



Guest Editors' Introduction

HIGH-PERFORMANCE INTERCONNECTS

..... We are pleased to introduce this special issue of *IEEE Micro*, featuring articles that capture the latest results on high-performance interconnection networks, including some of the best presentations from last summer's Hot Interconnects 13 at Stanford University.

Hot Interconnects provides a unique international forum for researchers and developers of state-of-the-art hardware and software architectures for interconnection networks at all scales, ranging from on-chip processor-memory interconnects to high-performance networks for the largest and most powerful supercomputers.

Interconnection networks are increasingly becoming the backbone of microprocessors, data centers, and supercomputers.

This special issue explores five important topics in the design space of high-performance networks: on-chip networks for multicore processors, with an accurate description of the Cell Broadband Engine (BE) internal network; the architectural evolution of Gigabit Ethernet in "Bridging the Ethernet-Ethernot Performance Gap"; Cray's SeaStar, a highly scalable network for supercomputers; a door to the future of optical interconnection networks, with some latest results from the Osmosis project; and a content-aware switch for Internet computing based on the Intel IXP 2400 in "A Network Processor-Based, Content-Aware Switch."

Perhaps the most revealing technological

trend is that the design issues (and the proposed solutions) are becoming very similar across all these domains, with many aspects in common.

For example, the on-chip network of the Cell processor and its communication mechanisms, such as remote direct memory access (RDMA) communication, are strikingly similar to those that cluster and supercomputer networks have adopted. Along the same line, the latest generation of Gigabit Ethernet network adapters is incorporating architectural solutions, such as protocol offload engines in the network interface, that are very popular in cluster computing.

The opening article, "Cell Multiprocessor Communication Network: Built for Speed," explores the Cell BE processor's on-chip network. A major trend in computer architecture is the implementation of highly integrated chips. This trend is driving the development of processors that can perform functions typically associated with entire systems. Rather than build monolithic processors, which are prohibitively expensive to develop and have high power consumption and limited return on technological investment, it is becoming more effective to build modular processors with multiple cores. These multicore system-on-a-chip processors, such as the Cell BE, can integrate several identical independent processing units on the same die, together with network interfaces, acceleration units, and

Fabrizio Petrini
Pacific Northwest National
Laboratory

Olav Lysne
Simula Research
Laboratory

Ron Brightwell
Sandia National
Laboratories

other specialized units. It is easy to predict that the multicore paradigm will become ubiquitous in the near future. Processors with potentially many cores—tens with the current state of the art, but possibly hundreds in the near future—will be the heart of gaming devices, desktops, supercomputers, and embedded devices.

In all multicore processors, a key technological challenge is the design of the internal, on-chip communication network. To bring to fruition the unprecedented computational power of the many available processing units, the network not only must provide very high performance—both in terms of latency and bandwidth—but it must also be able to resolve contention under heavy load, provide fairness, and hide as much as possible the physical distribution of the processing units.

The Cell BE uses a simple but effective strategy for interconnecting eight synergistic processing units with a control processor and memory and I/O controllers. These units can communicate using four rings that can deliver an impressive aggregate performance.

In the past decade, many high-performance networks have been introduced in the cluster-computing world. These networks provide many interesting features that enhance performance and scalability, such as native support for collective communication, extremely low-latency and high bandwidth, and feature-rich interfaces that allow direct user-access to smart communication-offload engines. These networks are typically adopted by some of the largest supercomputers, in particular those that require a high degree of communication coupling between thousands of processing nodes.

Ethernet is the de facto standard for local- and wide-area networks, a commodity technology that dominates the lower part of the Top 500 supercomputer list, but still it is not considered adequate for high-end computing. “Bridging the *Ethernet*-*Ethern*ot Performance Gap” provides a fresh look to the state of the art in the *Ethernet* and *Ethern*ot (anything else other than *Ethernet*) worlds.

The gap between these two worlds is narrowing. Architectural solutions that once were exclusive features of proprietary networks, such as protocol offload engines, are now an integral part of the latest 10-Gigabit

network adapters.

Recently, there has been a significant push toward the convergence of these two classes of technologies. To resist the wave of the commodity market, proprietary high-performance networks are progressively embracing the *Ethernet* protocol. Myricom (the vendor of the Myrinet network) recently introduced adapters that merge proprietary technology with the *Ethernet* wire-protocol. Similarly, Quadrics will release in the next few months a programmable network adapter that can support, in two distinct incarnations, its traditional proprietary protocol and 10-Gigabit *Ethernet*.

The authors of this article provide an up-to-date analysis of the state of the art and compare the communication protocols with a rich set of benchmarks and applications.

In “SeaStar Interconnect: Balanced Bandwidth for Scalable Performance,” the authors provide an overview of a new integrated network interface and router ASIC from Cray Inc. that provides the high-performance interconnect for Sandia National Laboratories’ Red Storm massively parallel processing machine. The SeaStar is a significant step forward for increasing the network bandwidth capability of a tightly coupled, distributed-memory platform, as it provides a peak bidirectional bandwidth capability of more than two gigabytes-per-second. The SeaStar is also one of the first high-performance networks to leverage AMD’s HyperTransport interface for directly attaching a network adapter to the memory bus.

Following the trend of other high-performance networks, like Infiniband, Myrinet, Quadrics and 10-Gigabit *Ethernet*, the SeaStar contains a programmable processor on the network interface that supports offloading of a significant portion of the network protocol stack. The authors describe their approach to utilizing this processor to handle typical functions associated with network protocol processing, as well as more advanced features like application-level message selection and additional functionality driven by managing some of the limited resources on the network interface.

The authors use traditional microbenchmark results to characterize the impact of offloading processing to the network interface

and also to analyze the overhead of network resource management strategies.

The next article in this issue describes several innovative approaches to meeting the demanding requirements of high-performance computing applications when designing and implementing an all-optical, data-path switch. In "Designing a Crossbar Scheduler for HPC Applications," the authors provide a detailed description of their approach to meeting the performance and reliability requirements for the Osmosis optical switch. They provide an overview of their prototype optical-switch environment and describe several new approaches to achieving the network latency, bandwidth, and reliability demands of a large-scale high-performance computing platform.

The article details a number of novel strategies, including a speculative transmission technique that significantly reduces average control path latency and a parallel-pipelined arbitration scheme that ultimately leads to increased throughput. While the techniques described in the article are driven by very stringent HPC requirements and the need to meet them in an all-optical environment, a number of the approaches are not specific to Osmosis and can be beneficial to crossbar schedulers in general.

Finally, many highly loaded Web sites use a server cluster to accommodate the incoming requests. To hide the distributed nature of the processing from the clients, most architects will put a switch in front of the cluster that forwards packets based on Layer-5 information. In "A Network Processor-Based Content-Aware Switch," we are presented with a solution for content-aware switching that is fully based on a network processor (NP).

Compared to a Linux-based switch, this solution eliminates the copying of data across the PCI-bus, and it opens for more efficient packet processing, since the NP's are typically optimized for packet processing. The authors present the current state of an ongoing project in implementing a content-aware switch on the Intel IXP 2400 NP. The article contains a detailed description of design choices. And, although the context of the article is application specific, it sheds light on a wide set of issues concerning the use of network processors.

MICRO

Fabrizio Petrini is a laboratory fellow in the Applied Computer Science Group of the Computational Sciences and Mathematics Division at Pacific Northwest National Laboratory. His research interests include various aspects of supercomputers, such as high-performance interconnection networks and network interfaces, multicore processors, job-scheduling algorithms, parallel architectures, operating systems, and parallel-programming languages. Petrini has a Laurea and a PhD in computer science from the University of Pisa, Italy. He is a member of the IEEE Computer Society.

Olav Lysne is a research director at Simula Research Laboratory and a professor in computer science at the University of Oslo. His early research was in the field of algebraic specification and term rewriting. His recent research interests include interconnects, focusing on effective routing, fault tolerance, and quality of service. In this field, he has served on program committees of several of the most renowned conferences, participated in a series of European projects, and published around 70 academic papers. Lysne received an MS and a DSc in computer science from the University of Oslo. He is a member of the IEEE.

Ron Brightwell is a principal member of the technical staff at Sandia National Laboratories. His research interests include high-performance scalable communication interfaces and protocols for system-area networks, and operating systems for massively parallel processing. Brightwell has a BS in mathematics and an MS in computer science from Mississippi State University. He is a member of the IEEE Computer Society and the ACM.

For further information on this or any other computing topic, visit our Digital Library at <http://www.computer.org/publications/dlib>.