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:

Networking and communication challenges for post-exascale systems*

Dhabaleswar PANDA‡, Xiao-yi LU, Hari SUBRAMONI

Department of Computer Science and Engineering, The Ohio State University, Ohio 43210, USA

E-mail: panda@cse.ohio-state.edu; luxi@cse.ohio-state.edu; subramon@cse.ohio-state.edu

Received Oct. 9, 2018; Revision accepted Oct. 15, 2018; Crosschecked Oct. 15, 2018

Abstract: With the significant advancement in emerging processor, memory, and networking technologies, exascale systems will become available in the next few years (2020–2022). As the exascale systems begin to be deployed and used, there will be a continuous demand to run next-generation applications with finer granularity, finer time-steps, and increased data sizes. Based on historical trends, next-generation applications will require post-exascale systems during 2025–2035. In this study, we focus on the networking and communication challenges for post-exascale systems. Firstly, we present an envisioned architecture for post-exascale systems. Secondly, the challenges are summarized from different perspectives: heterogeneous networking technologies, high-performance communication and synchronization protocols, integrated support with accelerators and field-programmable gate arrays, fault-tolerance and quality-of-service support, energy-aware communication schemes and protocols, software-defined networking, and scalable communication protocols with heterogeneous memory and storage. Thirdly, we present the challenges in designing efficient programming model support for high-performance computing, big data, and deep learning on these systems. Finally, we emphasize the critical need for co-designing runtime with upper layers on these systems to achieve the maximum performance and scalability.

Key words: Networking; Communication; Synchronization; Post-exascale; Programming model; Big data; High-performance computing (HPC); Deep learning; Quality of service (QoS); Accelerator

https://doi.org/10.1631/FITEE.1800631 **CLC number:** TP311

1 Introduction

Modern high-end computing (HEC) systems are enabling scientists and engineers to tackle major challenges in their respective domains and make significant contributions to their fields. Examples of such domains include astrophysics, earthquake analysis, weather prediction, nanoscience modeling, multi-scale and multi-physics modeling, biological computations, and computational fluid dynamics. In addition to traditional scientific computing, there is a significant demand for HEC in the fields of big-data analytics, machine learning, and deep learning.

HEC systems have made steady progress during the last two decades, and in the last decade significant attention has been paid toward designing and using exascale systems (ASCAC, 2010). The community is entering the pre-exascale period in 2018. As of June 2018, systems with capabilities of 200 PFlops, such as Summit (ORNL, 2018), have been announced. Multiple countries (US, China, and Japan) and the European Union are planning to design and develop exascale systems during 2020–2022.

As pre-exascale and exascale systems begin to be deployed and used, there will be continuous demand to run the next-generation applications with finer granularity, finer time-steps, and increased data

‡ Corresponding author

* Project supported by the National Science Foundation of the USA (Nos. IIS-1447804 and CNS-1513120)

ORCID: Xiao-yi LU, http://orcid.org/0000-0001-7581-8905

sizes. Based on the historical trends, the next-generation applications will require post-exascale systems during 2025–2035.

2 Envisioned architectures for post-exascale systems

Based on the upcoming technological trends, we envision some of the post-exascale systems to have the architectures outlined in Fig. 1. The systems will have a large number of dense nodes interconnected with network speeds of terabits per second or higher (Energy Government, 2011). The dense nodes will be heterogeneous with a mix of central processing units (CPUs), accelerators, co-processors, and field-programmable gate arrays (FPGAs)/application-specific integrated circuits (ASICs). The CPUs could also be a mix of strong and wimpy cores. The nodes will also have a large amount of memories with a combination of different technologies, such as non-volatile random-access memory and three-dimensional (3D) stacked memory. The components within a node will be connected with either wired or wireless interconnects. The on-chip interconnect technology would have matured to incorporate wireless interconnection among the components within a node (Sarvestani et al., 2018). Similarly, photonic technologies would have matured to be used within a node or a rack. This can provide a large number of concurrent communications among different components (CPUs, accelerators, and memories) without contention. As each node will be dense, the nodes will need to be connected to the overall network speed of terabits per second with multiple adapters/ports. This will facilitate a good balance between inter- and intra-node transfers.

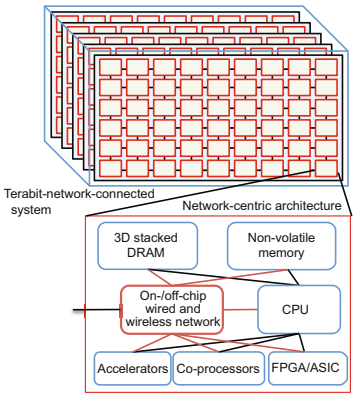

Fig. 1 Envisioned architectures for post-exascale systems

3 Heterogeneous networking technologies

As indicated in Section 2, we envision that post-exascale systems will use multiple networking technologies. These technologies can be used at different levels (intra-core, intra-node, intra-rack, and inter-rack) based on their power–performance trade-off. For example, a wireless interconnect might be optimal at the intra-core level. An optical photonic interconnect might be optimal at the intra-node and intra-rack levels. A standard electronic interconnect might be optimal at the inter-rack level. The wireless and photonic interconnects will introduce new features into the design of post-exascale systems. For example, both wireless and photonic technologies will allow the establishment of high-dimensional topologies within intra-core, intra-node, and intra-rack levels. These technologies will also facilitate one-to-many, many-to-one, and many-to-many communications in a flexible manner with a good performance. Thus, the co-existence of these heterogeneous technologies on a given system will provide much flexibility to the upper layers to design highly efficient communication and I/O stacks. As these networks will be operating at the multi-terabit speed, the communication mechanisms at different levels of the system must be coupled together to provide the best performance for end-to-end communication across any two processes in the system. For example, the wireless technology embedded inside the processor must be efficiently coupled with electronic/photonic technology across nodes/servers for efficient data transfer with negligible overhead. Otherwise, the benefits of these modern technologies will not be properly exploited. Therefore, the designers of post-exascale systems must consider these aforementioned issues and propose innovative solutions for them.

4 High-performance communication and synchronization protocols

As the high-performance networking technologies advance, capabilities are being increasingly offloaded to the network leading to an era of 'in-network computing'. Existing and emerging

schemes, such as core-direct and the Scalable Hierarchical Aggregation Protocol (SHArP) (Graham et al., 2016) from network vendors, are good examples of this trend. This trend is allowing different types of communication operations, such as point-to-point unicast, one-to-many multicast, many-to-one gather, and all-to-all communications to be efficiently implemented. Similarly, the wireless and photonic technologies described in Section 3 will provide efficient support for many-to-many communication without contention. As high-performance computing (HPC) systems aim toward becoming exascale and beyond, the previously mentioned techniques will become essential for achieving the best performance. Capabilities and middleware capable of using these schemes will be essential to propel scientific applications to a post-exascale performance. As communication networks become large, and emerging technologies, such as wireless and photonic, provide easy support for high-dimensional topologies, the necessity for communication and synchronization protocols to be 'network-topology-aware' will become very critical to minimize network contention. In this study, networks at each level of post-exascale systems must provide topology information with low overhead, which can be used by algorithms and communication protocols from upper-layers to provide the 'network-topology-aware' communication and minimize network contention.

5 Integrated support with accelerators and field-programmable gate arrays (FPGAs)

As post-exascale systems will use accelerators and FPGA-based components in an extensive manner, the communication mechanisms and protocols must also be re-designed to provide 'integrated' support for these devices in a heterogeneous environment. Such protocols must be designed to provide a high-performance communication with high productivity. For example, a wide array of scientific applications, such as MILC (Cui et al., 2007; Li et al., 2017), use message passing interface (MPI) datatypes to communicate non-contiguous data. However, state-of-the-art MPI libraries often simply 'pack' the data into contiguous buffers on the sender side, and the receiver 'unpacks' the data into the user buffer. This approach takes away CPU cycles from the application and provides a poor performance, especially for large messages, as it requires copying the data multiple times. To address the issue of unnecessary memory copies and the lack of overlapping opportunities in datatype processing, designs which offload the data packing/unpacking to FPGAs and smart adapters are essential. Solutions proposed by Mellanox BlueField (2017) from network vendors, such as Mellanox, are excellent examples of such emerging trends. Furthermore, the advent of new fields, such as deep learning, has introduced extreme and custom communication/computation requirements that must be addressed by runtimes. FPGAs, with the ability to satisfy the complex communications and computation requirements with high performance, are considered suitable for such requirements.

6 Fault-tolerance and quality-of-service (QoS) support

As post-exascale systems will be massive with millions of components, frequent failures will be common. Networks at each level must provide mechanisms for fault-tolerance and resiliency, and thus the upper levels of the software can be designed to make the software infrastructure fault-tolerant and resilient. Emerging applications, such as deep learning and streaming, are interactive in nature.

Furthermore, the interaction among multiple applications, which can potentially use the same interconnection network, must be considered. For instance, if a checkpoint operation is in progress at the same time as a dense communication, these two communication patterns are very likely to interfere with each other, resulting in a poor performance for both applications (Rajachandrasekar et al., 2012). In this study, providing QoS for different applications and different communication operations in the same application (Subramoni et al., 2010) will be important. It will also be critical for next-generation networking technologies to provide a quality-of-service (QoS) mechanism at the lowest level, and thus QoS-aware solutions for performance isolation across communication streams and jobs can be provided. As indicated in Section 3, post-exascale systems will have a set of heterogeneous networks. Although each network can

provide good QoS support, achieving an end-to-end QoS across heterogeneous networks will be challenging and will require innovative solutions.

7 Energy-aware communication schemes and protocols

Similar to exascale systems, the overall power consumption of post-exascale systems will also be crucial. In this study, the networking technologies must provide mechanisms for energy saving, such as automatic reduction in speed in the absence of data transfer and variable speeds with different energy trade-offs. Furthermore, the higher-level communication and synchronization protocols used by programming models, which move the data around post-exascale systems, and their runtime, must be designed with energy efficiency as one of the core objectives. This cannot be achieved through existing approaches, which consider runtime as a black-box, and use techniques such as CPU throttling and frequency-scaling based on high-level energy and performance measurements. A holistic approach encompassing a careful understanding of the power characteristics of the data movement phase, integrated design of the different communication protocols, and finally a co-design between the runtime and applications will be essential. Furthermore, it will be essential for the different software components on the HPC environment, such as scheduler, application, and runtime, to have complementary and adaptive energy-conserving strategies to achieve maximum savings while minimizing the impact on performance.

8 Software-defined networking (SDN)

During the past few years, the field of software-defined networking (SDN) has increasingly become more prominent. As modern HPC systems move toward supporting interactive workloads with data analytics and visualization, the requirements for persistent connectivity, such as those available in SDN with OpenFlow, are gaining increasing importance. The current-generation HPC interconnects, such as InfiniBand (Infiniband Trade Association, 1999. http://www. infinibandta.org), incorporate some SDN functionalities through their OpenSM subnet manager. However, per-flow resource management is not at a satisfactory level yet. Researchers have been exploring the incorporation of such resource management for InfiniBand. During the following decade, SDN technology will mature and will be a critical desired feature for post-exascale systems. This will allow post-exascale applications to take advantage of SDN features and functionalities. However, providing SDN functionalities across heterogeneous networks will be a challenge and will require innovative solutions.

9 Scalable communication protocols with the heterogeneous memory and storage

As post-exascale systems will use not only dynamic random-access memory (DRAM) but also other types of emerging memory technologies, such as non-volatile memory (NVM), the advanced features of NVM provide significant opportunities to design novel high-performance communication and high-throughput I/O subsystems for data-intensive applications. Technologies such as NVM express (NVMe)-over-fabrics (NVMf) (NVMe Express, 2016) have enabled offloading of various compute tasks associated with I/O operations to the network. The birth of the NVMe standard has changed the storage landscape. The lower latency and high scalability offered by this standard provide an unprecedented high performance. The emerging NVMf standard allows low-latency access to remote flash devices using NVMe commands, offering the possibility to fundamentally redesign storage systems. The next-generation communication and I/O runtime must be redesigned to consider NVM and NVMe technologies. We envision that the NVM- and NVMe-aware communication and I/O protocols with remote direct memory access (RDMA) can fundamentally change the landscape of the communication and I/O subsystem designs in post-exascale systems.

10 Programming model support for high-performance computing (HPC), big data, and deep learning

While MPI has been the common programming model for current-generation multi-petaflop systems for scientific computing applications, it has

inherent limitations for irregular applications. The new generation of partitioned global address space models (OpenSHMEM and UPC++) are attempting to alleviate the bottlenecks with a logical global address space. However, the programming models have inherent limitations for big data and deep learning. New generation of programming models, such as task-based support and efficient multi-threading schemes, must be designed to provide unified programming model support for HPC, big data, and deep learning on post-exascale systems. Owing to the requirements of supporting legacy applications, post-exascale systems may still choose to use existing programming models as the backbone, while significant and novel extensions may need to be proposed and added into current programming model standards, such as MPI and OpenMP. The associated challenges include: (1) How should the semantics of existing programming models be extended to enable HPC, deep learning, and big-data applications to extract the best possible performance? (2) Can these extended runtimes be used to co-design HPC, deep learning, and big-data applications to achieve the scale-up, scale-out, performance, and portability?

11 Hardware and software co-design for post-exascale systems and applications

To exploit the high-performance capabilities facilitated by the terabit-network-connected post-exascale systems, it will be vital to co-design the upper-layer HPC, big data, artificial intelligence (AI) applications, and the corresponding application frameworks with the underlying next-generation communication runtime designed for the post-exascale systems. As shown in Fig. 2, the design of the communication runtime and the co-design of several application-layer frameworks are not independent of each other, as proposed by Rahman et al. (2014), Lu et al. (2016), Shankar et al. (2016), and Biswas et al. (2018). The research and development process for the co-design on post-exascale systems will need to follow a spiral model. For instance, along with designing the advanced features of high-performance and energy-aware communication and synchronization protocols, heterogeneous memory and storage support, and programming model support, it will be critical to understand and identify the performance-governing patterns in the end applications, which will provide more insights into the bottlenecks specific to these applications as they are executed on next-generation post-exascale systems. For instance, tracing and profiling the data movement among different components, e.g., CPU, accelerators, heterogeneous memory, and FPGA/ASIC, will be used in designing the high-performance and energy-aware communication schemes described in Section 7. Furthermore, upon understanding predictable characteristics, the estimation of execution time for different computation and memory access patterns will form the basis for designing the data placement techniques and the data management schemes presented in Section 9.

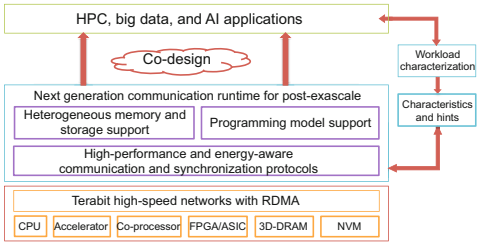

Fig. 2 Co-design challenges

In addition, from the aspect of emerging hardware trends, high-performance networks, such as InfiniBand and Omni-Path (Intel, 2016), have attempted to offload small subsets of computing tasks to the network, and the recent rapid scale-out of HEC systems has made the ability to offload more compute tasks to the network a fundamental need to achieve an exaFlOPS performance for end applications. Considering this target, vendors of high-performance networks have begun to incorporate various mechanisms to increasingly offload larger portions of computational tasks traditionally performed on the host to the network. An excellent example for this trend is the recent deployment of the SHARP technology (Graham et al., 2016) to offload reduction collectives on the Summit (1st on the TOP500 list) and Sierra (3rd on the TOP500 list). This computation offloading scheme has become popularly known as 'in-network computing'. We believe that such kind of in-network computing capabilities will still be available and will be even stronger than the current-generation schemes for future multi-terabit networks. These developments lead to the following broad challenges

for the hardware and software co-designs: (1) How can next-generation HPC, big data, and deep learning runtimes and frameworks be 'aware' of the computing capabilities of these emerging in-network computing technologies? (2) How can these systems be designed in the most optimized manner possible for post-exascale HPC, big data, and deep learning applications?

Finally, the co-design process will need to follow a positive feedback loop as shown in Fig. 2. The extracted application characteristics and demands must positively influence the communication runtime design. Furthermore, the runtime must deliver the best performance obtained from low-layer advanced hardware up to the end application layer. To make this co-design loop effective and efficient, there are many research and development challenges that must be solved from the collaborations among different communities, including architecture, HPC, big data, and AI.

12 Conclusions

In this paper, we have outlined a set of networking and communication challenges for post-exascale systems. These challenges include a range of diverse issues, such as heterogeneous networking technologies, communication protocols, integrated support with accelerators, fault-tolerance, QoS, energy-aware communication, and SDN functionalities. We have also discussed the challenges in designing scalable communication protocols with heterogeneous memory and storage. The impact of these challenges in designing next-generation programming model support for HPC, big data, and AI is emphasized. In addition, the critical need for co-designing many of these components is emphasized. Many of these challenges will require extensive research and development during the following decade. Innovative solutions arising from these research directions and close collaboration among the researchers in these areas will lead to successful design, deployment, and usage of post-exascale systems.

References

ASCAC Subcommittee on Exascale Computing, 2010. The Opportunities and Challenges of Exascale Computing. https://science.energy.gov/media/ascr/ascac/pdf/reports/Exascale_subcommittee_report.pdf

Biswas R, Lu XY, Panda DK, 2018. Accelerating tensorFlow with adaptive RDMA-based gRPC. 25th IEEE Int Conf on High Performance Computing, Data, and Analytic.

Cui YF, Moore R, Olsen K, et al., 2007. Enabling very-large scale earthquake simulations on parallel machines. In: Shi Y, van Albada GD, Dongarra J, et al. (Eds.), Computational Science. Springer Berlin Heidelberg, p.46-53.

Energy Government, 2011. Workshop on Terabits Networks for Extreme Scale Science. https://science.energy.gov/ /media/ascr/pdf/program-documents/docs/Terabit_networks_workshop_report.pdf

Graham RL, Bureddy D, Lui P, et al., 2016. Scalable Hierarchical Aggregation Protocol (SHArP): a hardware architecture for efficient data reduction. Proc 1st Workshop on Optimization of Communication in HPC, p.1-10. https://doi.org/10.1109/COMHPC.2016.006

Intel, 2016. Intel Omni-Path Architecture Driving Exascale Computing and HPC. https://www.intel.com/content/www/us/en/high-performance-computing-fabrics/omni-path-driving-exascale-computing.html

Li RZ, DeTar C, Gottlieb S, et al., 2017. MILC code performance on high end CPU and GPU supercomputer clusters. http://cn.arxiv.org/abs/1712.00143

Lu XY, Shankar D, Gugnani S, et al., 2016. High-performance design of Apache Spark with RDMA and its benefits on various workloads. Proc IEEE Int Conference on Big Data, p.253-262. https://doi.org/10.1109/BigData.2016.7840611

Mellanox BlueField, 2017. Multicore System on Chip. http://www.mellanox.com/related-docs/npu-multicore-processors/PB_Bluefield_SoC.pdf

NVMe Express, 2016. NVMe over Fabrics. http://www.nvmexpress.org/wp-content/uploads/NVMe_Over_Fabrics.pdf

ORNL, 2018. Summit: America's Newest and Smartest Supercomputer. https://www.olcf.ornl.gov/summit/

Rahman MWU, Lu XY, Islam NS, et al., 2014. HOMR: a hybrid approach to exploit maximum overlapping in MapReduce over high performance interconnects. Proc 28th ACM Int Conf on Supercomputing, p.33-42. https://doi.org/10.1145/2597652.2597684

Rajachandrasekar R, Jaswani J, Subramoni H, et al., 2012. Minimizing network contention in InfiniBand clusters with a QoS-aware data-staging framework. IEEE Int Conf on Cluster Computing, p.329-336. https://doi.org/10.1109/CLUSTER.2012.90

Sarvestani AMK, Bailey C, Austin J, 2018. Performance analysis of a 3D wireless massively parallel computer. *J Sens Actuat Netw*, 7(2):18. https://doi.org/10.3390/jsan7020018

Shankar D, Lu X, Islam N, et al., 2016. High-performance hybrid key-value store on modern clusters with RDMA interconnects and SSDs: non-blocking extensions, designs, and benefits. IEEE Int Parallel and Distributed Processing Symp, p.393-402. https://doi.org/10.1109/IPDPS.2016.112

Subramoni H, Lai P, Sur S, et al., 2010. Improving application performance and predictability using multiple virtual lanes in modern multi-core InfiniBand clusters. Int Conf on Parallel Processing.