The number of compute nodes in supercomputer systems grows steadily as we transition from petascale to exascale. Today’s petascale systems range from several thousands to tens of thousands of processing nodes. Exascale systems are expected to push the limit even further, with several hundreds of thousands of nodes and millions of computing cores. This extremely large component count poses daunting challenges to interconnection network designers.

The high-performance interconnection network design is a complex multidimensional space. The first obvious requirement is performance. We expect these networks to deliver submicrosecond end-to-end latencies, hundreds of Gigabytes per second of local link bandwidth combined with high global bisection bandwidth. But the designer must achieve these performance targets under demanding physical, architectural, technological, and economical constraints. The network routers have tight power budgets, and can be either integrated within the processing nodes to provide better communication performance or can be embedded in stand-alone routing switches. The network topology must obey strict wiring constraints when a large-scale, multi-rack supercomputer is deployed in a machine room. The network must be able to recover from different types of faults, using an arsenal of techniques from simple link retransmission to dynamic resource sparing and reallocation. The network interface also plays a central role and must be properly designed to match the network performance and functionality.

Although it’s crucial to carefully study the leading network designs and experiences currently driving petascale architectures, it’s also becoming more evident that no single country can drive such efforts in isolation. Hot Interconnects (HOTI) is an international conference that provides a unique forum for researchers and developers to discuss issues and experiences with the design and operation of large-scale interconnection networks.

This special issue presents improved and peer-reviewed versions of the best presentations of the 2011 edition of HOTI, focusing on large-scale network design. This extends the presentations of Hot Interconnects 2010 that include the IBM PERCS network.1
and the Cray Gemini network. This year we include contributions from Japan, China, and the United States, representing the leading network architectures.

“Tianhe-1A Interconnect and Message-Passing Services” by Min Xie et al. discusses the hardware and software design of the Chinese Tianhe-1A supercomputer, currently the second-fastest supercomputer in the Top 500 list (November 2011). The system consists of 7,168 compute nodes, each with two six-core Intel Xeon processors and an additional Nvidia GPU, thus posing very high demands on the communication bandwidth. The custom-designed network employs a fat-tree topology. The authors describe the network routing chip (NRC), network interface chip (NIC), and network topology. The article especially focuses on supporting the message passing interface (MPI) at large scale, including hardware support for offloading collective operations.

In “The Tofu Interconnect,” Yuichiro Ajima et al. contrast several interesting design tradeoffs faced during the design of Fujitsu’s Torus Fusion (Tofu) Interconnect. The network employs an unusual six-dimensional torus topology and has additional support for high-speed barrier synchronization and for other reduction operations. The interconnect has been demonstrated to work with more than 80,000 nodes. This system achieves almost 10 petaflops per second with an impressive 93 percent peak flop rate on High Performance Linpack (HPL) using more than 700,000 cores, rightfully claiming the number one position on the current Top 500 list. The article describes in detail the Tofu network interface (TNI) and network router (TNR), both integrated on the same chip, and gives initial communication benchmark results.

“The IBM Blue Gene/Q Interconnection Fabric” by Dong Chen et al. describes the design of the Blue Gene/Q interconnection network. This architecture is poised to scale to a peak performance of 20 petaflops per second and beyond. In addition to a very high message rate, an essential metric for exascale networks, the network also supports collective offload for application scalability. The network interface and router are integrated into the CPU chip and consume only 8 percent of the die area. Debugability and validation of the network primitives are essential during the network design, and this article provides insight on some of the techniques used to achieve both performance and reliability at scale.

We anticipate that many of the principles and techniques presented in this special issue will play an important role in the design of future large-scale network architectures.

References

Torsten Hoefler leads the modeling and simulation efforts of parallel petascale applications for the NSF-funded Blue Waters project at the University of Illinois at Urbana-Champaign. He also represents the University of Illinois in the Message Passing Interface (MPI) Forum, where he chairs the Collective Operations and Topologies working group. His research interests include performance-centric software development, specifically scalable networks, parallel programming techniques, and performance modeling. Hoefler has a PhD in computer science from Indiana University. He is a member of IEEE, the ACM, and the ACM Special Interest Group on High Performance Computing.

Patrick Geoffray is a senior software architect at Myricom. His research interests include high-speed interconnects, network interfaces, and communication layers. Geoffray has a PhD in computer science from the University of Lyon, France.

Fabrizio Petrini is a senior researcher in the Multicore Solution Department of the IBM T.J. Watson Research Center. His research
interests include multicore processors and supercomputers, including high-performance interconnection networks and network interfaces, fault tolerance, and job scheduling algorithms. Petrini has a PhD in computer science from the University of Pisa. He is a member of the IEEE Computer Society.

Jesper Larsson Träff is a full professor of parallel computing at the Vienna University of Technology (TU Wien). His research interests include interfaces, algorithms, and architectures for parallel computing. Larsson Träff has a PhD and a DSc in computer science from the University of Copenhagen. He is a member of the European Association for Theoretical Computer Science (EATCS) and the ACM.
Focus on Your Job Search

IEEE Computer Society Jobs helps you easily find a new job in IT, software development, computer engineering, research, programming, architecture, cloud computing, consulting, databases, and many other computer-related areas.

New feature: Find jobs recommending or requiring the IEEE CS CSDA or CSDP certifications!

Visit www.computer.org/jobs to search technical job openings, plus internships, from employers worldwide.

http://www.computer.org/jobs

The IEEE Computer Society is a partner in the AIP Career Network, a collection of online job sites for scientists, engineers, and computing professionals. Other partners include Physics Today, the American Association of Physicists in Medicine (AAPM), American Association of Physics Teachers (AAPT), American Physical Society (APS), AVS Science and Technology, and the Society of Physics Students (SPS) and Sigma Pi Sigma.