2.1 Tb/s at 45 Gb/s per channel.

There are strong motivations to co-integrate the optical transceivers with compute modules (CMP or GPU), as well as with memory packages. A single package allows cost reduction for OEM vendors, reduces the wiring complexity on boards, results in higher component density, and most importantly can reduce signal degradation between data source and optical transceiver. If transceivers and data sources are placed in close proximity, their communication can be simplified and greater power and area saving can be achieved. In 2012, Altera together with Avago demonstrated an FPGA VCSEL transceiver assembly using a package on package (PoP) approach. The optical aggregate bit-rate of the FPGA assembly reached 120 Gb/s. Recent packaging trends are aiming at a closer integration of transceivers and ICs within the same package. System in package products integrate several chips within one package by coupling them using a common interposer.

A silicon photonic interposer enables optical networks in-package either for high bit-rate communication of chips within the same package or at the same speed with peripherals as the package boundary is of no importance for optical signals. The highest level of integration is reached when the data source integrates optics on the same die, so called monolithic solutions. Monolithically integrated chips have the smallest parasitic loadings possible. Therefore, they show very high energy efficiencies. However, CMOS processes are not optimal for silicon photonic structures. In addition, optical structures cannot be arbitrarily reduced in size and a single modulator's size will remain in the micrometer range even as transistors continue to shrink in size. Hence, monolithic solutions are very costly if integrated with modern deep sub-micrometer CMOS processes. From a geometrical perspective it is a challenge to integrate a sufficient number of pins and transceivers into each die or package to carry all the data in and out. Both directly modulated VCSELs as well as silicon photonic transceivers can emit and receive light into and from fibers perpendicularly oriented to the chip plane. If a chip does not need to carry the data to the optical transceiver by a 2D interposer but instead can emit and receive on the top surface of the die or die stack itself, very high bit-rate densities can be achieved, independent from the overall packaging approach.

Technology Readiness Timeframe: Research and development is pushing forward the forefront of silicon photonics design and manufacturing. Progressively, an ecosystem of fabrication infrastructures, circuit design and automation software (EPDAs), researchers and industries is emerging. In 2015, the US Department of Defense initiated a national center of innovation specifically dedicated to nanophotonic system manufacturing (AIM Photonics) [31]. However, without specific investment, the adoption of photonic technologies in high-performance (exascale and be-

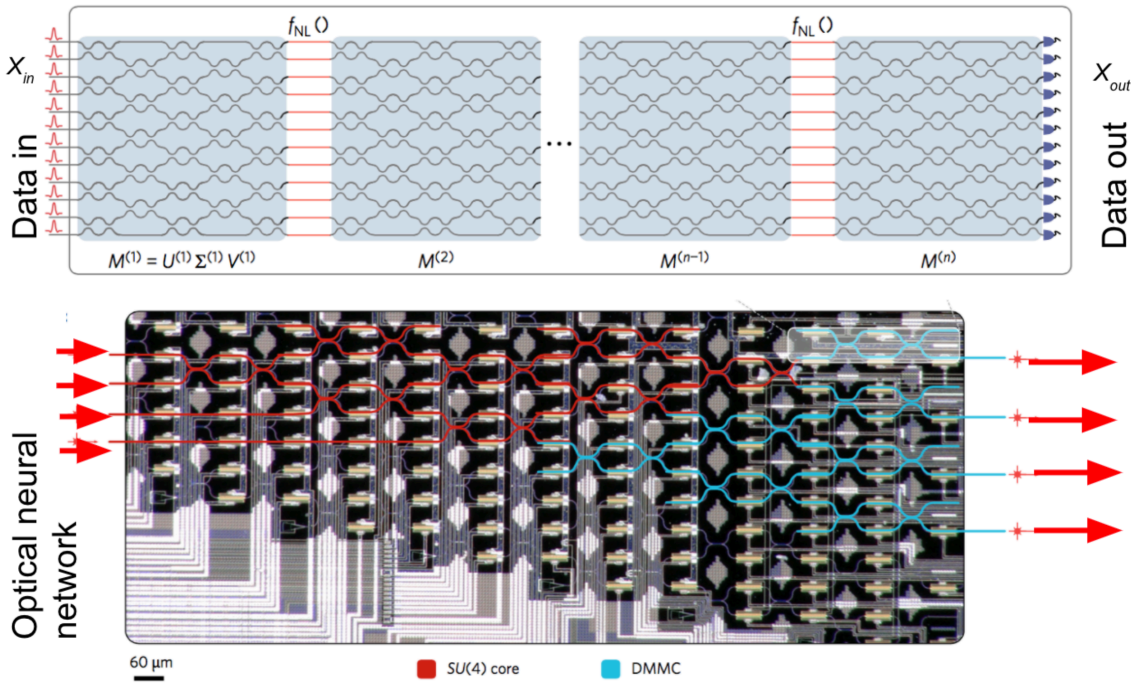


Figure 4.5: The optical NN consists of a series of n layers, each consisting of a matrix transformation M followed by an optical nonlinearity. The computation on an input vector Xin, encoded in the amplitudes of laser signals (left), occurs nearly instantaneously at the speed of light.

yond) interconnects over the next 5 years will largely build on the technologies currently developed for the commercial data center market where there is less emphasis on performance.

There are also some preliminary results showing the promise for using photonics for going beyond communication to enable a new kind of analog computing. An example is the recent development of a new architecture for an optical neural network (NN) that could bring significant advantages in computing speed, latency, and energy consumption [32, 33]. Recent experimental demonstrations show the core components of the architecture using a new class of fully programmable nanophotonic processor based on a CMOS-compatible silicon photonics architecture (see Figure 4.5). The key advantage for NNs is that the matrix transformation, which combines signals in neural networks, is performed optically at the speed of light. The number of operations needed to compute this transformation on N input signals scales linearly as N, whereas it scales as N^2 in a digital NN. In addition, the weight matrix – i.e., the strengths of connections between signals – can be encoded into a passive photonic circuit, whereas the digital NN requires the weight matrix to be accessed from memory. As a result, the optical NN promises significant advantages in speed and energy consumption.

4.4 Neuromorphic Computing

Neuromorphic computing covers a very broad set of approaches. In this section, we will give a brief overview and history to set the context, and highlight its most promising opportunities. Figure 4.6 shows a high-level comparison between conventional and neuromorphic computer architectures

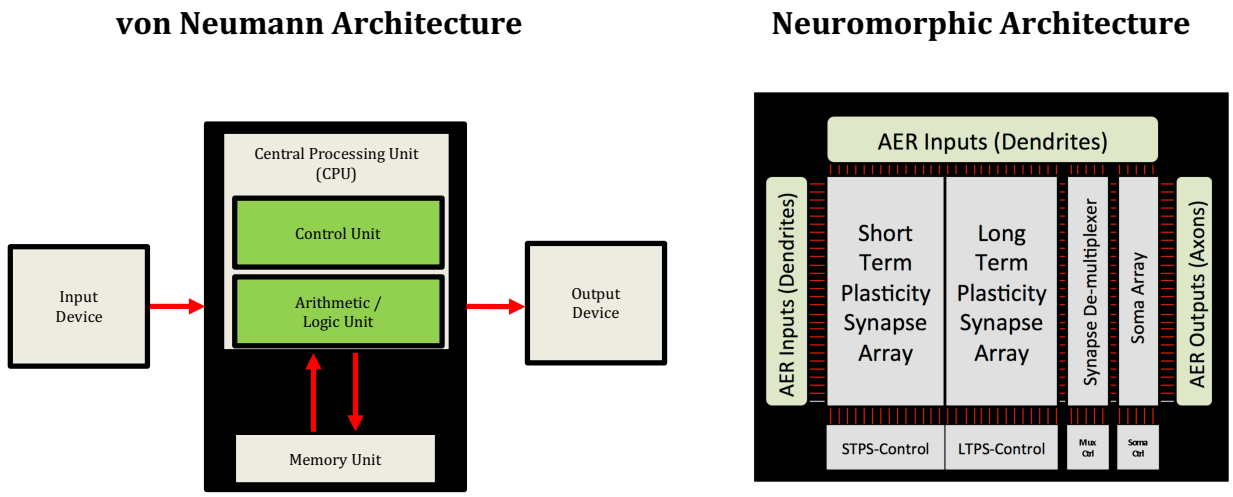


Figure 4.6: Comparison between conventional and neuromorphic computer architectures

taken from a recent DOE report [34]. The data path between the CPU and the memory unit serves is the so-called "von Neumann bottleneck". In contrast, a neural network based architecture combines synapses and neurons into a fine grain distributed structure that scales both memory (synapse) and compute (soma) elements as the systems increase in scale and capability, thus avoiding the bottleneck between computing and memory.

Generally speaking, neuromorphic computing refers to the implementation in hardware of circuits emulating, whether closely, or remotely, the behavior of the brain, in particular neurons and synapses. We need to distinguish two main trends, and purposes, of neuromorphic computing: (1) emulating the behavior of a subset of the brain, i.e., a number of neurons, (2) achieving brain-like functionality, such as object or speech recognition, i.e., actual applications. Until recently, most of the funding and efforts were targeted at approach (1). Some of the main programs include DARPA Synapse in the US, and the Human Brain Project in Europe. They resulted in architectures, such as IBM's TrueNorth processor and the SpiNNaker architecture from University of Manchester, UK, capable of emulating a billion or more spiking neurons. The overall goal of these approaches is that such architectures can be used as modeling tools by neuroscientists to emulate brain-like functionality. While the scientific value of such machines for neuroscience is a possibility, the approach hasn't as yet demonstrated significant successes in terms application functionality or efficiency. A key problem is that spiking neurons-based algorithms for actual tasks (e.g., object recognition) aren't competitive, for now, with machine-learning algorithms based on deep neural networks.

Artificial neural networks, more recently known as Deep Neural Networks (DNNs), form approach (2). The principle of artificial neural networks is to use "brain-inspired" operations that perform a sum of input neurons weighted by synapses, followed by a non-linear function. The history of artificial neural networks is long, and their success only recent, due to the current availability of large volumes of training data and compute power for multiple application domains. After an initial excitement in the 1950s with the Perceptron, there was a spike of enthusiasm and interest with Multi-Layer Perceptrons (MLP) in the 1980s/1990s. Interest in the Perceptron model declined as they were outperformed by algorithms with seemingly better properties, such as Support Vector Machines (SVMs). It's only after GPUs enabled training of large enough networks with enough training data, that researchers were able to show how powerful these approaches are. Today, DNNs

are at, or close to, human-level performance for non-trivial tasks such as object recognition, speech recognition, translation, etc. As a result of their growing popularity, it has become sensible for companies, such as Google, to implement ASICs to efficiently support such algorithms. Google has publicly disclosed using TPUs/Cloud TPUs in its data centers, Microsoft and Amazon have disseminated FPGAs for the same purpose, and NVIDIA is actively supporting the usage of DNNs in self-driving cars.

Technology Readiness Timeframe: Going forward, we can expect DNN algorithms to be broadly used, both in data centers, and in devices, from phones to self-driving cars, and many others, and as a result, many companies are expected to propose ASICs efficiently supporting them.

4.5 Quantum Computing

Quantum computing is a model of computation that proposes to exploit the quantum mechanical nature of specific physical phenomenon to provide advantages relative to so-called classical computing, i.e., the familiar use of CMOS and other digital logic. Whereas N digital bits contain one N-bit state, N entangled quantum bits (qubits) contain 2^N states upon which operations can be simultaneously applied. Quantum computing was originally conceived of as a way to use quantum mechanical phenomenon to solve problems in modeling other quantum mechanical properties of materials. The range of potential applications for which quantum computing offers advantages relative to classical computing has since expanded, including factoring composite integers (Shor), search (Grover), and optimization (quantum annealing). A complete list of known quantum algorithms and the speedups they offer can be found at [35].

Quantum computing today is a promising technological direction, but one which will still require significant research and development effort before becoming a tool that can be applied for broader scientific discovery. Since the advent of Shor's algorithm, there has been substantial investment in quantum computing worldwide, first by governments, and more recently, commercial interests. The range of potential applications for which quantum computing offers advantages relative to classical computing has grown, and now including the simulation of physical systems for applications in materials science and quantum chemistry, training of machine learning models, solving of semi-definite programs, and solving linear systems of equations. In addition, there has been an interesting side effect of quantum computing research, the development of new, quantum-inspired classical algorithms.

The announcement of USA's National Quantum Initiative Act in 2018 has increased the allocation of funds in DOE and other agencies towards advancements in quantum computing, with the promise of continued future investment in this direction. There are many opportunities for DOE, and in particular, ASCR, to contribute to these advances. Quantum speedups, i.e., algorithms with better scaling properties relative to traditional computing, have been discovered for a variety of scientific problems of interest to DOE. These range from problems in chemistry and physics, to data analysis and machine learning, and to fundamental mathematical operations. Further investigation by the ASCR mathematics and computer science research programs will both broaden and strengthen these capabilities.

The above-mentioned quantum algorithms are supported by theoretical proofs of their scaling properties. However, without the existence of suitable quantum computers, they cannot yet be exploited to accelerate time to discovery. Therefore, DOE SC, working with other offices such as BES, can work on the development of materials and devices to make it possible to realize such machines in the future, at scales where they will offer true computational advantage relative to

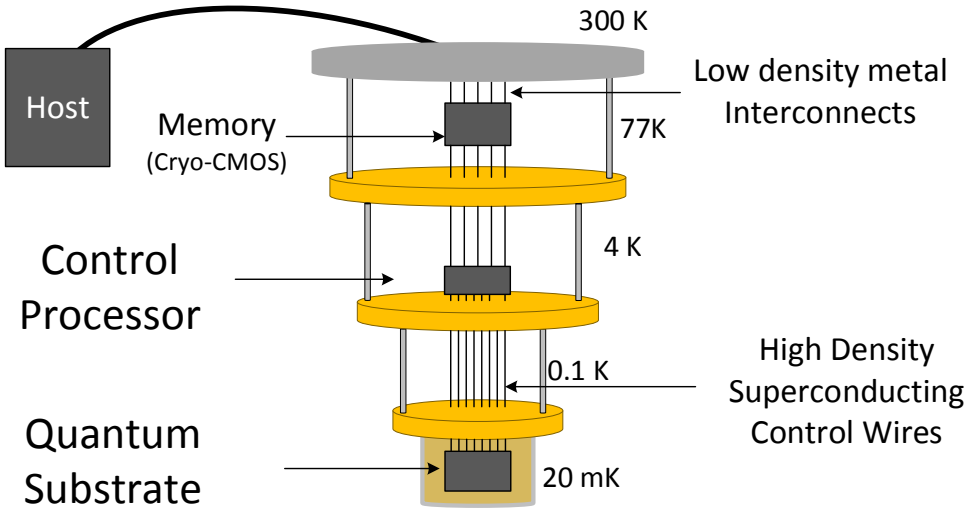


Figure 4.7: Thermal hierarchy for host and control processes connected to a quantum substrate

classical machines.

Prototypes of small quantum systems, be they specialized annealing devices (e.g., D-wave's machines), or even general purpose computers (e.g., machines from Google, IBM, Intel, Microsoft), are beginning to appear. DOE ASCR's facilities division can take a leading role in evaluating such devices, and making them accessible to the broader scientific community, which in general won't have the capability to house such devices. Quantum computing systems need to be isolated form the external world, so as to maximize coherence. In many of the leading paths toward physical realization, Helium-3 dilution refrigerators inside of Faraday cages are used for thermal and electromagnetic isolation, as illustrated in Figure 4.7. As a result, there are fundamental challenges in creating quantum computing testbeds that go beyond the quantum substrate, *e.g.,* a thermal hierarchy is needed to bridge the large thermal gradient across a host processor operating at an ambient temperature (300°K), a cryogenic control processor operating at 4°K and the quantum substrate operating (say) at 20°mK. It may be possible in the future to create quantum devices that require less extreme cooling, and it is possible that a scalable quantum computing system will require integration between multiple types of quantum bits, not all of which require such extreme cooling. Nevertheless, it will likely be a long time before the devices can be broadly deployed within the scientific computing community.

Technology Readiness Timeframe: Quantum computing is evolving from a theoretical curiosity in the 1980s to a tantalizingly close possibility today. Specialized devices, such as open system, adiabatic quantum annealers are available today, but still have fundamental challenges to overcome before becoming useful [36]. General purpose machines, albeit with limitations on size and error correction, are also starting to appear (*e.g.,* devices being developed by Google, IBM, Intel, and Microsoft). It is reasonable to expect that they will scale in the post Moore's Law time frame to be able to solve problems of interest to DOE, such as electronic state calculations. They will likely serve as specialized accelerators for problems beyond the reach of classical computing, and DOE will need to learn how to integrate them into its increasingly heterogeneous, post-Moore's Law scientific computing infrastructure. This ranges from mathematical and computer science problems of how to extract from a larger problem components suitable for quantum computation, to practical questions such as the communication interfaces that would allow integration of a quantum computer

with the rest of DOE's computing infrastructure.

When powerful quantum computers become available, capable of uniquely solving some of the nation's problems in science and engineering, they may still remain unapproachable to the vast majority of scientists and engineers who have not been trained to use them. Development of suitable programming languages and tools will need to accompany the systems themselves, in a way analogous to the development of such tools for classical computing, which started six decades ago with FORTRAN. ASCR research investments can build upon and extend early efforts to develop such tools for quantum computing, including IARPA's Quantum Computer Science program, the Microsoft Quantum Development Kit, and IBM's QISKit. It will also be necessary to invest in creating a new quantum workforce, training scientists to frame their problems suitably, so as to use the new quantum computing environments.

4.6 Analog Computing

Analog computing is the use of a physical process that is of reasonable efficiency to compute an analogous process that shares the same physical relationships. A simple example is the electronic-hydraulic analogy for Ohm's law [37]. Electronic analogous systems are particularly well suited to solving systems of partial differential equations – an approach that was used extensively prior to the emergence of digital computers [38]. Digital computing surpassed analog computing due to its ability to represent quantities to much higher dynamic range and precision than were then (and now) possible in analog electronics. There are several reasons for this, including the manufacturing process variations that impact the signal-to-noise ratio (SNR) and accuracy of differential amplifiers in analog computing, and the limits of metrology even in the case of infinite SNR [39].

The recent interest in data-driven science has led to the creation and adoption of a new generation of machine learning techniques that do not require the relatively high level of precision associated with classical scientific and engineering applications, such as the solution of PDEs. This is reflected in the addition of half-precision (16-bit) to the IEEE 754 floating point standard, and its implementation in new devices such as the Nvidia Volta GPU. For such applications, that do not need high precision and can perhaps tolerate modest errors, analog computing offers the possibility of much greater performance and energy efficiency, as mentioned in Section 4.3.2. There are many possible physical phenomena that can be revisited in this regard (e.g., the use of arrays of resistors for multiplication and lenses for Fourier transforms), many of which include techniques in use before the emergence of general purpose digital computing.

A second approach to analog computing is via modeling physical processes that naturally reconfigure themselves according to the theory of thermodynamics [40, 41]. We believe this approach to analog computing holds great promise as well. In its simplest form, a thermodynamic computer (TDC) is a system that uses the thermodynamics of annealing near equilibrium to find (near) optimal solutions to complex problems. Examples include using analog electronics to perform annealing [42, 43] as well as the development of quantum annealers mentioned earlier, e.g., D-Wave Systems' *Orion, One, Two* and *2X* quantum annealers [44]. As observed in [41], TDCs are related to neuromorphic unsupervised learning techniques including Helmholtz machines [45] and variational autoencoders [46]. These approaches are able to, "learn optimal encodings of the underlying structure in unlabeled data."

Generalizing from this, a new class of computational devices that spontaneously organize are emerging. These TDCs are open, non-equilibrium, thermodynamic systems that evolve their organization in response to the thermodynamics in the environment. Formalization of these ideas has emerged recently from work in non-equilibrium statistical physics and related fluctuation theo-

rems [47–50]. However, the idea of thermodynamic evolution challenges many long-standing philosophical and technical assumptions in the field of computing and beyond.

A generalized TDC architecture is a networked fabric of thermodynamically evolvable cores (ECs) embedded in a reconfigurable network of connections, as shown in Figure 4.8. Energy is the "language" of the network and time-efficient communication is critical. It is the job of the entire system, both the network and the ECs, to move energy from inputs to outputs with minimal loss. Losses within the TDC create variations that cause reconfigurations to naturally occur.

A TDC can be programmed to solve a specific problem. The "problem" is defined by the structure of the energy / information in the environment. Programmers preconfigure some of the ECs to define constraints. Dissipation within the network creates fluctuations over many length and time scales and thereby "search" for solutions over a very large state space. Structure precipitates out of the fluctuating state and entropy production increases in the environment as free energy flows through the network and dissipation decreases.

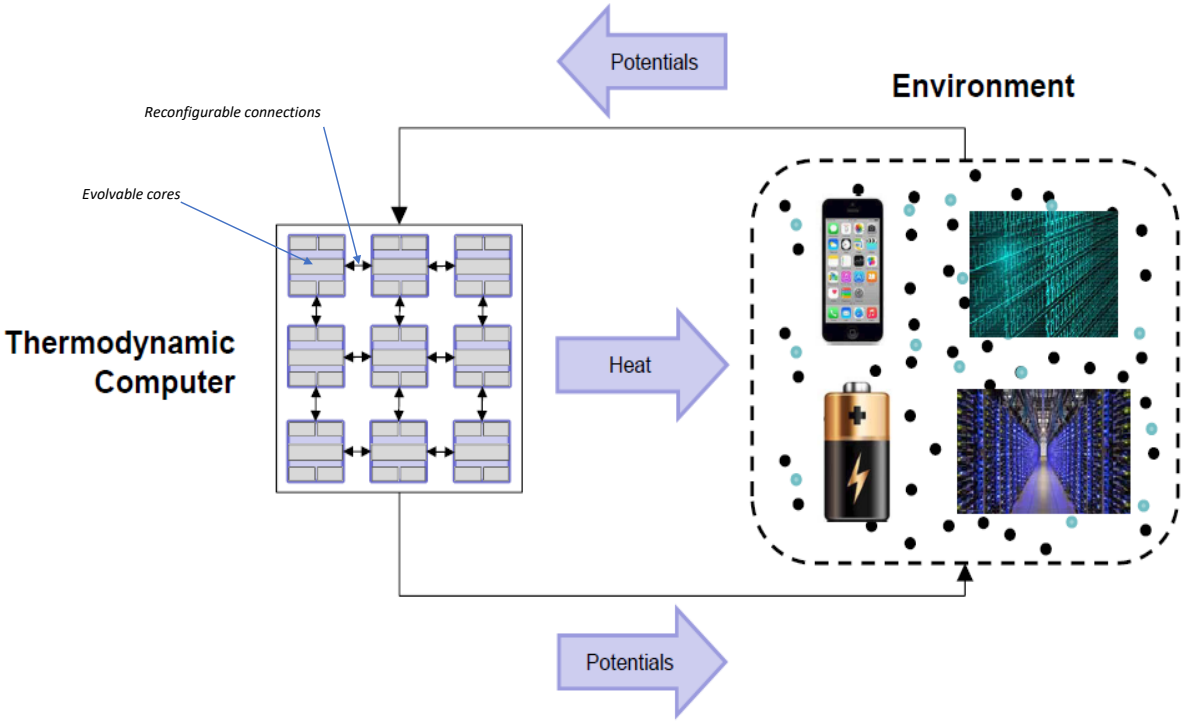


Figure 4.8: Example high-level architecture of a thermodynamic computer. (*Courtesy of T. Hylton, with permission*)

Technology Readiness Timeframe: Electronic analogous computing predates modern digital computing, but the low precision of these systems lead to their demise. The analog content of smartphones and automotive is significant and has led the growth of the analog semiconductor industry. However, the percentage of the analog segment is under 5% of the overall semiconductor industry (digital and analog) [51]. Of the analog segment, only a small fraction is itself dedicated to amplifiers [52]. DOE critical applications have potential uses of analog computing, but the current commercial market pressures are unlikely to improve SNR (i.e., effective bit precision) without incentives and R&D investment, perhaps arising from new applications of commercial importance.

In contrast, annealing approaches that leverage analog processes to solve optimization problems are gaining traction, especially with respect to quantum-based annealing. There is overlap between electronic analog annealing and the active field of neuromorphic unsupervised learning. The latter is causing rapid development of the former [42, 43]. These technologies exist in the marketplace and the quantum-based annealing approach is being applied to DOE problems already.

Extensions of annealing approaches to a more general thermodynamic computing paradigm are currently beginning to emerge from theoretical studies into the realm of early prototypes and proofs of concept [41]. TDC is in the same state that gate-based Quantum computing was a decade ago. The potential is high for TDC to succeed where QC cannot: energy and power constrained systems. However, TDC also requires significant R&D to move forward.

4.7 Application Challenges

While the new hardware technologies discussed in this chapter provide many exciting opportunities for future science applications, there will undoubtedly be very significant challenges for science applications to leverage these technologies. As previously discussed in Chapter 3, previous technology transitions have forced the developers of scientific and engineering applications to explicitly exploit dramatically increasing levels of parallelism. The form of parallelism that is exploited evolves, to reflect contemporaneous HPC architectures, but the basic tenet has held true for the last five decades, since the introduction of vector mainframes. With the end of Dennard scaling, and the cessation of clock frequency growth, increased capability now comes from exponentially increasing parallelism, and developers already uncover these levels of parallelism in the algorithms, explicitly represent it, and then choreograph the interaction of millions of concurrent operations. This is a daunting task today, and will only grow as we transition to exascale, where the number of independent operations will increase to be on the order of billions, with extreme levels of heterogeneity in post-exascale computing. The challenges abound, and there is need for mathematical and computer science research to address them, so as to make post-exascale systems accessible to as broad a swath of the computational science community as possible. We are already faced with the challenges of design for adaptability, heterogeneity, dynamic data and work partitioning, and remote and asynchronous execution. Looking to the future, there are also the core challenges of designing scientific applications for reconfigurable logic, memory centric and silicon photonics technologies (among others).

It is anticipated that exascale systems will have $O(10^9)$ ALUs. The parallelism needed to go beyond exascale will surely be even greater. Research into mathematical algorithms that can both create and sustain this level of parallelism, without excessive synchronization is critically needed. Simple operations in familiar algorithms, like computing residuals or Courant numbers threaten to become computational bottlenecks due to the need to coordinate their computation amongst all processors. New algorithms that scale effectively, yet are also robust enough to solve a broad range of problems need to be invented.

Mapping new or existing applications to post-exascale and post-Moore computing systems will be increasingly challenging. As discussed earlier in this chapter, increasingly heterogeneous components will be incorporated into systems, to maximize both computing power and energy efficiency. Choosing among the diverse components of one computing environment will be challenging, and porting amongst multiple such systems even more so. New execution models will need to be created, with abstractions for components that we do not have today, e.g., quantum-based accelerators and ephemeral FPGA-based functional units. Programming systems will need to assist developers face these application challenges by creating and mapping new programming abstractions to diverse

machines, and providing tools for both functional and performance debugging that allow users to understand if their programs are running correctly, and with adequate performance, and where to fix them when they are not. Some of these needs were also identified in the five Priority Research Directions identified in Extreme Heterogeneity workshop report [2]: 1) Maintaining and Improving Programmer Productivity, 2) Managing System Resources Intelligently, 3) Modeling and Predicting Performance, 4) Enabling Reproducible Science Despite Diverse Processors and Non-Determinism, and 5) Facilitating Data Management, Analytics, and Workflows.

Finally, quantum and analog computing represent qualitatively different approaches from the other technologies, and it is difficult to predict at this time if and how applications for these technologies will be integrated into our HPC ecosystem. At the same time, these technologies are presently highly specialized, and their application base will likely start small, so general concerns of integration are not pressing at this time. Further, the investments accompanying the National Quantum Initiative Act promise to accelerate early breakthroughs related to quantum computing.

4.8 Open Platforms

As increasingly diverse hardware architectures proliferate, co-exist, and interact with traditional instruction set architectures, there is an increased need for the development of open platforms with open interfaces. Some of the key issues to be addressed by open interfaces include:

- resource allocation, protection, and coordination,
- efficient management of multiple memory domains with varying characteristics,
- memory address translation management,
- cache management optimizations,
- extreme scale file and storage system demands, and
- security in the presence of "bare metal" directly attached and network-accessible collections of accelerators.

On the hardware front, these open interfaces could help support the development and integration of new hardware protocols for communication, coherence, and synchronization among processing units, as well as novel, tightly integrated accelerators/co-processors, some of which may be the outcome of open source hardware development [53]. We observe that the presence of open interfaces and open source hardware components focuses, rather than restricts, the role of proprietary hardware innovation. On the software front, open interfaces could enable new innovations in system software to support both distributed computations as well as distributed data stores to hold the growing experimental and observational science data.

As a recent example of the benefits of open interfaces, we can look at the tremendous success in identifying and designing new scientific software abstractions and libraries that make the use of neuromorphic platforms almost turnkey for application developers. Open source software libraries such as TensorFlow, Caffe, and others [54] have enabled many scientists to integrate machine learning into their computational workflows. The emerging importance and the growing hardware support for fast low-precision computations has spurred a new effort for batched and low precision BLAS [55]. All of these developments are being integrated seamlessly into our computing ecosystem, building on decades of experience with open source software in the HPC community.

Chapter 5

Findings

5.1 Need for clarity in future HPC roadmap for science applications

The challenges associated with post-exascale and post-Moore computing are receiving significant attention from multiple government agencies and initiatives including DARPA, DOE, IARPA, NSF and NSCI. However, while some of these efforts are focused on particular application domains (e.g., high-performance data analytics) there is currently a lack of clarity as to what the future high performance computing roadmap is for science applications. The subcommittee believes that Science will need to prepare for a period of uncertainty and exploration in future HPC technologies and computing paradigms, akin to the exploration in the 1990s before our current Massively Parallel Processing (MPP) paradigm emerged as dominant successor to vector parallelism. However, it is exactly because of this uncertainty that there is a need to focus on strategy and planning activities so as to better anticipate and update, on an ongoing basis, what the future HPC roadmap possibilities will be for science applications.

5.2 Extreme heterogeneity with new computing paradigms will be a common theme in future HPC technologies

As discussed in Chapter 4, there is a great diversity in the technologies that are expected in the post-exascale and post-Moore eras. These technologies include new forms of heterogenous processors, heterogeneous memories, near-memory computation structures, new interconnect technologies (including silicon photonics), and non-von Neumann computing elements based on analog, neuromorphic and quantum technologies. This diversity in computing paradigms has been appropriately labeled as "extreme heterogeneity" in an ASCR workshop held in 2018 [2] and related discussions. The subcommittee believes that there is value in focusing on extreme heterogeneity as a common theme in future HPC technologies, so as to enable a broader view of post-Moore computing rather than focusing solely on point solutions such as neuromorphic computing and quantum computing. At the same time, there are compelling research challenges in moving these point solutions forward so that they can be integrated in future platforms that exhibit extreme heterogeneity.

5.3 Need to prepare applications and system software for extreme heterogeneity

As discussed in the report, different applications have responded to past technology transitions (e.g., from vector to MPP, terascale to petascale, petascale to exascale) in different ways. We are rapidly approaching a period of significant redesign and reimplementation of applications that is expected to surpass the disruption experienced by the HPC community when transitioning from vector to MPP platforms. As a result, scientific teams will need to prepare for a phase when they are simultaneously using their old codes to obtain science results while also developing new application frameworks based on the results of new applied math and computer science research investments. In order to improve productivity, application developers will further need to rely more heavily on external and evolving software capabilities: expanded use of libraries, code transformation tools and evolving language standards. These software dependencies need to be sustainably supported in order for application teams to readily adopt and rely upon them.

5.4 Need for early testbeds for future HPC technologies

Given the wide diversity of technologies expected in the post-Moore era, accompanied by radically new computing paradigms in many cases, there is a need for building and supporting early testbeds for future HPC technologies that are broadly accessible to the DOE community, so as to enable exploration of these technologies through new implementations of science (mini-)applications, e.g. [56].

Timing-realistic emulation can also serve as a valuable evaluation tool to assess hardware designs prior to and during realization to physical implementations. The degree of fidelity and method of emulation depends on the architecture being studied. For example, experiments with asymmetric memory latencies for read and write operations could be tested on existing systems that can change memory timing through control registers. Alternatively, an FPGA emulator could insert delays in soft logic to mimic characteristics of new memories [57]. Novel computation blocks or microarchitecture implemented on FPGAs can serve as a surrogate that eventually is replaced by the actual hardware in the testbed.

These explorations could yield new computational motifs that are better aligned with the new computing paradigms. There are multiple instances of individual research groups at DOE laboratories creating early testbeds (e.g., [58–61]), but administration of these testbeds is necessarily ad hoc, due to their being supported by researchers, and lacks the support for broad accessibility that is typical for DOE computing facilities. Collaborations between DOE laboratories and universities (e.g., [62]) can help improve accessibility, with universities undertaking early explorations (e.g., [63]) to help identify technologies that may be deserving of hosting as testbeds in DOE Facilities, while also contributing to the development of researchers who can use these testbeds.

5.5 Open hardware is a growing trend in future platforms

With extreme heterogeneity, there is a growing trend towards building hardware with open interfaces so as to integrate components from different hardware providers. The motivation behind this trend is to enable new approaches to System-on-Chip (SoC) design that can more easily integrate components form different vendors.

There is also a growing interest in building "open source" hardware components through recent movements such as the RISC-V foundation. Despite many obstacles in building production-strength

hardware components through an open source approach (e.g., lack of EDA tools that are used for building proprietary hardware), open source hardware promises to be a growing trend in the future, which could help support the creation of hardware components (e.g., on-chip accelerators and interconnects) that are customized to the needs of science while being integrated with proprietary components from hardware vendors. In the opinion of the subcommittee, the presence of open interfaces and open source hardware components focuses, rather than restricts, the role of proprietary hardware innovation. For the purpose of this report, the term "open hardware" encompasses both open interfaces for proprietary components as well as open source hardware.

5.6 Synergies between HPC and mainstream computing

Though this report has focused on future high performance computing requirements from the perspective of science applications, there are notable synergies between future HPC and mainstream computing requirements. Some of them have been called out in the paragraphs on Technology Readiness for the different technologies described in Chapter 4, e.g., there is already a growing commercial use of reconfigurable logic in mainstream platforms . One application area where these synergies are already being leveraged, and will undoubtedly grow in the future, is in the area of data-intensive applications and data analytics (e.g., the use of neuromorphic computing and other accelerators for deep learning). As observed in a past ASCAC study [64], there are also notable synergies between the data-intensive computing and high-performance computing capabilities needed for science applications.

Chapter 6

Recommendations

6.1 Office of Science's Role in Future HPC Technologies

Recommendation 1: The DOE Office of Science should play a leadership role in developing a post-Moore strategy/roadmap/plan, at both the national and international levels, for high performance computing as a continued enabler for advancing Science.

The findings in this study have identified the urgency of developing a strategy, roadmap and plan for high performance computing research and development in the post-exascale and post-Moore eras, so as to ensure continued advancement of Science in the future. Though there are multiple government agencies that are stakeholders in post-Moore computing, the subcommittee recommends that the DOE Office of Science play a leadership role in developing a post-Moore strategy/roadmap/plan for advancing high performance computing in the service of Science. As in past years, this leadership role should span both the national and international levels.

There are many aspects to leadership in this regard. As was done for exascale computing, it is important for DOE to raise public awareness of the upcoming post-Moore challenges, and its impact on different science domains, well in advance of the start of the post-Moore computing era. However, unlike exascale computing, it will also be important to set expectations that different post-Moore technologies will have different time horizons, which will require a more agile and adaptive planning methodology than what is currently required in the Exascale Computing Project. In addition, engagement with existing technology roadmap efforts (such as IRDS) should play a key role in establishing DOE's strategy as to which timeframes are appropriate for adopting different post-Moore technologies. Finally, international competitiveness dictates that DOE Office of Science maintain its role in ensuring USA's continued worldwide leadership in high performance computing.

6.2 Investing in Readiness of Science Applications for post-Moore era

Recommendation 2: DOE should invest in preparing for readiness of science applications for new computing paradigms in the post-Moore era

The findings in this study have identified the challenges involved in preparing applications for past technology disruptions, and the fact that these disruptions will require exploration of new computing paradigms as we move to extreme heterogeneity in the era of post-Moore computing. The subcommittee recommends that the Office of Science, work with other offices of DOE to ensure that sufficient investment is made with adequate lead time to prepare science applications for the post-Moore era. While the adaptations that ECP application teams are starting to make

for supporting current and emerging heterogeneous execution environments is good preparation for some of the anticipated post-exascale technologies, additional investments will be needed to explore the newer computing paradigms that will emerge in the post-exascale and post-Moore timeframes.

There are multiple dimensions to investing in the readiness of science applications. First, preparing applications for new computing paradigms will be critical in the post-Moore era. It is observed that, while the Exascale Computing Project (ECP) has been structured to achieve the important goal of delivering an exascale system early in the next decade, it has also dampened efforts to explore the new paradigms that will be necessary for post-exascale and post-Moore computing. This dampening was intensified when the ECP delivery timeline was reduced, and there is additional risk that pressure to deliver to the deadline will further narrow research exploration as part of ECP efforts. Thus, investing in application readiness will also require renewed investments in research in the areas of applied mathematics (e.g., exploring new models of computer arithmetic) and algorithms, which in turn will need to be tightly coupled with the development of new computation and data models in different science domains that will be necessary for the new computing paradigms. Second, this investment will require continued partnership between the Office of Science and other DOE offices, as is done in SciDAC and other joint programs. Third, a clear methodology will need to be established for making migration vs. rewrite decisions for different applications in different timeframes, as new technologies are adopted. Finally, the Office of Science should invest in organizing early workshops on post-Moore application readiness, as was done for exascale application readiness.

6.3 Investing in Research related to Platforms with Open Hardware interfaces and components

Recommendation 3: DOE should invest in research to help foster an ecosystem with open hardware interfaces and components as part of the future HPC technology roadmap

The findings in this study have identified a growing trend in the use of open hardware interfaces and components, which is expected to increase in the post-exascale and post-Moore eras, relative to current and past approaches for hardware acquisition. In the interest of future Science needs, the subcommittee recommends that the Office of Science foster this ecosystem by investing in research related to platforms with open hardware components, i.e., platforms built using open interfaces that support high-performance and reliable integration of open hardware components with proprietary components from different hardware providers.

There are many reasons behind this recommendation. First, post-Moore hardware will require more innovation and agility in hardware design than in past decades, and an open platform approach will help foster this innovation while also mitigating risks associated with selecting a single vendor for hardware acquisition. There is a long history of DOE-sponsored research influencing industry hardware standards, and it is reasonable to expect that DOE's investment in this research will in turn influence future standards for open hardware platforms Second, the trend towards extreme heterogeneity in post-Moore computing reinforces the importance of integrating hardware components developed by different hardware providers. While these components will continue to be proprietary in many cases, it will be important to allow for the possibility of also integrating open source hardware components where appropriate. (The subcommittee recognizes that there are many obstacles to enabling the use of open source hardware components in production systems, but also sees an analogy here with the early skepticism to the use of open source software components that are now commonplace in production systems.) Finally, research investment is necessary because existing approaches to open interfaces are highly impoverished in both perfor-

mance and reliability; new approaches are needed to overcome these limitations so as to ensure that leadership-class HPC hardware can be built for future science applications by tightly integrating the best technologies from different hardware providers (proprietary or open source).

6.4 Investing in Research related to System Software

Recommendation 4: DOE should invest in research to help advance system software technologies to support post-Moore computing

The findings in this study have identified the need for advancing system software to meet the requirements of post-Moore computing. In the interest of future Science needs, the subcommittee recommends that the Office of Science ensure this advancement by investing in research related to open source and proprietary system software for future HPC technologies. In terms of synergies with mainstream computing, many of the system software capabilities needed to map science applications on future HPC systems will also be beneficial to commercial computing. The DOE should support active and sustained efforts to contribute to relevant software projects to ensure that HPC concerns such as performance isolation, low latency communication, and diverse wide area workflows are addressed in the design and adoption of system software for future HPC platforms.

There are many reasons behind this recommendation. First, over the past decades, DOE investments have helped ensure a successful history of using advances in system software to enable production DOE applications to run on leadership HPC systems. However, the current system software stack are built on technology foundations that are more than two decades old, and are ill-prepared for the new computing paradigms anticipated in post-Moore computing, e.g., new storage technologies to hold the every-increasing experimental and observational science datasets, tighter integration of accelerators and co-processors than in the past, and new hardware consistency models for communication, coherence, and synchronization among different hardware components. Second, the combination of open hardware platforms and open source system software will enable software/hardware co-design to occur with the agility needed in post-Moore timeframe. Finally, system software has a longer history of reducing the impact of hardware disruptions on application software, and this role will be even more important in the context of future HPC technologies.

6.5 Early Testbeds in DOE Computing Facilities

Recommendation 5: DOE computing facilities should prepare users for post-Moore computing by providing and supporting early access to testbeds and small-scale systems

The findings in this study have identified the need for providing users of DOE computing facilities early access to timing accurate emulators, testbeds and small-scale systems that are exemplars of systems expected in the post-Moore computing roadmap. The subcommittee recommends that the Office of Science's computing facilities address this need by acquiring such emulators, testbeds and small-scale systems, and providing and supporting access to these systems by current HPC users. The investments in Recommendations 2, 3, 4 will help create a community of researchers that can assist computing facilities staff in training activities related to these early testbeds. This recommendation is synergistic with the conclusions of a recent ASCR workshop on facility requirements for supporting computer science research [65].

There are multiple facets to this recommendation. The acquisition of such testbeds will require building relationships with hardware providers who are exploring new post-Moore technologies, some of whom may not have had past relationships with DOE facilities. The subcommittee believes that creating these new relationships will help foster a broader ecosystem of partners for future HPC

systems. Further, to address the need for educating HPC users on future technologies, the support for these testbeds will need to extend beyond system support, and also include training, workshops, as well as fostering of user groups for different systems. The subcommittee also recognizes that labor costs (personnel, training, etc.) will be a more significant fraction of the cost of deploying a testbed small-scale system, relative to the labor cost fraction in leadership facilities, but believes that this human investment is important for recruiting, growing and retaining talent (as discussed in the next recommendation). Finally, the subcommittee understands that this recommendation for DOE facilities must not distract from current exascale commitments, and trusts that investment in small-scale future HPC testbeds will be possible in the pre-exascale timeframe, with the goal of increased investments in this direction in the post-exascale era.

6.6 Recruiting, Growing and Retaining Talent for the post-Moore era

Recommendation 6: Recruit and grow workforce members who can innovate in all aspects of mapping applications onto emerging post-Moore hardware, with an emphasis on recognizing top talent in this area

The findings in this study have identified the need for significant innovation in support of the enablement of science applications on post-Moore hardware. The subcommittee recommends that DOE national laboratories prioritize the recruiting and nurturing of top talent in all aspects of mapping applications onto emerging post-Moore hardware, including skills and talent related to development of science applications, applied mathematics research, system software research, and hardware research for future platforms.

The context for this recommendation lies in observations that have been made in past ASCAC studies with respect to the increasing challenge of retaining talent in computing-related areas, give their high demand in the commercial sector. This challenge will continue to increase as companies start to develop their post-Moore computing strategies. However, the subcommittee believes that DOE national laboratories have unique opportunities to build a talent pipeline in this area, because it is expected that the DOE labs will explore post-Moore technologies in an earlier timeframe than many industry labs, which can be attractive to technical personnel who are passionate about working with cutting-edge technologies. Building the necessary workforce pipeline will require prioritization of post-Moore technologies in all avenues related to recruiting, growth and retention, including CSGF fellowships, postdoctoral appointments (including prestigious named postdoctoral fellowships), LDRD-funded projects, and recognition (through awards and other channels) of top talent in this area. In addition, building partnerships in post-Moore technology areas with interested and qualified faculty members in academia through established mechanisms, such as recruiting their students for internships, hosting them for sabbaticals, and joint faculty appointments, can further help with strengthening the talent pipeline that will be needed in DOE laboratories in the post-Moore era.

Chapter 7

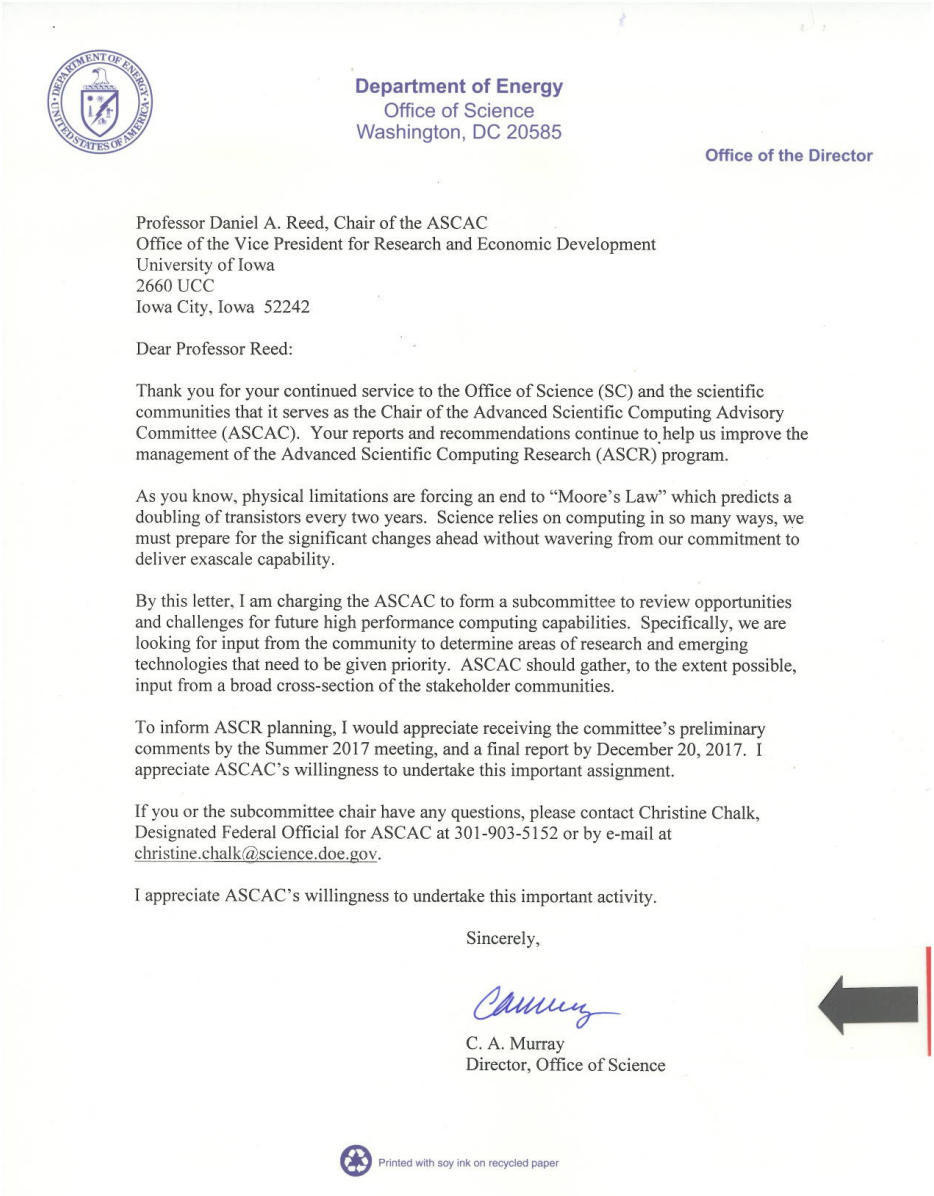
Conclusions

This report reviewed opportunities and challenges for future high performance computing capabilities, with a focus on the use of computing for the advancement of Science. The review drew from scientific publications, presentations, reports and expert testimony. The report includes key findings and recommendations from the perspective of the post-exascale and post-Moore timeframes. While the subcommittee appreciated the timeliness of the charge, we acknowledge that a single study cannot provide a comprehensive answer to identifying research opportunities and challenges for future HPC capabilities in the post-exascale and post-Moore timeframes, which span multiple decades, and trust that there will be follow-on studies to elaborate further on these challenges and opportunities as details of emerging HPC technologies become clearer in the coming years.

An overarching concern that emerged from the subcommittee's findings and recommendations is that DOE has lost considerable momentum in funding and sustaining a research pipeline in the applied math and computer science areas that should have been the seed corn for preparing for these future challenges, and it is therefore critical to correct this gap as soon as possible. While the subcommittee understands the paramount importance of DOE's commitment to deliver exascale capability, it is also critical to fund research and development that look beyond the ECP time horizon. The recommendations in this report highlight areas of research and emerging technologies that need to be given priority in this regard (application readiness, open hardware platforms, system software), as well as supporting activities that are essential for success (post-Moore strategy leadership, early testbeds in DOE facilities, and recruitment, growth and retention of top talent in post-Moore technology areas). While these recommendation areas were identified from the perspective of this study, the subcommittee firmly believes that sustaining a research pipeline in the applied math and computer science areas in general is also of paramount importance to ASCR's future.

Appendix A

Charge to Subcommittee



Department of Energy
Office of Science
Washington, DC 20585

Office of the Director

Professor Daniel A. Reed, Chair of the ASCAC
Office of the Vice President for Research and Economic Development
University of Iowa
2660 UCC
Iowa City, Iowa 52242

Dear Professor Reed:

Thank you for your continued service to the Office of Science (SC) and the scientific communities that it serves as the Chair of the Advanced Scientific Computing Advisory Committee (ASCAC). Your reports and recommendations continue to help us improve the management of the Advanced Scientific Computing Research (ASCR) program.

As you know, physical limitations are forcing an end to "Moore's Law" which predicts a doubling of transistors every two years. Science relies on computing in so many ways, we must prepare for the significant changes ahead without wavering from our commitment to deliver exascale capability.

By this letter, I am charging the ASCAC to form a subcommittee to review opportunities and challenges for future high performance computing capabilities. Specifically, we are looking for input from the community to determine areas of research and emerging technologies that need to be given priority. ASCAC should gather, to the extent possible, input from a broad cross-section of the stakeholder communities.

To inform ASCR planning, I would appreciate receiving the committee's preliminary comments by the Summer 2017 meeting, and a final report by December 20, 2017. I appreciate ASCAC's willingness to undertake this important assignment.

If you or the subcommittee chair have any questions, please contact Christine Chalk, Designated Federal Official for ASCAC at 301-903-5152 or by e-mail at christine.chalk@science.doe.gov.

I appreciate ASCAC's willingness to undertake this important activity.

Sincerely,

C. A. Murray
Director, Office of Science

Printed with soy ink on recycled paper

Appendix B

Subcommittee Members

The ASCAC Subcommittee on Future High Performance Computing Capabilities consisted of the following members:

- Keren Bergman, Columbia University, ASCAC member.
- Tom Conte, Georgia Institute of Technology.
- Al Gara, Intel Corporation.
- Maya Gokhale, Lawrence Livermore National Laboratory.
- Mike Heroux, Sandia National Laboratories.
- Peter Kogge, University of Notre Dame.
- Bob Lucas, Information Sciences Institute.
- Satoshi Matsuoka, Tokyo Tech., ASCAC member.
- Vivek Sarkar, Georgia Institute of Technology, ASCAC member (subcommittee chair).
- Olivier Temam, Google.

Appendix C

Bibliography

[1] T. M. Conte, E. P. DeBenedictis, P. A. Gargini, and E. Track. Rebooting computing: The road ahead. *Computer*, 50(1):20–29, Jan. 2017.

[2] Jeffrey S. Vetter, Ron Brightwell, Maya Gokhale, Pat McCormick, Rob Ross, John Shalf, Katie Antypas, David Donofrio, Travis Humble, Catherine Schuman, Brian Van Essen, Shinjae Yoo, Alex Aiken, David Bernholdt, Suren Byna, Kirk Cameron, Frank Cappello, Barbara Chapman, Andrew Chien, Mary Hall, Rebecca Hartman-Baker, Zhiling Lan, Michael Lang, John Leidel, Sherry Li, Robert Lucas, John Mellor-Crummey, Paul Peltz Jr., Thomas Peterka, Michelle Strout, and Jeremiah Wilke. Extreme Heterogeneity 2018: DOE ASCR Basic Research Needs Workshop on Extreme Heterogeneity, December 2018.

[3] Gordon E. Moore. Cramming more components onto integrated circuits. *Electronics*, 38(8), April 1965.

[4] G. E. Moore. Progress in digital integrated electronics [technical literaiture, copyright 1975 ieee. reprinted, with permission. technical digest. international electron devices meeting, ieee, 1975, pp. 11-13.]. *IEEE Solid-State Circuits Society Newsletter*, 11(3):36–37, Sept 2006.

[5] R.H. Dennard, F.H. Gaensslen, V.L. Rideout, E. Bassous, and A.R. LeBlanc. Design of ion-implanted mosfet's with very small physical dimensions. *Solid-State Circuits, IEEE Journal of*, 9(5):256–268, October 1974.

[6] et al. P. Gargini. Ieee international roadmap for devices and systems. Technical report, 2017.

[7] Haohuan Fu, Junfeng Liao, Nan Ding, Xiaohui Duan, Lin Gan, Yishuang Liang, Xinliang Wang, Jinzhe Yang, Yan Zheng, Weiguo Liu, Lanning Wang, and Guangwen Yang. Redesigning cam-se for peta-scale climate modeling performance and ultra-high resolution on sunway taihulight. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '17, pages 1:1–1:12, New York, NY, USA, 2017. ACM.

[8] Martin Berzins, Justin Luitjens, Qingyu Meng, Todd Harman, Charles A. Wight, and Joseph R. Peterson. Uintah: A scalable framework for hazard analysis. In *Proceedings of the 2010 TeraGrid Conference*, TG '10, pages 3:1–3:8, New York, NY, USA, 2010. ACM.

[9] Michael Feathers. *Working Effectively with Legacy Code*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2004.

[10] David E. Shaw, J. P. Grossman, Joseph A. Bank, Brannon Batson, J. Adam Butts, Jack C. Chao, Martin M. Deneroff, Ron O. Dror, Amos Even, Christopher H. Fenton, Anthony Forte,

Joseph Gagliardo, Gennette Gill, Brian Greskamp, C. Richard Ho, Douglas J. Ierardi, Lev Iserovich, Jeffrey S. Kuskin, Richard H. Larson, Timothy Layman, Li-Siang Lee, Adam K. Lerer, Chester Li, Daniel Killebrew, Kenneth M. Mackenzie, Shark Yeuk-Hai Mok, Mark A. Moraes, Rolf Mueller, Lawrence J. Nociolo, Jon L. Peticolas, Terry Quan, Daniel Ramot, John K. Salmon, Daniele P. Scarpazza, U. Ben Schafer, Naseer Siddique, Christopher W. Snyder, Jochen Spengler, Ping Tak Peter Tang, Michael Theobald, Horia Toma, Brian Towles, Benjamin Vitale, Stanley C. Wang, and Cliff Young. Anton 2: Raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '14, pages 41–53, Piscataway, NJ, USA, 2014. IEEE Press.

[11] Max M. Shulaker, Gage Hills, Nishant Patil, Hai Wei, Hong-Yu Chen, H. S. Philip Wong, and Subhasish Mitra. Carbon nanotube computer. *Nature*, 501:526 EP –, 09 2013.

[12] M. Gokhale, W. Holmes, A. Kopser, S. Lucas, R. Minnich, D. Sweely, and D. Lopresti. Building and using a highly parallel programmable logic array. *IEEE Computer*, pages 81–89, January 1991.

[13] P. Bertin, D. Roncin, and J. Vuillemin. Programmable active memories: a performance assessment. In G. Borriello and C. Ebeling, editors, *Research on Integrated Systems: Proceedings of the 1993 Symposium*, pages 88–102, 1993.

[14] Seth Copen Goldstein, Herman Schmit, Matthew Moe, Mihai Budiu, Srihari Cadambi, R. Reed Taylor, and Ronald Laufer. Piperench: A co/processor for streaming multimedia acceleration. In *Proceedings of the 26th Annual International Symposium on Computer Architecture*, ISCA '99, pages 28–39, Washington, DC, USA, 1999. IEEE Computer Society.

[15] E. Waingold, M. Taylor, D. Srikrishna, V. Sarkar, W. Lee, V. Lee, J. Kim, M. Frank, P. Finch, R. Barua, J. Babb, S. Amarasinghe, and A. Agarwal. Baring it all to software: Raw machines. *Computer*, 30(9):86–93, Sep 1997.

[16] Guangming Lu, Ming-hau Lee, Hartej Singh, Nader Bagherzadeh, Fadi J. Kurdahi, and Eliseu M. Filho. *MorphoSys: a reconfigurable processor targeted to high performance image application*, pages 661–669. Springer Berlin Heidelberg, Berlin, Heidelberg, 1999.

[17] MathStar. Mathstar corp. https://en.wikipedia.org/wiki/MathStar, accessed 2017.

[18] Ambric. Ambric corp. https://en.wikichip.org/wiki/ambric/am2000, accessed 2017.

[19] Norman P. Jouppi, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates, Suresh Bhatia, Nan Boden, Al Borchers, Rick Boyle, Pierre-luc Cantin, Clifford Chao, Chris Clark, Jeremy Coriell, Mike Daley, Matt Dau, Jeffrey Dean, Ben Gelb, Tara Vazir Ghaemmaghami, Rajendra Gottipati, William Gulland, Robert Hagmann, Richard C. Ho, Doug Hogberg, John Hu, Robert Hundt, Dan Hurt, Julian Ibarz, Aaron Jaffey, Alek Jaworski, Alexander Kaplan, Harshit Khaitan, Andy Koch, Naveen Kumar, Steve Lacy, James Laudon, James Law, Diemthu Le, Chris Leary, Zhuyuan Liu, Kyle Lucke, Alan Lundin, Gordon MacKean, Adriana Maggiore, Maire Mahony, Kieran Miller, Rahul Nagarajan, Ravi Narayanaswami, Ray Ni, Kathy Nix, Thomas Norrie, Mark Omernick, Narayana Penukonda, Andy Phelps, Jonathan Ross, Amir Salek, Emad Samadiani, Chris Severn, Gregory Sizikov, Matthew Snelham, Jed Souter, Dan Steinberg, Andy Swing, Mercedes Tan, Gregory Thorson, Bo Tian, Horia Toma, Erick Tuttle, Vijay Vasudevan, Richard Walter, Walter Wang, Eric

Wilcox, and Doe Hyun Yoon. In-datacenter performance analysis of a tensor processing unit. *CoRR*, abs/1704.04760, 2017.

[20] Brian Van Essen, Chris Macaraeg, Ryan Prenger, and Maya Gokhale. Accelerating a random forest classifier: multi-core, gp-gpu, or fpga. *IEEE International Symposium on Field Programmable Custom Computing Machines (FCCM)*, April 2012.

[21] Maya Gokhale and Paul S. Graham. *Reconfigurable Computing: Accelerating Computation with Field Programmable Gate Arrays*. Springer Verlag, 2005.

[22] Christian de Schryver. *FPGA Based Accelerators for Financial Applications*. Springer Publishing Company, Incorporated, 1st edition, 2015.

[23] Stefano Cherubin, Giovanni Agosta, Imane Lasri, Erven Rohou, and Olivier Sentieys. Implications of reduced-precision computations in hpc: Performance, energy and error. *International Conference on Parallel Computing (ParCo)*, September 2017.

[24] Khronos. Opencl. https://www.khronos.org/opencl/, accessed 2017.

[25] Zachary Jacobs, Keith Morgan, Michael Caffrey, Joseph Palmer, and Lauren Ho. LANL CubeSat Reconfigurable Computer (CRC). August 2010. Presented at CubeSat Summer Workshop 2010.

[26] Vladimir Marjanović, José Gracia, and Colin W Glass. Performance modeling of the hpcg benchmark. In *High Performance Computing Systems. Performance Modeling, Benchmarking, and Simulation*, pages 172–192. Springer International Publishing, 2014.

[27] Sébastien Rumley, Madeleine Glick, Simon D. Hammond, Arun Rodrigues, and Keren Bergman. *Design Methodology for Optimizing Optical Interconnection Networks in High Performance Systems*, pages 454–471. Springer International Publishing, Cham, 2015.

[28] D. M. Kuchta, T. N. Huynh, F. E. Doany, L. Schares, C. W. Baks, C. Neumeyr, A. Daly, B. Kögel, J. Rosskopf, and M. Ortsiefer. Error-free 56 gb/s nrz modulation of a 1530-nm vcsel link. *Journal of Lightwave Technology*, 34(14):3275–3282, July 2016.

[29] B. G. Lee, A. Biberman, J. Chan, and K. Bergman. High-performance modulators and switches for silicon photonic networks-on-chip. *IEEE Journal of Selected Topics in Quantum Electronics*, 16(1):6–22, Jan 2010.

[30] N. Ophir, C. Mineo, D. Mountain, and K. Bergman. Silicon photonic microring links for high-bandwidth-density, low-power chip i/o. *IEEE Micro*, 33(1):54–67, Jan 2013.

[31] AIM Photonics. Aim photonics web site. `http://www.aimphotonics.com`.

[32] Nicholas C Harris, Gregory R Steinbrecher, Mihika Prabhu, Yoav Lahini, Jacob Mower, Darius Bunandar, Changchen Chen, Franco NC Wong, Tom Baehr-Jones, Michael Hochberg, et al. Quantum transport simulations in a programmable nanophotonic processor. *Nature Photonics*, 11(7):447–452, 2017.

[33] Yichen Shen, Nicholas C. Harris, Scott Skirlo, Mihika Prabhu, Tom Baehr-Jones, Michael Hochberg, Xin Sun, Shijie Zhao, Hugo Larochelle, Dirk Englund, and Marin Soljačić. Deep learning with coherent nanophotonic circuits. *Nature Photonics*, 11, 06 2017.

[34] Ivan K. Schuller and Rick Stevens. Neuromorphic Computing: From Materials to Systems Architcture. Report of a Roundtable Convened to Consider Neuromorphic Computing Basic Research Need.s. October 2015.

[35] Stephen Jordan. Quantum algorithm zoo. `http://math.nist.gov/quantum/zoo`.

[36] Tameem Albash, Victor Martin-Mayor, and Itay Hen. Temperature scaling law for quantum annealing optimizers. *Physical review letters*, 119(11):110502, 2017.

[37] A. Esposito. A simplified method for analyzing circuits by analogy. *Machine Design*, pages 173–177, October 1969.

[38] A. S. Jackson. *Analog Computation*. McGraw-Hill, 1960.

[39] A. K. Dewdney. On the spaghetti compute and other analog gadgets for problem solving. *Scientific American*, 250(6):19–26, June 1984.

[40] T. Hylton. On thermodynamics and the future of computing. IEEE, November 2017.

[41] N. Ganesh. A thermodynamic treatment of intelligent systems. IEEE, November 2017.

[42] B. W. Lee and B. J. Sheu. *Hardware Annealing in Analog VLSI Neurocomputing*. Kluwer Academic Publishers, 1991.

[43] J. C. Lee, B. J. Sheu, W. C. Fang, and R. Chellappa. Vlsi neuroprocessors for video motion detection. *IEEE Transactions on Neural Networks*, 4(2):178–191, Mar 1993.

[44] https://www.dwavesys.com/quantum-computing.

[45] D. Peter et al. The helmholtz machine. *Neural Computation*, 7(5):889–904, 1995.

[46] Lei Xu, Michael I. Jordan, and Geoffrey E Hinton. An alternative model for mixtures of experts. In G. Tesauro, D. S. Touretzky, and T. K. Leen, editors, *Advances in Neural Information Processing Systems 7*, pages 633–640. MIT Press, 1995.

[47] Jeremy L. England. Dissipative adaptation in driven self-assembly. *Nature nanotechnology*, 10(11):919–923, November 2015.

[48] Gavin E. Crooks. Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Physical Review E*, 60(3):2721, March 1999.

[49] Rosemary J. Harris and Gunther M. Schutz. Fluctuation theorems for stochastic dynamics. *Journal of Statistical Mechanics: Theory and Experiment*, 2007(07):P07020, July 2007.

[50] Nikolay Perunov, Robert A. Marsland, and Jeremy L. England. Statistical physics of adaptation. *Physical Review X*, 6(2):021036, 2016.

[51] Global semiconductor sales increase 5 percent year-over-year in october; industry forecast revised upward, December 2016.

[52] Will analog be as good tomorrow as it was yesterday. *McKinsey on Semiconductors*, 2011.

[53] RISC-V Foundation. Risc-v website. `https://riscv.org`.

[54] Wikipedia. Wikipedia website. https://en.wikipedia.org/wiki/Comparison_of_deep_learning_software.

[55] Innovative Computing Laboratory. Icl website. http://icl.utk.edu/bblas.

[56] A. Danalis, G. Marin, C. McCurdy, J.S. Meredith, P.C. Roth, K. Spafford, V. Tipparaju, and J.S. Vetter. The scalable heterogeneous computing (shoc) benchmark suite. In *ACM Workshop on General-Purpose Computation on Graphics Processing Units (GPGPU)*, pages 63–74, Pittsburgh, Pennsylvania, 2010. ACM.

[57] Scott Lloyd and Maya Gokhale. Evaluating the feasibility of storage class memory as main memory. *International Symposium on Memory Systems MEMSY16*, October 2016.

[58] CENATE. Cenate web site. https://cenate.pnnl.gov.

[59] LLNL. Brain-inspired supercomputer web site. https://www.llnl.gov/news/lawrence-livermore-and-ibm-collaborate-build-new-brain-inspired-supercomputer.

[60] LLNL. Catalyst web site. https://computation.llnl.gov/computers/catalyst.

[61] D-Wave. Los alamos d-wave 2x web site. https://www.dwavesys.com/press-releases/los-alamos-national-laboratory-orders-1000-qubit-d-wave-2x-quantum-computer.

[62] J.S. Vetter, R. Glassbrook, J. Dongarra, K. Schwan, B. Loftis, S. McNally, J. Meredith, J. Rogers, P. Roth, K. Spafford, and S. Yalamanchili. Keeneland: Bringing heterogeneous gpu computing to the computational science community. *IEEE Computing in Science and Engineering*, 13(5):90–95, 2011.

[63] CRNCH. Rogues gallery website. http://crnch.gatech.edu/rogues-gallery.

[64] J. Chen, Alok Choudhary, S. Feldman, B. Hendrickson, C. R. Johnson, R. Mount, V. Sarkar, V. White, and D. Williams. *Synergistic Challenges in Data-Intensive Science and Exascale Computing: DOE ASCAC Data Subcommittee Report*. Department of Energy Office of Science, March 2013. Type: Report.

[65] J. Vetter, A. Almgren, P. DeMar, K. Riley, K. Antypas, D. Bard, R. Coffey, E. Dart, S. Dosanjh, and R. Gerber. Advanced scientific computing research exascale requirements review. an office of science review sponsored by advanced scientific computing research, september 27-29, 2016, rockville, maryland. Technical report, Argonne National Lab.(ANL), Argonne, IL (United States). Argonne Leadership Computing Facility, 2017.