# Design of a High-Speed Optical Interconnect for Scalable Shared-Memory Multiprocessors

THE ARCHITECTURE PROPOSED HERE REDUCES REMOTE MEMORY ACCESS LATENCY BY INCREASING CONNECTIVITY AND MAXIMIZING CHANNEL AVAILABILITY FOR REMOTE COMMUNICATION. IT ALSO PROVIDES EFFICIENT AND FAST UNICAST, MULTICAST, AND BROADCAST CAPABILITIES, USING A COMBINATION OF AGGRESSIVELY DESIGNED MULTIPLEXING TECHNIQUES. SIMULATIONS SHOW THAT THIS ARCHITECTURE PROVIDES EXCELLENT INTERCONNECT SUPPORT FOR A HIGHLY SCALABLE, HIGH-BANDWIDTH, LOW-LATENCY NETWORK.

Avinash Karanth Kodi
Ahmed Louri
University of Arizona

Large-scale distributed shared-memory multiprocessors (DSMs) provide a shared address space by physically distributing the memory among different processors.[1] A fundamental DSM communication problem that significantly affects scalability is an increase in remote memory latency as the number of system nodes increases. Remote memory latency, caused by accessing a memory location in a processor other than the one originating the request, includes both communication latency and remote memory access latency over I/O and memory buses. This long latency can degrade overall system performance by as much as 50 percent.[1]

Although DSM systems frequently use latency-tolerating or latency-hiding techniques to reduce remote latency, these techniques require extra bandwidth and greatly increase memory traffic by fetching more data than needed.[1] Moreover, as Figure 1a shows, the increasing performance gap between processor and off-chip clock rates further deteriorates DSM system performance. As seen in Figure 1a, the CPU bandwidth is computed by multiplying the expected increase in clock rates with the speed of L2 cache data access rate.

The projected CPU, serial off-chip, memory and I/O speeds were obtained from the "International Technology Roadmap for Semiconductors," 2003 ed.; http://public.itrs.net/Files/2003ITRS/Home2003.htm.

0272-1732/05/$20.00 © 2005 IEEE Published by the IEEE computer Society

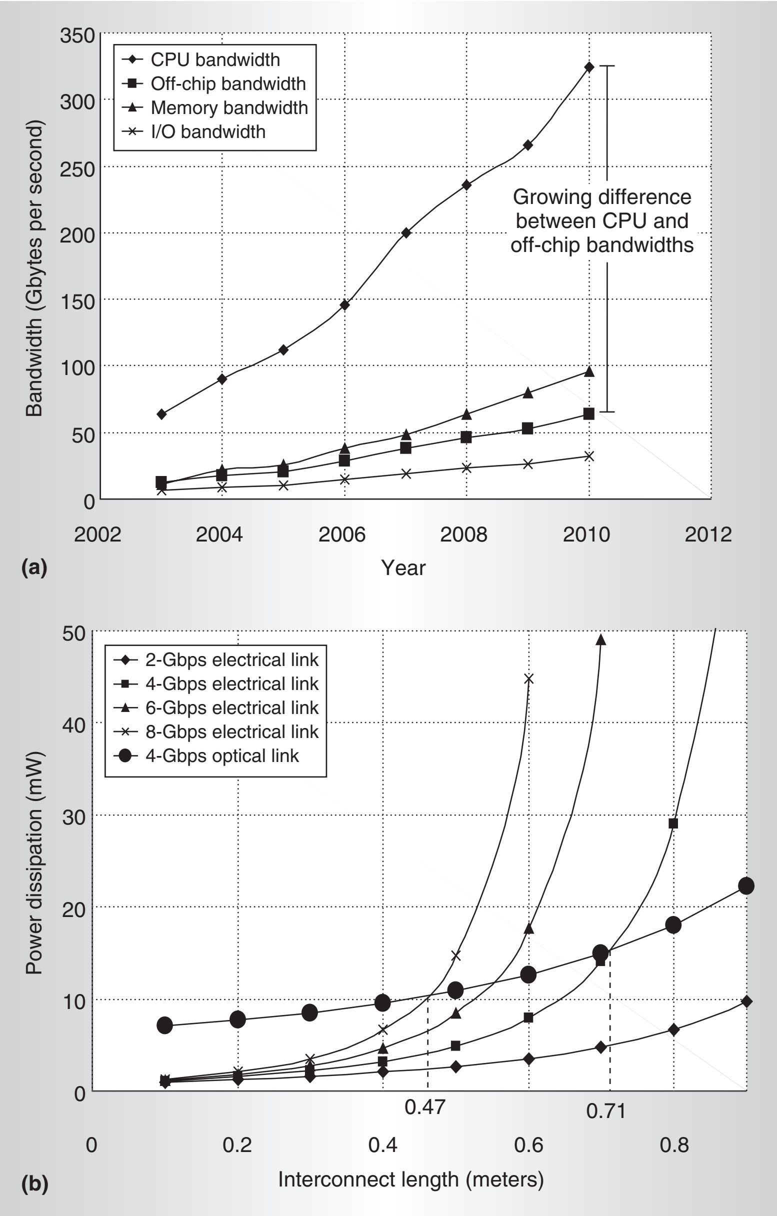

Figure 1. Predicted bandwidth comparison based on ITRS 2003 (a); power dissipation of electrical and optical links for various interconnect lengths (b).

Difficulties with electrical signaling at multi-GHz rates over electrical interconnects limit the performance of conventional peripheral-component-interface (PCI)- and PCI-X-based electrical solutions.[2] Although newer serial, point-to-point I/O technologies such as HyperTransport[3] and PCI-Express[4] achieve scalable bandwidth, further attempts to increase off-chip bandwidths by using equalization at higher data rates reduce noise tolerance and increase power dissipation.[5,6]

Figure 1b shows the power dissipated at various interconnect lengths, using simultaneous bidirectional low-swing current mode with a bipolar differential signaling scheme on a high-performance Getek board at 2, 4, 6, and 8 Gbps. Electrical interconnect power rises with length and bit rates because it becomes more attenuated and suffers a greater impact from unattenuated as well as fixed noise sources, thereby limiting DSM systems from reaching their full potential.

Smaller DSM systems ranging from 4 to 8 nodes usually interconnect via a single switch. An enlarged system requires a hierarchy of switches, which causes a significant routing or switching delay in the additional switching stages, which in turn increases remote latency.[7] Huang et al. reported that scaling from a medium- to a large-scale multiprocessor increases memory access latency by 60 percent.[8] Collective operations such as multicast and broadcast algorithms require synchronization among various processors using electrical interconnects and can be very time consuming as system size increases.[1] Future high-performance DSM systems will utilize commercial off-the-shelf processors that require aggregate computational and communication bandwidths on the order of 4 to 40 terabits per second.[2] Thus, lack of sufficient bandwidth (both memory and communication) will be the fundamental obstacle to future scalable DSM systems.

A technology that provides higher bandwidths, less cross talk, no electromagnetic interference, and lower latencies, with lower power requirements than current electronics-based interconnect, is optical interconnect.[9,10] Figure 1b shows the power dissipation of an optical link using a vertical-cavity surface-emitting laser (VCSEL) and pin photodetector at 4 Gbps. This figure presents an analysis similar to that of Cho, Kapur, and Saraswat,[5] but we extended it to include the optical link for the VCSEL model. Optical interconnects are superior because they have lower attenuation and noise levels, but they need extra power for conversion from electronics to optics. Because this is a fixed penalty, optical interconnects become beneficial at longer lengths. As Figure 1b shows, the critical length at which optics become beneficial decreases from 0.71 meters at 4 Gbps to 0.47 meters at 8 Gbps. Moreover, significant developments

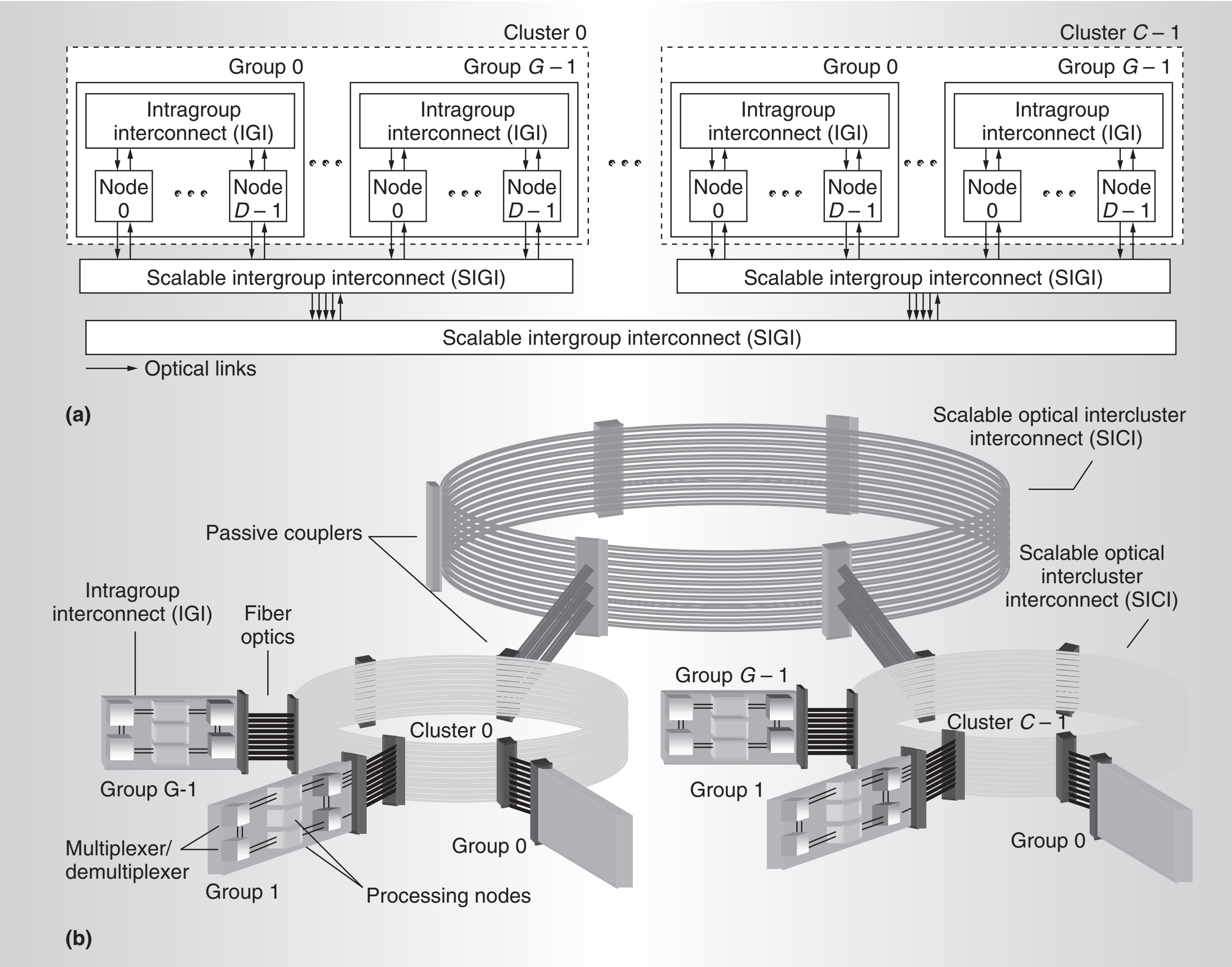

Figure 2. Architectural overview of Rapid (a); conceptual diagram of Rapid network (b).

in optical and optoelectronic devices and packaging technologies have made optical interconnects a viable and cost-effective option for building scalable networks.[2,11]

This article proposes an integrated solution that reduces remote memory access latency in DSMs and still lets us scale the network significantly using low-latency, high-bandwidth optical technology for both board-to-board and backplane communication. The interconnect technology combines optical components and a novel architecture called Rapid (reconfigurable and scalable all-photonic interconnect for distributed shared memory). Rapid significantly reduces the critical remote memory latency in high-performance DSMs by

- increasing connectivity, maximizing channel availability, and providing scalable bandwidth, using a decentralized wavelength allocation scheme along with wavelength-division-multiplexing (WDM), time-division-multiplexing (TDM), and space-division-multiplexing (SDM) techniques;
- implementing an efficient multicast and broadcast functionality, which helps reduce the part of memory latency associated with implementing synchronization operations; and
- using a switchless topology, based on passive optical-interconnect technology, which reduces cost and significantly improves performance.

## Overview

We define a Rapid network as a 3-tuple ($C$, $G$, $D$), where $C$ is the total number of clusters,

$G$ is the total number of groups per cluster, and $D$ is the total number of nodes per group. Each node is identified as R($c$, $g$, $d$), where $0 \leq d \leq D - 1$; $0 \leq g \leq G - 1$; $0 \leq c \leq C - 1$. (R is from R($c$, $g$, $d$) indicating the node with cluster, group and node numbers. Upper case indicates the total number of clusters, groups and nodes. Lower case is used for indicating the numbers.) The total number of nodes in Rapid is the multiplicative factor $N = C \times D \times G$.

Figure 2 shows the Rapid architecture. In Figure 2a, 0 to $D - 1$ nodes connect to form a group, and 0 to $G - 1$ groups connect to form a single cluster. All nodes connect via passive couplers to two subnetworks: a scalable intragroup interconnect (IGI) and a scalable intergroup interconnect (SIGI). The SIGI further connects to the scalable intercluster interconnect (SICI) to increase the architecture's scalability. We separate intragroup (local) and intergroup/intercluster (remote) communications to provide a more efficient implementation for both types. Every node in Rapid has two sets of tunable transmitters and fixed receivers for intra- and intergroup communication.

Figure 2b shows a conceptual diagram of the Rapid network. We used optical waveguides for interconnects on the board and optical fiber with multiplexers and demultiplexers for interconnects from the board to the SIGI. As Figure 2b shows, there are three ways to scale Rapid: adding more nodes, adding more groups, or replicating the existing network to form a new cluster. Although the figure shows the architecture as a hierarchical ring system, we drew it this way only for clarity. Rapid actually has a point-to-point topology, as explained later.

### Wavelength allocation and routing

The number of wavelengths employed for intragroup communication in an R($c$, $g$, $d$) system equals the maximum number of nodes ($D$) in each group of the system; that is, we assign every node a wavelength on which it can receive signals. Therefore, we can perform distinct wavelength allocation in different groups by assigning every node a unique wavelength on which it can receive optical packets from other intragroup nodes.

Figure 3 shows the remote wavelength assignment scheme in an R(1, 3, 4) system. For remote communication, different wavelengths from various groups are selectively merged into separate channels and provide high connectivity. Remote wavelengths are denoted $\lambda_i^{(j,k)}$, where $i$ is the wavelength, $j$ is the group number, and $k$ is the number of the cluster from which the wavelength originates. For example, consider group 2's transmitter. All nodes R(0, 2, $d$) have tunable transmitters, so the nodes can transmit on any wavelength $\lambda_i^{(2,0)}$, $i = 0, 1, 2, 3$. Any node in group 2 can communicate with itself on $\lambda_0^{(2,0)}$, with group 0 on $\lambda_1^{(2,0)}$, and with group 1 on $\lambda_2^{(2,0)}$. The cluster that group 2 can communicate with is cluster 1 on $\lambda_3^{(2,0)}$. The fiber channel that transmits $\lambda_0$ is called the home channel for that particular group (shown as a dashed line for group 2). All signals originating from a particular group are demultiplexed and then selectively multiplexed with different home group channels. For group 2, the multiplexed signal on the home channel, $(\lambda_0^{(2,0)} + \lambda_1^{(0,1)} + \lambda_2^{(0,0)} + \lambda_3^{(1,0)})$, is demultiplexed at the group 2 receiver. Because the receivers are fixed, $\lambda_i$ is received by node R(0, 2, $I - 1$). For remote traffic, the number of wavelengths necessary to obtain the appropriate connectivity is $G$; that is, $(G - 1)$ wavelengths are required for communication with every other group, and one more wavelength $\lambda_0$ for multicast communication. We have described the multicast/broadcast implementation elsewhere.[12]

Rapid uses the time division multiple-access (TDMA) protocol with preallocation to prevent the collision of requests.[12] Remote intergroup communication takes place when the source and destination nodes are in different groups. For R(1, $g$, $d$), a single optoelectronic conversion implements complete connectivity for a network of any size; thus, the diameter of R(1, $g$, $d$) is 2. This is possible because the wavelength assignment algorithm designed for remote group communication permits high connectivity. For intercluster communication, the multiplexed signals from different clusters are demultiplexed at group-cluster interface 0, as Figure 3 shows. The demultiplexed signal then merges with different group home channels. The wavelengths originating from different groups are then selectively demultiplexed to other clusters, thus providing high connectivity. The R($c$, $g$, $d$) configuration's maximum diameter is 4. This configuration trades

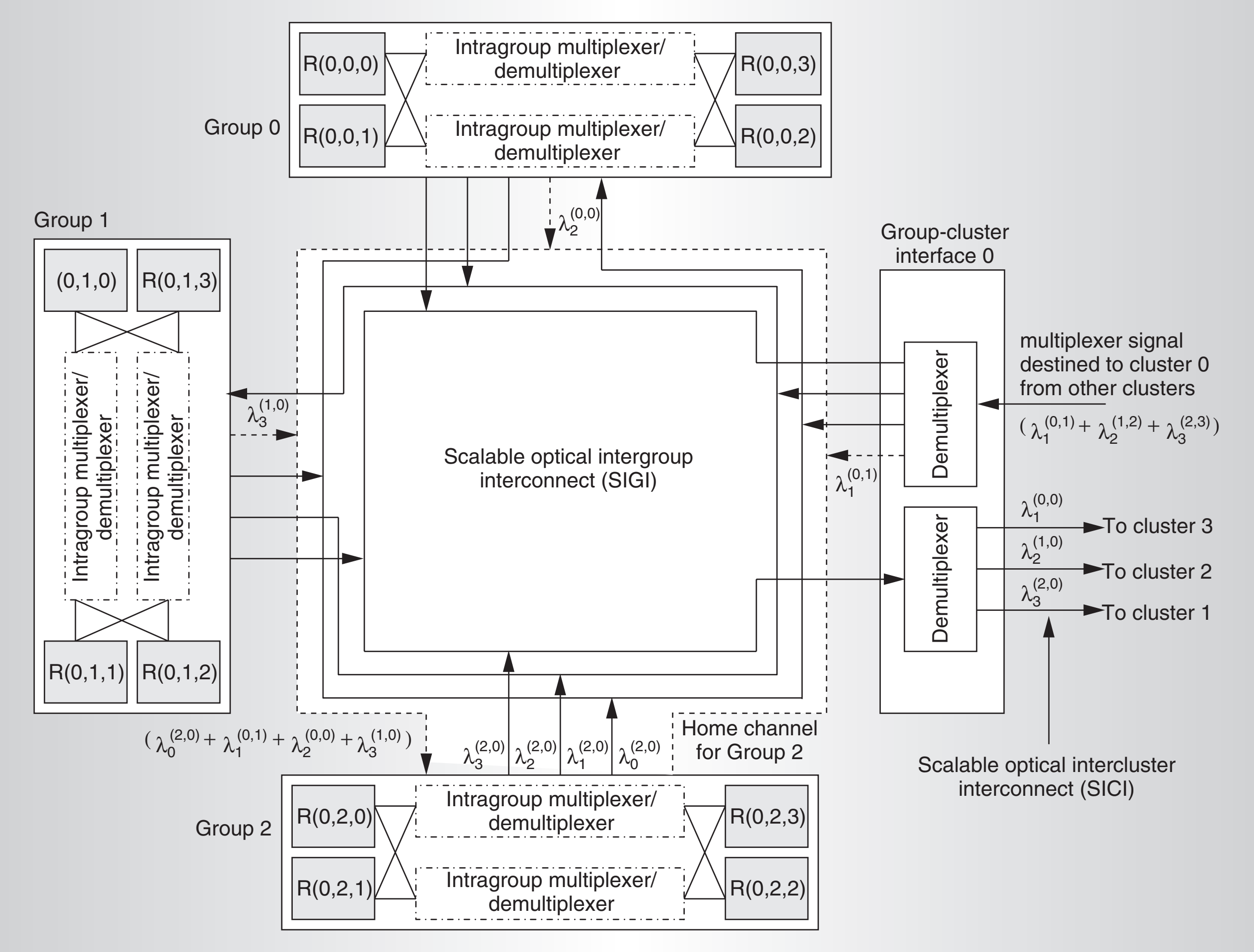
Group 0
R(0,0,0)
Intragroup multiplexer/ demultiplexer
R(0,0,3)
R(0,0,1)
Intragroup multiplexer/ demultiplexer
R(0,0,2)
$\lambda_2^{(0,0)}$
Group 1
(0,1,0)
R(0,1,3)
Intragroup multiplexer/ demultiplexer
Intragroup multiplexer/ demultiplexer
R(0,1,1)
R(0,1,2)
$\lambda_3^{(1,0)}$
Scalable optical intergroup interconnect (SIGI)
Group-cluster interface 0
Demultiplexer
Demultiplexer
multiplexer signal destined to cluster 0 from other clusters
$(\lambda_1^{(0,1)} + \lambda_2^{(1,2)} + \lambda_3^{(2,3)})$
$\lambda_1^{(0,1)}$
$\lambda_1^{(0,0)}$
To cluster 3
$\lambda_2^{(1,0)}$
To cluster 2
$\lambda_3^{(2,0)}$
To cluster 1
$(\lambda_0^{(2,0)} + \lambda_1^{(0,1)} + \lambda_2^{(0,0)} + \lambda_3^{(1,0)})$
$\lambda_3^{(2,0)}$ $\lambda_2^{(2,0)}$ $\lambda_1^{(2,0)}$ $\lambda_0^{(2,0)}$
Home channel for Group 2
Scalable optical intercluster interconnect (SICI)
Group 2
R(0,2,0)
Intragroup multiplexer/ demultiplexer
R(0,2,3)
R(0,2,1)
Intragroup multiplexer/ demultiplexer
R(0,2,2)

Figure 3. Functional diagram of a Rapid R(*c*, *g*, *d*) network in which *d* = 4 (nodes), *g* = 3 (groups), and *c* = 1 (cluster).

off wavelength usage for latency in smaller systems. For example, by using four wavelengths and passive optical components, R(*c*, *g*, *d*) can accommodate 64 nodes, whereas the R(1, *g*, *d*) configuration requires 16 wavelengths. With 16 wavelengths, R(*c*, *g*, *d*) has the potential to scale to as many as 4,096 nodes. However, R(1, *g*, *d*) has a lower latency because it has a smaller diameter than the R(*c*, *g*, *d*) configuration.

## Performance evaluation

We evaluated the Rapid R(1, *g*, *d*) configuration's performance using the RSIM simulator and compared it with a mesh interconnection for the Splash-2 benchmark suite. Unfortunately, because of the complexities of full-system simulation, we could not simulate systems consisting of more than 64 processors. To evaluate the R(*c*, *g*, *d*) configuration, we also used CSIM, a process-oriented, discrete-event model simulator using synthetic traffic workloads, and compared Rapid with several scalable electrical interconnect topologies.

### RSIM simulation

The Rice Simulator for ILP Multiprocessors (RSIM) models a mesh-based multiprocessor interconnection network subsystem, including contention at all resources. On the RSIM, we designed the Rapid network with WDM, implemented the token ring for group channel allocation, and modified the network interface for R(1, *g*, *d*). The network interface modification included adding a queue between remote receive and local send when packet forwarding was needed.

*Benchmarks.* In this study, we used five Splash-2 benchmarks, covering a spectrum of

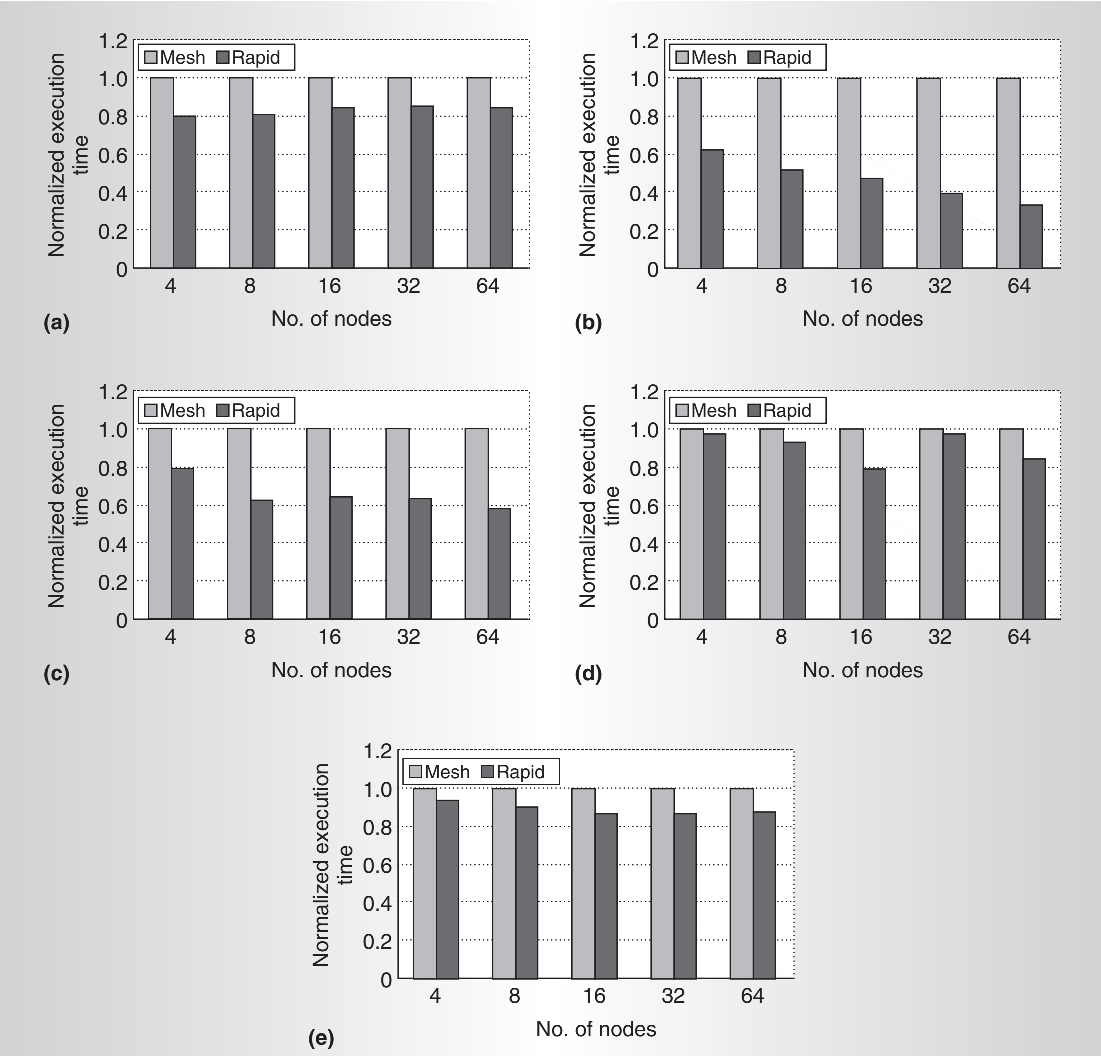

Figure 4. Normalized execution time for Rapid and mesh networks on RSIM simulator for workloads ranging from 4 to 64 nodes: FFT (a), Radix (b), Water (c), Ocean (d), and LU (e).

memory-sharing and access patterns. These included

- FFT with an input data set of 64,000 points,
- Radix with 1 million integers and 1,024 radices,
- Water-nsquared with 512 molecules,
- Ocean with 258 × 258 blocks, and
- LU with 256 × 256 and 16 × 16 blocks.

*Simulation parameters.* Each node of the simulated network contained a 1-GHz processor and had two cache levels: L1 (16-Kbyte, direct-mapped) and L2 (64-Kbyte, 4-way set-associative). Each node had four miss status holding registers (MSHRs). The L1 hit time was 1 ns, and the access time to the pipelined L2 cache was 15 ns. Memory access time was 70 ns with 4-way interleaving. We simulated a wormhole-routed, bidirectional, 2D mesh network with an 8-byte flit size and a 16-byte nondata size. The router speed was 500 MHz, the router's internal bus width was 64 bits, and the channel speed was 10 GHz. For the optical network, we assumed a channel speed of 10 GHz, based on current optical technology. At 10-Gbps data rates, transmission of an 8-byte address request took about 6.4 ns, and for a 64-byte cache line, the transmssion of the data took about 51.2 ns. We modeled optical-to-electrical and electrical-to-optical delays of 12.8 ns.

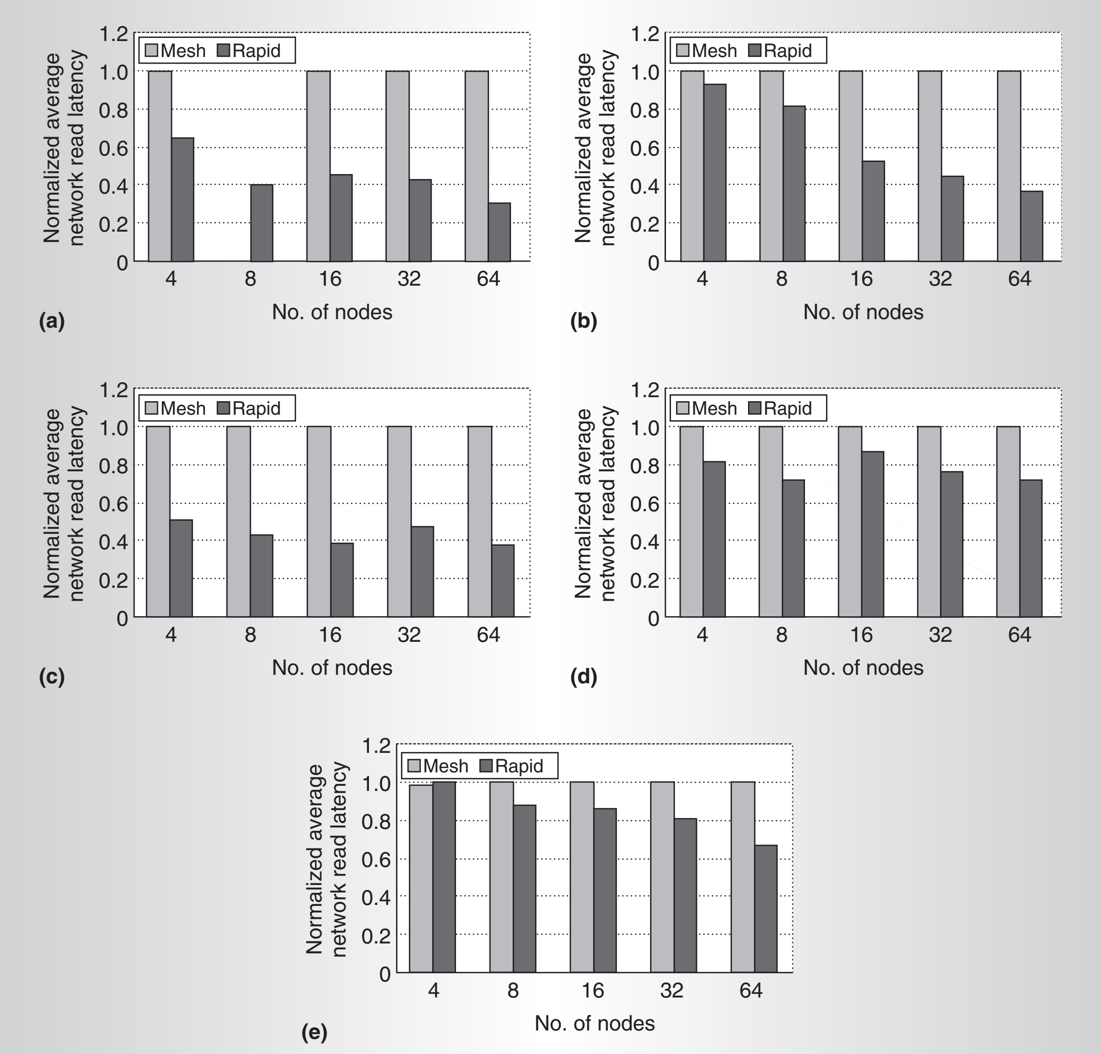

Figure 5. Normalized average network read latency for Rapid and mesh networks on RSIM simulator for workloads ranging from 4 to 64 nodes: FFT (a), Radix (b), Water (c), Ocean (d), and LU (e).

### RSIM simulation results

Figure 4 shows the normalized execution time for the five applications. We normalized the simulated time in clock cycles to the maximum of the two networks for each simulated number of nodes. In FFT, Rapid outperformed the mesh network by almost 20 percent in all simulation runs. This is attributable to larger protocol processor occupancies and the amount of contention in FFT. In the Radix and Water applications, Rapid outperformed the mesh network by more than 40 percent. Rapid outperformed the mesh network by more than 20 percent in Ocean and by 20 to 40 percent in LU. LU spends most of its time on synchronization points, resulting in hot spots on nodes.

Figure 5 shows the normalized average network read latency for the same applications. We normalized the average network read latency in clock cycles to the maximum of the two networks for each simulated run. Rapid showed an improvement of more than 60 percent over mesh interconnects in the FFT, Radix, and Water applications with 64 nodes. We attribute this to the architecture design that maximized channel availability and reduced queuing and waiting delays for channel allocation. In Ocean with 64 nodes, Rapid outperformed mesh by more than 20 percent, and in LU with 64 nodes, by more than 30 percent.

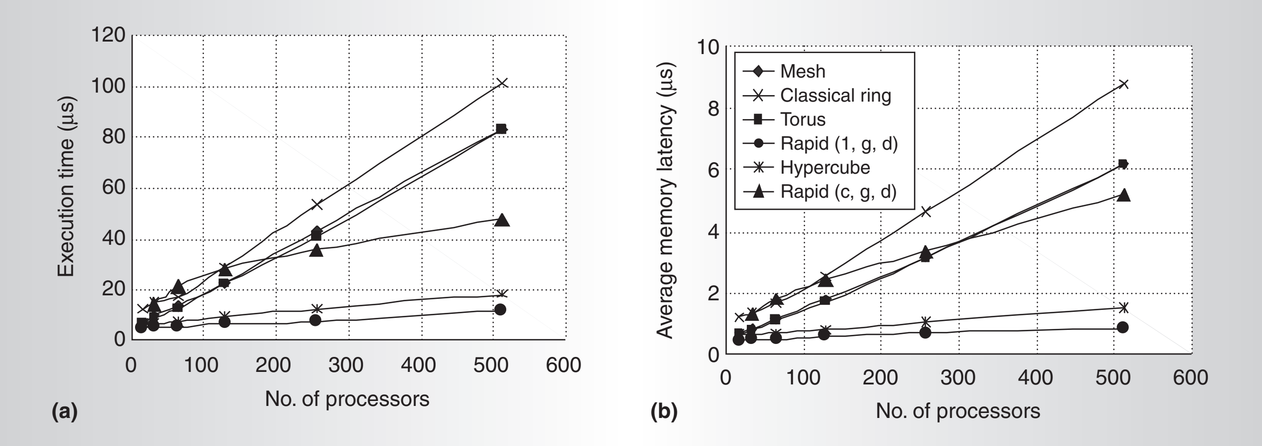

Figure 6. CSIM simulation results for classical-ring, mesh, torus, hypercube, Rapid R(1, *g*, *d*), and R(*c*, *g*, *d*) networks: execution time (a) and remote memory latency (b).

## CSIM simulation results

Using the CSIM simulator, we evaluated Rapid R(*c*, *g*, *d*)'s performance and compared it with electrical topologies such as the 2D mesh, the 2D torus, the hypercube, and the classical ring. We have detailed our simulation methodology in another publication.[12] Figures 6a and 6b show execution times and average memory latency for various topologies. Rapid R(1, *g*, *d*) outperforms all networks by maximizing channel availability, but it needs more wavelengths as system size increases. R(*c*, *g*, *d*) has a higher latency than most networks for small system configurations. However, as system size increases, memory latency in R(*c*, *g*, *d*) increases slowly, providing reasonable performance because its diameter doesn't change with an increased number of processors. These results show that R(1, *g*, *d*) can reduce latency for smaller system configurations by using more wavelengths and maintaining a low diameter. Moreover, R(*c*, *g*, *d*) can scale to very large configurations yet provide low latency by using minimal wavelengths.

This research is focused on developing new innovative architectures using optical interconnect technology in order to determine the insertion points in the hierarchy of parallel computing systems where optical technology can be beneficial. We intend to develop an end-to-end system modeling and simulation framework to evaluate the performance of Rapid and to design a small scale prototype of Rapid. MICRO

## Acknowledgments

This research is supported by NSF grants CCR-0000518 and CCR-0309537 and an Intel grant.

## References

1. D.E. Lenoski and W.D. Weber, *Scalable Shared Memory Multiprocessing*, Morgan Kaufmann, 1995.
2. B.E. Lemoff et al., "Maui: Enabling Fiber-to-the-Processor with Parallel Multiwavelength Optical Interconnects," *J. Lightwave Technology*, vol. 22, no. 9, Sept. 2004, pp. 2043-2054.
3. *Hypertransport Technology I/O Link*, tech. report 25012A, Advanced Micro Devices, 2001.
4. *White Paper: PCI Express Ethernet Networking*, tech. report 254108-002, Intel, 2003.
5. H. Cho, P. Kapur, and K.C. Saraswat, "Power Comparison between High-Speed Electrical and Optical Interconnects for Interchip Communication," *J. Lightwave Technology*, vol. 22, no. 9, Sept. 2004, pp. 2021-2033.
6. O. Kibar et al., "Power Minimization and Technology Comparison for Digital Free-Space Optoelectronic Interconnections," *J. Lightwave Technology*, vol. 17, no. 4, Apr. 1999, pp. 546-555.
7. *White Paper: The Cray Xd1 High Performance Computer*, tech. report WP-0020404, Cray, 2004.
8. D. Huang et al., "Optical Interconnects: Out of the Box Forever?" *IEEE J. Selected Topics in Quantum Electronics*, vol. 9, no. 2,

Mar.-Apr. 2003, pp. 614-623.

9. D.A.B. Miller, "Rationale and Challenges for Optical Interconnects to Electronic Chips," *Proc. IEEE*, vol. 88, no. 6, June 2000, pp. 728-749.
10. J.H. Collet et al., "Architectural Approaches to the Role of Optics in Mono and Multiprocessor Machines," *Applied Optics, Special Issue on Optics in Computing*, vol. 39, no. 5, 2000, pp. 671-682.
11. R.R. Patel et al., "Multiwavelength Parallel Optical Interconnects for Massively Parallel Processing," *IEEE J. Selected Topics in Quantum Electronics*, vol. 9, no. 2, Mar.-Apr. 2003, pp. 657-666.
12. A.K. Kodi and A. Louri, "A Scalable Architecture for Distributed Shared Memory Multiprocessors Using Optical Interconnects," *Proc. 18th Int'l Parallel and Distributed Processing Symp.* (IPDPS 04), IEEE Press, 2004, pp. 11-20.

**Avinash Karanth Kodi** is pursuing a PhD at the University of Arizona, Tucson. His research interests include design of high-speed optical interconnects for shared-memory multiprocessors, parallel processing, and cache coherence protocols. Kodi has a BEng in electronics and communication from Manipal Institute of Technology, Mangalore University, India, and an MS in computer engineering from the University of Arizona.

**Ahmed Louri** is a professor of electrical and computer engineering and chairman of the computer engineering program at the University of Arizona, Tucson. He is also the director of the university's Optical Networking and Parallel Processing Laboratory. His research interests include computer architecture, parallel processing, optical computing systems, and optical interconnection networks. Louri has MS and PhD degrees in computer engineering, both from the University of Southern California. He is a senior member of the IEEE.

Direct questions and comments about this article to Ahmed Louri, Electrical and Computer Engineering Dept., University of Arizona, 1230 E. Speedway Blvd., Tucson, AZ 85721; louri@ece.arizona.edu.