Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Perspective:

Moving from exascale to zettascale computing: challenges and techniques*

Xiang-ke LIAO, Kai LU†‡, Can-qun YANG, Jin-wen LI, Yuan YUAN, Ming-che LAI, Li-bo HUANG, Ping-jing LU, Jian-bin FANG, Jing REN, Jie SHEN

College of Computer, National University of Defense Technology, Changsha 410073, China

†E-mail: kailu@nudt.edu.cn

Received Aug. 16, 2018; Revision accepted Sept. 14, 2018; Crosschecked Oct. 15, 2018

Abstract: High-performance computing (HPC) is essential for both traditional and emerging scientific fields, enabling scientific activities to make progress. With the development of high-performance computing, it is foreseeable that exascale computing will be put into practice around 2020. As Moore's law approaches its limit, high-performance computing will face severe challenges when moving from exascale to zettascale, making the next 10 years after 2020 a vital period to develop key HPC techniques. In this study, we discuss the challenges of enabling zettascale computing with respect to both hardware and software. We then present a perspective of future HPC technology evolution and revolution, leading to our main recommendations in support of zettascale computing in the coming future.

Key words: High-performance computing; Zettascale; Micro-architectures; Interconnection; Storage system; Manufacturing process; Programming models and environments

https://doi.org/10.1631/FITEE.1800494 **CLC number:** TP311

1 Introduction

High-performance computing (HPC) has been playing an important role in key scientific fields such as cosmology, geoscience, life science, and medical science. Since the first release of the Top500 list in 1993, we have noted a significant development of HPC in terms of computing capability to meet the ever increasing computing requirements. Exascale (10^{18} operations/s) computing systems are expected to be deployed in 2020–2022. To this end, the United States, Japan, Europe, and China have proposed exascale computing projects to accelerate delivery of a capable exascale computing ecosystem.

In Fig. 1, the projected performance shows that the performance increase of the No. 1 systems slowed down around 2013, and it was the same for the sum performance. As the speed of performance increase slows down, the HPC systems are likely to reach the next milestone, zettascale computing (10^{21} operations/s), by 2035. We note that this goal agrees with that in Xu et al. (2016). Thus, it can be predicted that the next 10 years after the exascale era, i.e., 2020–2030, will be a critical time-frame of moving from exascale to zettascale computing.

Nonetheless, we will face unprecedented challenges in the enabling technologies during this time window. This is because enabling the performance increases is a multifaceted systematic problem relating to computer architectures, manufacturing process, programming models, and runtime systems. In this study, we analyze the technical HPC challenges of moving from exascale to zettascale computing, overview the innovative technical approaches to overcome these barriers, and then propose our suggestions for realizing a smooth transition and delivering sustainable HPC systems in the post-exascale era.

‡ Corresponding author

* Project supported by the National Key Technology R&D Program of China (No. 2016YFB0200401)

ORCID: Kai LU, http://orcid.org/0000-0003-2284-7897

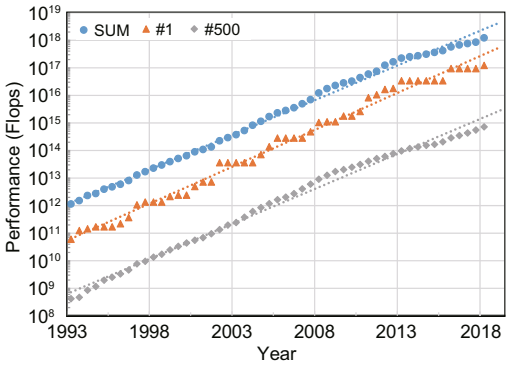

Fig. 1 Projected performance of the Top500 high-performance computing systems from 1993 to 2020. The plot is based on a data source publicly available on the Top500 website (https://www.top500.org/)

2 Future technical challenges in high-performance computing

Before exascale computing, the performance of HPC systems increases with the advancement of the semiconductor technology, which follows Moore's law. With Moore's law approaching its limit, we will face many more unprecedented challenges than ever in realizing a sustainable zettascale computing system. These challenges are created by process limit, power consumption, communication, memory and storage, reliability, and programming.

2.1 Challenges in the manufacturing process

In the past decades, each new manufacturing process leads to smaller, faster transistors, and the regular introduction of new techniques keeps Moore's law ticking along Waldrop (2016). At present, the semiconductor process technology has been extended to below 10 nm. The challenge of further upgrading the microprocessor process is that it is difficult to arrange transistors at a smaller scale with the existing material and technology. These small-scale devices cannot be simply analyzed with the knowledge of classical physics; the theory of quantum mechanics is necessary. This makes the circuit design more complex. In addition, considering the atom size, the volume of some devices cannot be reduced, which further limits the reduction of the chip size. To sum up, the traditional technology roadmap based on feature size reduction may end after three to four technology generations, so the design of post-exascale high-performance computers will face unprecedented challenges.

2.2 Challenges in power consumption

At present, all the leading countries in supercomputing have developed their exascale computing projects, aiming to reach a power efficiency of 30–50 Gflops/W, which leads to a power consumption budget of 20–30 MW (Lucas et al., 2014). However, in the latest Green500 in June 2018 (https://www.top500.org/green500/lists/2018/06/), the ZettaScaler system with the highest energy efficiency can reach only 18.4 Gflops/W. Therefore, energy consumption will be a major technical obstacle to the implementation of future high-performance computers. Because the Dennard scaling law broke down earlier than expected, it is foreseeable that the power consumption of exascale computer systems will be unbearable if the silicon process and the current mainstream architecture are still used. Therefore, we need to explore new technology, such as low-power devices and components, low-power heterogeneous computer architectures, energy-aware system scheduling and compilation technology, and low-power systems and cooling technology, to deal with the power challenge of future high-performance computing.

2.3 Challenges in interconnection

The performance increase of HPC comes mainly from the enhancement of the per-node computing capability and the increase in the number of computing nodes. With the increase of single-node computing capacity, the node communication bandwidth must be increased accordingly to build a more balanced system. Take Tianhe-2 as an example. Its single-node computing performance is 3 Tflops, and the node communication bandwidth is 112 Gb/s. Thus, we can calculate that the node bandwidth-performance ratio is 0.037. In future exascale computing systems, the single-node computing power should be around 10 Tflops. To maintain the node bandwidth-performance ratio at 0.04, the node communication bandwidth needs to increase to 400 Gb/s, which is much higher than what can be reached according to the current SerDes performance growth trend. In the post-exascale era, high-performance computing systems will definitely face many more severe challenges such as unbearable system power consumption, difficulty in network topology implementation, significant increase in transmission delay,

difficulty in ensuring system reliability, and difficulty in increasing the density of interconnection network engineering (Lucas et al., 2014).

2.4 Challenges in the storage system

The significant separation between the processor and memory industries has led to the situation in which their developments are severely out of sync (Wulf and McKee, 1995). As a result, current processors and memory systems present a significant performance gap. In the past 20 years, the processor performance increased by 55% each year, whereas the memory access performance increased by only 10% each year. Consequently, the latency of current processors is around 0.3 ns, whereas that of current memory systems is as long as 90 ns, and their performance gap keeps increasing. Apparently, this huge performance gap has been greatly restricting the sustainable development of HPC.

In the post-exascale era, billions of processes will concurrently run I/O operations, accessing hundreds of peta-byte (PB) data and requiring dozens of TB/s bandwidth. It will bring great challenges to storage systems in terms of scalability, fault tolerance, and usage efficiency. With the integration of high-performance computing, intelligent computing, and big data analytics, workloads running on supercomputers will be more complex and diverse. As a result, new challenges are put forward in the architecture design of future HPC storage systems.

2.5 Challenges in reliability

With the growing scale of high-performance computers, the software stack is becoming more and more complex, and its reliability has become a major concern. At the scale of 10P, the average meantime between failures (MTBF) is only around five hours (Glosli et al., 2007; Schroeder and Gibson, 2007). As the failure rate of a single processor remains relatively stable, the system failure rate increases as the scale of the system expands and is proportional to the number of processors in the system. Accordingly, when the system continues to develop to exascale or zettascale, the average time of failure will be far less than one hour, and its reliability problem will be more serious. In addition, in the actual use of high-performance computers, failure of the computing task caused by system software vulnerability or programming errors is also an important reason for system unavailability. Therefore, effective failure detection and diagnosis will be one of the key technologies for high-performance computers in the future.

2.6 Challenges in programming

Future HPC computing systems will have a total of tens of millions of parallelism degrees. It is a great challenge for programming models to manage such a large amount of parallelism in terms of programming, debugging, and tuning. On one hand, programming models have to explicitly represent the heterogeneous and hierarchical parallelism, have control of data locality, and provide an interface to control power capping and reliability, to minimize the overhead of data movement and unlock performance potentials. On the other hand, we have to develop domain-specific frameworks and/or libraries to improve the level of abstraction and enhance programmers' productivity.

New programming models come along with the emergence of new computing devices. For example, quantum computing will likely to be used to solve the unsolvable problems in domains such as quantum encryption, prime number decomposition, and scene planning (Chong et al., 2017). In post-exascale computing, it is likely that quantum computers will co-exist with classical computers. To this end, vendors have to provide programmers with quantum programming languages, quantum compilers, and runtime. These new programming models differ from classical ones, which steepens the learning curve. Therefore, we have to compensate for the difference in programming models as much as we can, thus lowering the learning curve from classical programming paradigms to new ones.

3 Future high-performance computing technology evolution and revolution

To address the technical challenges in the zettascale computing era, researchers from industry and academia have to work together in both hardware and software to enable a sustainable zettascale system. In the 2020 to 2030 time-frame, we will have to investigate new enabling technologies in micro-architecture, interconnection, storage, manufacturing process, programming model, and runtime environment.

3.1 Architecture

Dedicated processors will become the most attractive high-performance processing units for their high computing efficiency, but the utilization of a dedicated accelerator usually means partial loss of general programmability. Considering that future large-scale computer construction is limited mainly to the processor speed and its efficiency, we argue that heterogeneous computing architectures combined with general-purpose central processing units (CPUs) and dedicated accelerators will still be the main approach, regardless of how they are realized. Further, since conventional HPC applications and emerging intelligent computing applications (such as deep learning) will both exist in the future, the processor design should take mixed precision arithmetic into consideration to support a large variety of application workloads.

Currently, dedicated processors include graphics processing units (GPUs), many integrated cores (MICs), and field programmable gate arrays (FPGAs). GPUs employ single-instruction multiple-thread architecture, which is good at processing data-intensive applications. MIC co-processors incorporate many general-purpose processor cores, which can use conventional programming models such as OpenMP, and thus programmers have a shallow learning curve. At the same time, we can use FPGAs to develop customized systems that are still programmable. Since there is no unified accelerator architecture, we argue that many accelerator forms may co-exist (Kim et al., 2017).

In 2020–2030, the processor architectures will evolve to adapt to applications by incorporating new advanced technologies such as 2.5D/3D integration. This can be used to reduce the difficulty of heterogeneous integration, and to relieve the memory wall. Such architecture could use more hierarchies of memory, non-volatile memory, and near-data processing to improve memory performance. In addition, the integration of inter-chip interconnection networks can reduce the communication latency and improve communication bandwidth, which can effectively relieve the communication wall. At the same time, the continuous decrease in transistor cost will provide more space for more heterogeneous components.

New computing devices based on the non-Von Neumann structure will become an important branch of future microprocessor architectures. As Moore's law reaches its limit, the traditional computing paradigm based on the Turing machine and the Von Neumann structure cannot provide unlimited computational power. Thus, we will have to investigate other ways such as quantum computing, biological computing, and optical computing, at the end of the complementary metal-oxide-semiconductor (CMOS) process. Many prototypes that follow new computing paradigms are still at the experimental stage and will take some time to get into mass production. We argue that the new computing paradigm will appear in the form of accelerators in the future and give new impetus to the development of HPC.

3.2 High-performance interconnecting technology

A balanced, scalable, and low-cost HPC interconnect system will be designed to achieve flexible allocation of network resources based on silicon photonics interconnects (Rumley et al., 2015). High-performance interconnect technology has been continuously advancing with increase in the communication line rate, and has continuously led the development of the international computer industry interconnection technology. The electrical interconnection rate will exceed 56 Gb/s and will adopt a higher pulse amplitude modulation (PAM) coding format and an ultra-short or very-short-range-transmission standard. The electrical interconnection rate will be increased by shortening the electrical signal distance. The optical interconnection will adopt the optical coarse wavelength division multiplexing (CWDM) transmission standard. The trend of 'replacing electronic switching with optical switching' is even clearer. The encoding method, drive capability, and optoelectronic mode of the 112 Gb/s high-speed interface will all undergo tremendous changes and will have a significant impact on the entire HPC interconnect system. In the post-exascale HPC era, the CWDM technology will use more optical channels, and new types of interconnect materials such as new photonic crystals and carbon nanotubes will gradually emerge. The interconnect speed will increase to 400 Gb/s, and the chip throughput will be in hundreds of Tb/s. At that time, we need to design a more balanced, scalable, and low-cost HPC interconnect system by redistributing the chip transistor resources and link resources.

High-density opto-electronic integration enables deep integration of interconnects and computation/storage, thus reducing latency and improving density power consumption. The huge gap between interconnects and computation/storage always exists in HPC. To achieve gains in performance and energy efficiency, it is likely that we will integrate interconnects and computation/storage into a single device with the single-chip 3D integration technology (Vinaik and Puri, 2015). Nowadays, the photonic integration technology has been evolving from a single functional device to large-scale integrated chips. Small- and medium-scale photonic integration technology has matured and has been widely used commercially. The integration degree of large-scale photonic integration has reached hundreds of components. Compared with the widely used discrete components, photonic integrated products have significant advantages in terms of size, power consumption, cost, and reliability, and are the mainstream development direction of optical devices in the future. To minimize the opto-electronic gap, the opto-electronic integrated process is gradually becoming compatible with the application-specific integrated circuit (ASIC) chip design process, which lays the foundation for deep opto-electronic integration.

An all-optical interconnecting system will be realized, with an aim to increase the scalability of HPC systems and reduce interconnection costs. In the post-exascale HPC era, the data traffic in HPC systems might increase explosively and the behaviors of applications might change during runtime. Therefore, we need all-optical switching technologies and software-defined network management to dynamically redistribute fiber resources according to the target application. In this way, we can dynamically reconfigure HPC interconnection topologies, increase HPC scalability, and reduce interconnection costs.

3.3 Emerging storage technology

Future HPC storage systems can have a much richer hierarchy. At the memory level, the traditional methods, which improve storage bandwidth by increasing the clock frequency of memory or width of the storage bus, are approaching their physical limits (Wilkes, 1995). The emerging 2.5D/3D stack memory technology brings new opportunities to break the memory wall (Jacob et al., 2009). In particular, the hybrid memory cube (HMC) and high bandwidth memory (HBM) are expected to be widely used in future HPC systems (Jeddeloh and Keeth, 2012). At the storage level, various types of non-volatile storage media (NVM) have emerged with the development of microelectronics and nanoelectronics (Xu et al., 2014). For example, 3D XPoint, co-developed by Intel and Micron, is regarded as the most promising NVM storage product in large-scale commercial use (Kolli et al., 2016). Due to the constraints in price, capacity, and programming modes, NVM has not been used independently in current HPC systems. Rather, they tend to use a hierarchical hybrid storage architecture of NVM+SSD+HDD. In the post-exascale era, the newly emerging high-density and low-power memory technologies have potential to dominate the market and gradually replace the traditional dynamic random access memory (DRAM) technology. The NVM technology will bring fundamental changes to the architecture and software stack of HPC storage systems, thus effectively mitigating the bottleneck in memory and I/O.

Integrating storage and interconnection can significantly improve the efficiency of future storage systems. Network attached memory (NAM) has become a hot research topic in exascale computing. Typical examples include The Machine, Summit, and ExaNoDe. In these systems, CPUs and large NVMs are interconnected via a storage network to implement storage pooling. Hewlett-Packard (HP) has built a revolutionary memory-driven computer, i.e., The Machine, with the universal memory architecture (Keeton, 2015). In this system, CPU cores, GPUs, network interfaces, and other customized processing units are all interconnected through a memory network. The experiment results demonstrate that this architecture can improve the overall performance by five times and transaction processing performance by 100 times, while using only 1.25% of the original power consumption. Furthermore, future optical interconnection technology can attach new NVM resources to the high-speed interconnection network, thus achieving the pooling of distributed storage resources (Mishra et al., 2013). We regard this as an effective way to mitigate the bottlenecks in memory and I/O.

The development of memristor and quantum computing enables a real integration of computing and storage (Xu et al., 2014). The HP laboratory proposed a 'state logic' model in 2010, which uses the

resistance states of the memristor to represent input and output data, and runs logic operations. In this way, the memristor further integrates storage and computation. In the post-exascale era, memristor-based storage systems will enter a practical stage, and related programming models will come as follows. The emerging quantum computation is another potential option to tackle the challenges in memory and I/O.

3.4 New manufacturing process

Computer engineering and technology is a general term for the enabling and implementation technologies in high-performance computer systems, and plays a key role in ensuring the reliable implementation and stable operations of the system (Ábrahám et al., 2015). Higher computing, storage, and interconnect density demands require that the capability of the engineering process be improved in various aspects such as computing devices, interconnections, printed circuit boards (PCBs), and cooling. Academia and industry have reached a consensus that a reliable, usable, and extensible system can be built only when we keep the system scale and the energy consumption at the current level (Cavin et al., 2012). This means that high-performance computers will continue to evolve towards higher density, in terms of computing, assembly, and energy consumption. In 2020–2030, we expect that the density per node will increase to 0.5–0.8 Pflops, and the energy consumption density will be around 0.8 Tflops/W.

With the continuous increase of system energy consumption, investigating efficient cooling technologies has become a key factor in developing sustainable HPC systems. Considering the pros and cons of various cooling technologies (e.g., air cooling, liquid cooling, and mixed cooling) and the trend of the current industry, future HPC systems will probably adopt more efficient liquid cooling technologies and particularly the evaporative cooling technology. We need to keep the power usage effectiveness (PUE) ratio at around 1.1. In the post-exascale era, the new cooling technology will enable us to recycle the heat produced by the HPC system, to further reduce the energy consumption of the cooling system. The system PUE ratio should be kept at around 1. In the future, the 2.5D/3D integrated circuit will adopt the chip-packaging cooling technique. At the same time, the cooling granularity will be extended from the cabinet/board level to the chip level, thus aiming to achieve the accurate targeted heat dissipation.

In addition, the optical technology will be widely used in high-speed signal design. With the rapid increase of the interconnection signal rate (bandwidth), the PCB loss of high-speed electrical signals is increasing, resulting in the deterioration of signal quality. With the continuous penetration of light into traditional electrical signals, the trend shows that light will replace electricity from the interconnection of cabinet, rack, board, and chip, to the silicon photonic interconnection.

3.5 Programming models and environments

Programming models are taken as the interface of using high-performance computing resources and the bridge between upper applications and underlying hardware. Thus, these models can provide effective support to make full use of HPC computing potential. From the projected trend of HPC performance, we argue that mining parallelism is still the most important way of enabling a sustainable computing system, unless there is a significant increase in single-chip performance.

In the long term, MPI+X will still be the most commonly used parallel programming model, where MPI is used as the message passing interface across large-scale computing nodes and the X models (e.g., OpenMP, OpenACC, OpenCL, and CUDA) are used as the intra-node threading models (Fang et al., 2011; Shen et al., 2013). These intra-node models have their own unique advantages in terms of performance, portability, and productivity (Diaz et al., 2012). Provided that there is no clear indication that one model would replace another as the unified standard, we believe that MPI+X will still be the mainstream programming paradigm for future HPC systems. Based on this belief, the paradigm will evolve to adapt to hardware and application changes. This is because future computing systems will feature increasingly more heterogeneity and hierarchy in computing units and memory systems. The programming model needs to be aware of such changes to fully unlock the hardware potential. On the other hand, this MPI+X paradigm must be extended to realize portability in both function and performance. This is particularly true when we take into account the large legacy code in scientific computing.

New programming paradigms will emerge and

develop with new computing devices. For example, recent advances in quantum device fabrication offer the hope of its utility (Chong et al., 2017), but a gap still exists between the reliability requirements of quantum algorithms and the hardware size. To bridge the gap, vendors (e.g., Microsoft, Intel, and Google) have invested a lot in quantum devices and quantum programming. In particular, Microsoft has released quantum programming language Q# and its supporting tool (https://docs.microsoft.com/en-us/quantum/quantum-qr-intro?view=qsharp-preview). Provided that classical computers and quantum computers are good at solving different problems, we need to investigate how to incorporate new programming models into the classical ones at the software level and how they can collaborate to meet future multi-scale computing needs.

Domain-specific frameworks and library tools will be the main ways to improve productivity. This can be achieved by hiding the architectural details and improving the level of abstraction. Then we can realize rapid development and deployment of applications, thus effectively improving productivity. With the continuous expansion of application types and scales, we expect that the conventional scientific computing and the new intelligent computing will further enrich the application layer. Techniques (such as machine learning) will be used to auto-tune various workloads during runtime (Zhang et al., 2018). To summarize, domain-oriented frameworks will be the bond between future complex applications and HPC systems, as well as the connection between domain experts and system programmers.

4 Suggestions for zettascale computing

We think that the roadmap for future HPC development can be summarized in three phases. From now to 2025, the system performance will increase to 2–3 Eflops. From 2025 to 2030, the performance will scale to 50–80 Eflops. Finally, until the end of 2035, we are likely to reach zetta-flops. Most HPC systems will still be built based on CMOS chips until 2025. From petascale to exascale, the novel heterogeneous architecture was the key factor in enabling performance boost. With the technology evolution and revolution, specialized computing devices based on the non-Von Neumann structure will be developed and will play an important role in enabling zettascale computing by 2035.

Table 1 lists the expected main technical metrics of a zettascale computing system. With a peak performance of 1 Zflops, the power consumption will reach the level of 100 MW, leading to a power efficiency of 10 Tflops/W. Note that we refer to the double-precision 64-bit floating-point operations (mainly used today) for a clear comparison with the present HPC systems. The performance per node will be at the level of 10 Pflops. The bandwidth between nodes will increase to 1.6 Tb/s, and the system I/O bandwidth will be around 10–100 PB/s. The storage capacity will be scaled to 1 ZB to hold an extremely large volume of data. The whole system will take around 1000 m^2.

Table 1 Zettascale metrics

Metric	Value
Peak performance	1 Zflops
Power consumption	100 MW
Power efficiency	10 Tflops/W
Peak performance per node	10 Pflops/node
Bandwidth between nodes	1.6 Tb/s
I/O bandwidth	10–100 PB/s
Storage capacity	1 ZB
Floor space	1000 m^2

To realize these metrics, micro-architectures will evolve to consist of more diverse and heterogeneous components. Many forms of specialized accelerators (including new computing paradigms like quantum computing) are likely to co-exist to boost high performance computing in a joint effort. Enabled by new interconnect materials such as photonic crystals, fully optical-interconnecting systems may come into use, leading to more scalable, high-speed, and low-cost interconnection. The storage system will be more hierarchical to increase data access bandwidth and to reduce latency. The 2.5D/3D stack memory and the NVM technology will be more mature. With the development of material science, the memristor may be put into practice to close the gap between storage and computing, and the traditional DRAM may end life. To reduce power consumption, cooling will be achieved at multiple levels, from the cabinet/board level to the chip level. The programming model and software stack will also evolve to suit the new hardware models. Except for the MPI+X programming model, new programming models for new computing paradigms and new computing devices

will be developed, with the balance of performance, portability, and productivity in mind. Conventional HPC applications and emerging intelligent computing applications will co-exist in the future, and both hardware and software layers need to adapt to this application workload evolution (Asch et al., 2018).

To effectively support the transition from exascale computing to zettascale computing, we should do as follows: (1) We should focus on developing collaborative design for heterogeneous computing architectures, develop customization technology for specific applications, and strengthen the exploration of intelligent computing technologies and next-generation enabling technologies; (2) We need to investigate new optoelectronic integration and key component design in advance, and accelerate research into integrated network architecture and common optimization technologies of HPC; (3) We need to keep pace with the development trend of the new non-volatile storage technology and explore the integration of computing and storage; (4) We should focus on studying high-density engineering technologies, low-energy-consumption targeted cooling technologies, and integrated energy consumption dynamic management; (5) We need to develop transparent programming and software framework support to hide the programming changes brought by hardware changes; (6) For future development of high-performance computing in China, we should focus on the development of a domestic processor architecture, build a sound and robust high-performance software environment, and ensure the healthy and sustainable development of the domestic high-performance computing industry.

References

Ábrahám E, Bekas C, Brandic I, et al., 2015. Preparing HPC applications for exascale: challenges and recommendations. 18th Int Conf on Network-Based Information Systems, p.401-406.

Asch M, Moore T, Badia R, et al., 2018. Big data and extreme-scale computing: pathways to convergence—toward a shaping strategy for a future software and data ecosystem for scientific inquiry. *Int J High Perform Comput Appl*, 32(4):435-479. https://doi.org/10.1177/1094342018778123

Cavin RK, Lugli P, Zhirnov VV, 2012. Science and engineering beyond Moore's law. *Proc IEEE*, 100:1720-1749. https://doi.org/10.1109/JPROC.2012.2190155

Chong FT, Franklin D, Martonosi M, 2017. Programming languages and compiler design for realistic quantum hardware. *Nature*, 549(7671):180-187. https://doi.org/10.1038/nature23459

Diaz J, Muñoz-Caro C, Niño A, 2012. A survey of parallel programming models and tools in the multi and many-core era. *IEEE Trans Parall Distrib Syst*, 23(8):1369-1386. https://doi.org/10.1109/TPDS.2011.308

Fang J, Varbanescu AL, Sips HJ, 2011. A comprehensive performance comparison of CUDA and OpenCL. Int Conf on Parallel Processing, p.216-225. https://doi.org/10.1109/ICPP.2011.45

Glosli JN, Richards DF, Caspersen KJ, et al., 2007. Extending stability beyond CPU millennium: a micron-scale atomistic simulation of Kelvin-Helmholtz instability. ACM/IEEE Conf on Supercomputing, p.1-11. https://doi.org/10.1145/1362622.1362700

Jacob P, Zia A, Erdogan O, et al., 2009. Mitigating memory wall effects in high-clock-rate and multicore CMOS 3D processor memory stacks. *Proc IEEE*, 97(1):108-122. https://doi.org/10.1109/JPROC.2008.2007472

Jeddeloh J, Keeth B, 2012. Hybrid memory cube new DRAM architecture increases density and performance. Int Symp on VLSI Technology, p.87-88. https://doi.org/10.1109/VLSIT.2012.6242474

Keeton K, 2015. The machine: an architecture for memory-centric computing. 5th Int Workshop on Runtime and Operating Systems for Supercomputers, p.1. https://doi.org/10.1145/2768405.2768406

Kim NS, Chen D, Xiong J, et al., 2017. Heterogeneous computing meets near-memory acceleration and high-level synthesis in the post-Moore era. *IEEE Micro*, 37(4):10-18. https://doi.org/10.1109/MM.2017.3211105

Kolli A, Rosen J, Diestelhorst S, et al., 2016. Delegated persist ordering. 49th Annual IEEE/ACM Int Symp on Microarchitecture, p.1-13. https://doi.org/10.1109/MICRO.2016.7783761

Lucas R, Ang J, Bergman K, et al., 2014. Top10 exascale research challenges. Department of Energy Office of Science. https://science.energy.gov/~/media/ascr/ascac/pdf/meetings/20140210/Top10reportFEB14.pdf

Mishra S, Chaudhary NK, Singh K, 2013. Overview of optical interconnect technology. *Int J Sci Eng Res*, 3(4):364-374.

Rumley S, Nikolova D, Hendry R, et al., 2015. Silicon photonics for exascale systems. *J Lightw Technol*, 33(3):547-562. https://doi.org/10.1109/JLT.2014.2363947

Schroeder B, Gibson GA, 2007. Understanding failures in petascale computers. *J Phys*, 78(1):012022. https://doi.org/10.1088/1742-6596/78/1/012022

Shen J, Fang J, Sips HJ, et al., 2013. An application-centric evaluation of OpenCL on multi-core CPUs. *Parall Comput*, 39(12):834-850. https://doi.org/10.1016/j.parco.2013.08.009

Vinaik B, Puri R, 2015. Oracle's Sonoma processor: advanced low-cost SPARC processor for enterprise workloads. IEEE Hot Chips 27 Symp, p.1-23. https://doi.org/10.1109/HOTCHIPS.2015.7477455

Waldrop MM, 2016. The chips are down for Moore's law. *Nature*, 530(7589):144-147. https://doi.org/10.1038/530144a

Wilkes MV, 1995. The memory wall and the CMOS endpoint. *SIGARCH Comput Archit News*, 23(4):4-6. https://doi.org/10.1145/218864.218865

Wulf WA, McKee SA, 1995. Hitting the memory wall: implications of the obvious. *SIGARCH Comput Archit News*, 23(1):20-24. https://doi.org/10.1145/216585.216588

Xu W, Lu Y, Li Q, et al., 2014. Hybrid hierarchy storage system in Milkyway-2 supercomputer. *Front Comput Sci*, 8(3):367-377. https://doi.org/10.1007/s11704-014-3499-6

Xu Z, Chi X, Xiao N, 2016. High-performance computing environment: a review of twenty years of experiments in China. *Nat Sci Rev*, 3(1):36-48. https://doi.org/10.1093/nsr/nww001

Zhang P, Fang JB, Tang T, et al., 2018. Auto-tuning streamed applications on Intel Xeon Phi. IEEE Int Parallel and Distributed Processing Symp, p.515-525. https://doi.org/10.1109/IPDPS.2018.00061