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

Avinash Karanth Kodi and Ahmed Louri
Department of Electrical and Computer Engineering
University of Arizona
Tucson, AZ - 85721, USA
E-mail:louri@ece.arizona.edu

Abstract

This paper proposes a highly connected optical interconnect based architecture that maximizes the channel availability for future scalable parallel computers such as Distributed Shared Memory (DSM) multiprocessors and cluster networks. As the system size increases, various messages (requests, responses and acknowledgments) increase in the network resulting in contention. This results in increasing the remote memory access latency and significantly affects the performance of these parallel computers. As a solution, we propose an architecture called RAPID (Reconfigurable and scalable All-Photonic Interconnect for Distributed-shared memory), that provides low remote memory access latency by providing fast and efficient unicast, multicast and broadcast capabilities using a combination of aggressively designed WDM, TDM and SDM techniques. We evaluated RAPID based on network characteristics and by simulation using synthetic traffic workloads and compared it against other networks such as electrical ring, torus, mesh and hypercube networks. We found that RAPID outperforms all networks and satisfies most of the requirements of parallel computer design such as low latency, high bandwidth, high connectivity, and easy scalability.

1 Introduction

Large-scale distributed shared-memory (DSM) architectures provide a shared address space supported by physically distributing the memory among different processors[1, 2]. The key strength of DSM systems is that communication occurs implicitly as a result of conventional memory access instruction (i.e. loads and stores) which makes them easier to program. One of the fundamental communication problem in DSM systems that significantly affects scalability, is the increase in remote memory access latency as the number of processors increase in the system. Latency reducing[3] techniques and latency hiding techniques[1, 2, 3] are commonly used to tolerate large remote latencies. However, these successful and efficient latency-tolerating techniques require much more bandwidth, and create much more memory traffic in the network. In addition, every transaction in a DSM system consists of a request, response (data) and several acknowledgment messages. As the system size increases, more processors are injecting more messages (both transaction related messages and latency tolerating requests) into the network that causes network contention[4] for various shared resources. Moreover, communication paradigms such as multicast and broadcast algorithms (essential for synchronization and to reduce hot spots) are generally more complex to implement and expensive (in terms of latency) using electrical interconnects.

One technology that has the potential for providing higher bandwidths and lower latencies at lower power requirements than current electronic-based interconnects is optical interconnects[5, 6]. The use of optics has been recognized widely as a solution to overcome many fundamental problems in high-speed and parallel data communications. This paper proposes an integrated solution to solve the remote memory access latency in DSMs and still be able to scale the network significantly using low-latency, high-bandwidth optical technology. The proposed architecture, called RAPID dramatically reduces the critical remote memory latency in high-performance DSMs, (1) by increasing the connectivity and maximizing the channel availability, (2) by using a decentralized media access protocol and wavelength re-use schemes, and (3) by using passive optical interconnects, thereby making the architecture cheaper and faster.

2 Architecture Overview

In this section, we describe and explain the design of RAPID architecture. A RAPID network is defined by a 3-tuple:(D,G,C) where C is the total number of clusters, G is the total number of groups per cluster and D is the

0-7803-8686-8/04/$20.00 ©2004 IEEE

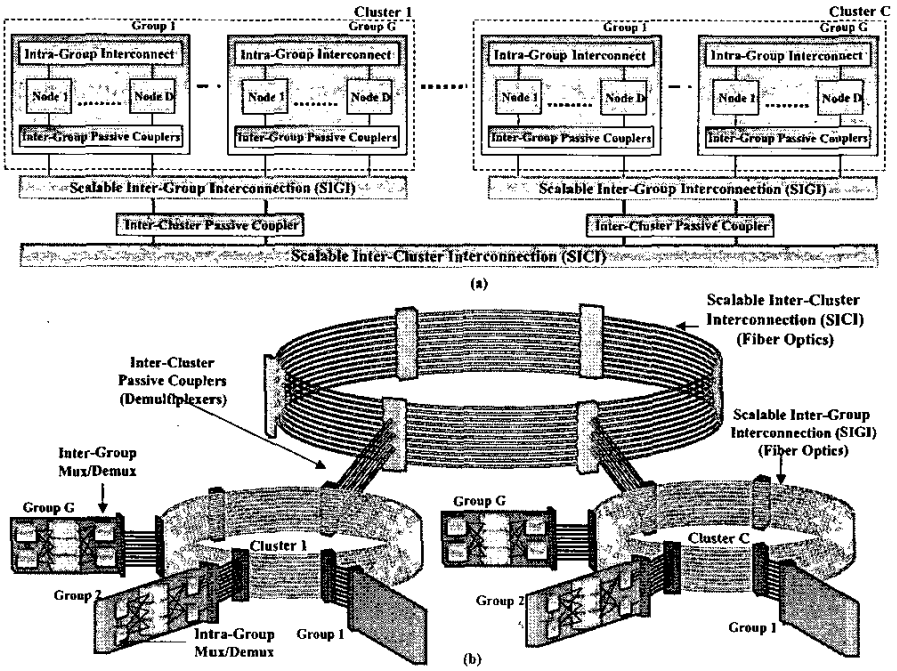

Figure 1. shows the RAPID network. Figure 1(a) shows D nodes connected using Intra Group Interconnect, G groups connected using SIGI and C clusters connected using SICI interconnects. Figure 1(b) shows the conceptual diagram of RAPID network.

total number of nodes per group. Each node is identified as R(d,g,c) where $1 \leq d \leq D$; $1 \leq g \leq G$; $1 \leq c \leq C$ such that $G \leq D\text{-}1$ and $C \leq D$. This condition enables every group to communicate to every other group/cluster.

Figures 1(a) and 1(b) show the RAPID architecture. In fig.1(a) each node in RAPID network, contains the processor and its caches, a portion of the machines physically distributed main memory, and a node controller. 1 up to D nodes are connected together to form a group. All nodes are connected to two sub-networks; a scalable Intra-Group interconnection (IGI) and a scalable Inter-Group Interconnection (SIGI) via the Inter-Group Passive Couplers (IGPC). SIGI is further connected to the Scalable Inter-Cluster Interconnection (SICI) using the Inter-Cluster Passive Couplers to increase the scalability of the architecture. We have separated intra-group (local) and inter-group/inter-cluster (remote) communications from one another in order to provide a more efficient implementation for both communications. Figure 1(b) shows the conceptual diagram of RAPID network.

Figure 2 shows the functional diagram of RAPID. As seen, the figure shows D = 4 (nodes), G = 4 (groups) and C = 1 (cluster). Within a group, all nodes are connected to multiplexers and demultiplexers for intra- and inter-group/inter-cluster communication. We will use this system to discuss the wavelength allocation, message routing for both local and remote communication and, the design of RAPID to support multicast and broadcast communications.

Wavelength Assignment and Routing for Group Communication: We propose an efficient wavelength assignment strategy based on wavelength re-use and spatial division multiplexing (SDM) techniques[10]. The proposed methodology allows wavelengths to be re-used when they are spatially separated, that is, when they are used at the local (intra-group) level, remote (inter-group) level or remote (inter-cluster) level. The number of wavelengths employed for intra-group communication equals the maximum number of nodes, D located in each group of the system. Figure 2(a) shows an example of the intra-group wavelength assignment and shows group 1 of cluster 1. The wavelengths located next to each node correspond to the wavelength that each node receives on. This same wavelength assignment applies to all groups in the system i.e. R(1,g,c) will always receive on λ_1 for intra-group communication. For example, for node R(1,1,1) to transmit to node R(3,1,1) in group 1, node R(1,1,1) would simply transmit on the wavelength assigned to node R(3,1,1) (e.g. λ_3). Therefore, distinct wavelength allocation in different groups is possible by assigning an unique wavelength to every node at which

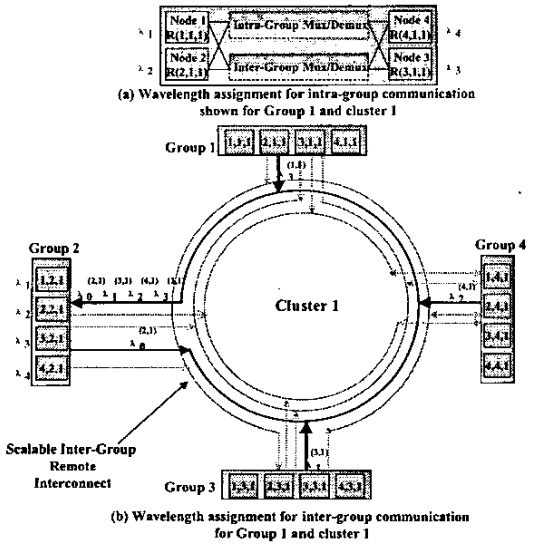

Figure 2. Wavelength assignment for intra- and inter-group communication.

it can receive optical packet from other intra-group nodes.

For the remote wavelength assignment scheme, we study two cases: (1) a single cluster configuration R(d,g,1) and (2) multi-cluster configuration R(d,g,c). In our remote wavelength assignment scheme shown in Figure 2(b) for R(d,g,1), the objective here is to selectively merge different wavelengths from various groups to provide high connectivity and at the same time to maximize the channel utilization. All nodes within the source group are assigned a unique wavelength at which the nodes can transmit to communicate with any destination group. At the destination group, each node receives optical signals at a unique wavelength as shown by the wavelength located next to each node in Figure 2(b) for group 2. Remote wavelengths are indicated by $\lambda_i^{(j,k)}$, where i is the wavelength, j is the group number and k is the cluster number from which the wavelength originates. In Figure 2(b), any node in group 3 can communicate with group 2 on $\lambda_1^{(3,1)}$, any node in group 4 can communicate with group 2 on $\lambda_2^{(4,1)}$ and any node in group 1 can communicate with group 2 on $\lambda_1^{(1,1)}$. For clarity, only the wavelengths received by group 2 are shown in bold in fig 2(b). Note here that, the wavelength $\lambda_0^{(2,1)}$ is the wavelength at which group 2 communicates with itself. This wavelength is used to multicast transaction requests to all nodes within a group. Now, the multiplexed signal received by group 2 is demultiplexed and node R(1,2,1) receives signals on wavelength $\lambda_1^{(3,1)}$, node R(2,2,1) receives signals on wavelength $\lambda_2^{(4,1)}$, node R(3,2,1) receives signals on wavelength $\lambda_3^{(1,1)}$ and the signal on wavelength $\lambda_0^{(2,1)}$ is broadcast to every node within group 2. For remote traffic, the number of wavelengths required to obtain the connectivity mentioned above, is G i.e. $(G-1)$ wavelengths are required to communicate with every other group and 1 wavelength for multicast communication. This gives us the criteria, that the there should exist at least $G-1$ nodes within a group to receive data from other groups. The maximum number of wavelengths then required for either intra-group or remote inter-group communication for R(d,g,1) configuration is, simply D. This represents an order of magnitude reduction in the total number of wavelengths required compared to a straight forward wavelength assignment where each group is associated with a distinct wavelength.

Remote inter-group communication takes place when both the source and destination nodes are on different groups, $\text{R(x,g,1)}_{source} \neq \text{R(y,h,1)}_{destination}$. Now, node $R(x,g,1)$ can transmit the packet on a specific wavelength to group h. The destination node in group h which can receive the packet from group g may not be node y (the intended destination). To illustrate this, consider Figure 2(b). Let the source node be R(1,3,1) (group 3) and the destination node be R(3,2,1) (group 2). The source node can transmit to group 2 on wavelength $\lambda_1^{(3,1)}$. The destination node which receives packets for remote communication in group 2 on wavelength $\lambda_1^{(3,1)}$ is R(1,2,1). Node R(1,2,1) then uses the intra-group interconnection to forward the packet to node R(3,2,1) on wavelength λ_3. In some cases, source node R(x,g,1) may directly transmit to destination node R(y,h,1). As in the previous example, if the destination was node R(1,2,1), then node R(1,3,1) could directly transmit on $\lambda_1^{(3,1)}$ which is received by node R(1,2,1), the intended destination.

Wavelength Assignment and Routing for Inter-Cluster Communication: Inter-cluster communication for configuration R(d,g,c), we extend the basic configuration of R(d,g,1) shown in Figure 2(b) by replacing group 4 and connecting cluster 1 to SICI using low loss passive bi-directional demultiplexers as shown in Figure 3. For inter-cluster communication, different wavelengths from different clusters are selectively merged and and dropped at every cluster by demultiplexing the signals propagating on the SICI. Remote inter-cluster communication takes place when both the source and destination nodes are on different clusters, $\text{R(x,g,c)}_{source} \neq \text{R(y,h,d)}_{destination}$. To illustrate this, consider that the source node is R(2,3,1) and the destination node is R(4,3,2). Cluster 2 is reachable by group 1 within cluster 1 by using the same wavelength assignment as explained above. Node R(2,3,1) transmits on wavelength $\lambda_2^{(3,1)}$ and is received by the intermediate node 1a shown by R(2,1,1). In order to communicate to cluster 2, node R(2,1,1) transmits on wavelength $\lambda_1^{(1,1)}$. This

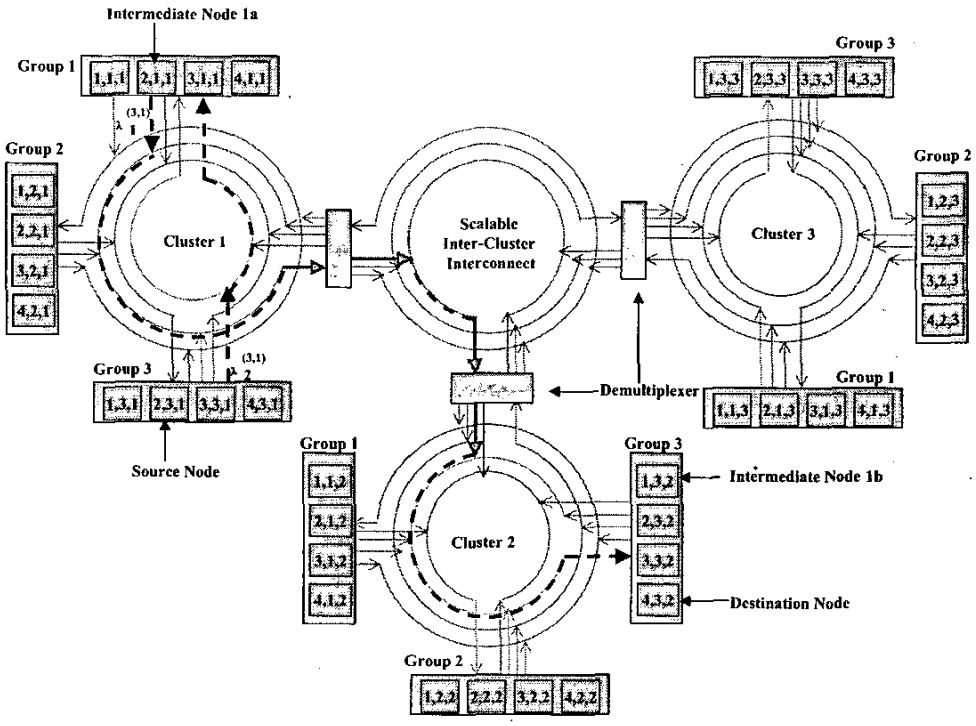

Figure 3. Scalable Inter-Cluster Interconnection.

path is shown by the bold dotted lines to indicate the transmission from group 1 in cluster 1 to group 3 in cluster 2. Note, that the signals from cluster 1 is first demultiplexed and merged with SICI selectively such that different wavelengths are multiplexed to different fibers. At cluster 1, the multiplexed signals are again demultiplexed, and wavelength $\lambda_1^{(1,1)}$ is received by intermediate node 1b R(1,3,2) as shown in Figure 3. In order to reach the intended destination R(4,3,2), node R(1,3,2) transmits using the intra-group interconnection on wavelength λ_4. The maximum diameter of R(d,g,c) is 4. From the previous example, if the intended destination was located in a group other than group 3, then it would require an additional hop. The configuration R(d,g,c) trade-offs wavelength usage to latency for smaller system sizes. For example, by using 4 wavelengths and passive optical components, R(d,g,c) can accommodate 64 nodes, where as in R(d,g,1) configuration, to design a network with 64 nodes, 16 wavelengths are required. With 16 wavelengths, R(d,g,c) has the potential to scale to as many as 4096 nodes. However, R(d,g,1) has a lower latency due to lower diameter than the R(d,g,c) configuration.

Time division multiple access (TDMA) protocol is used as a control mechanism to achieve mutual exclusive access to the shared local and remote communication channels[7, 8]. In this paper, we consider an optical token based TDMA protocol with pre-allocation to prevent collision of requests by different processors. A novel media access protocol is discussed for RAPID so as to minimize the remote access latency. The optical tokens generated for inter-group/inter-cluster communications are shared among the nodes locally connected and not among all nodes. This is a significant feature of the proposed network, as the queuing time to transmit the packets reduces considerably. In RAPID, under worst case scenario, a node waits only for $D-1$ transmissions of the packet to a particular remote destination group/cluster before it can transmit its request, thereby significantly reducing the remote group latency. We generate two sets of token for every intra-group g; one set of D tokens are shared for inter-group/inter-cluster communications and the other set of $(G \leq D)$ tokens are shared for inter-group communications. These local and global token are shared by the intra-group nodes connected to the concerned group i.e. R(g). In order to prevent collision of requests, a processor can transmit an address request, response or an acknowledgement to another processor (local/remote) depending on the token received. The implementation details of RAPID have been published elsewhere[9].

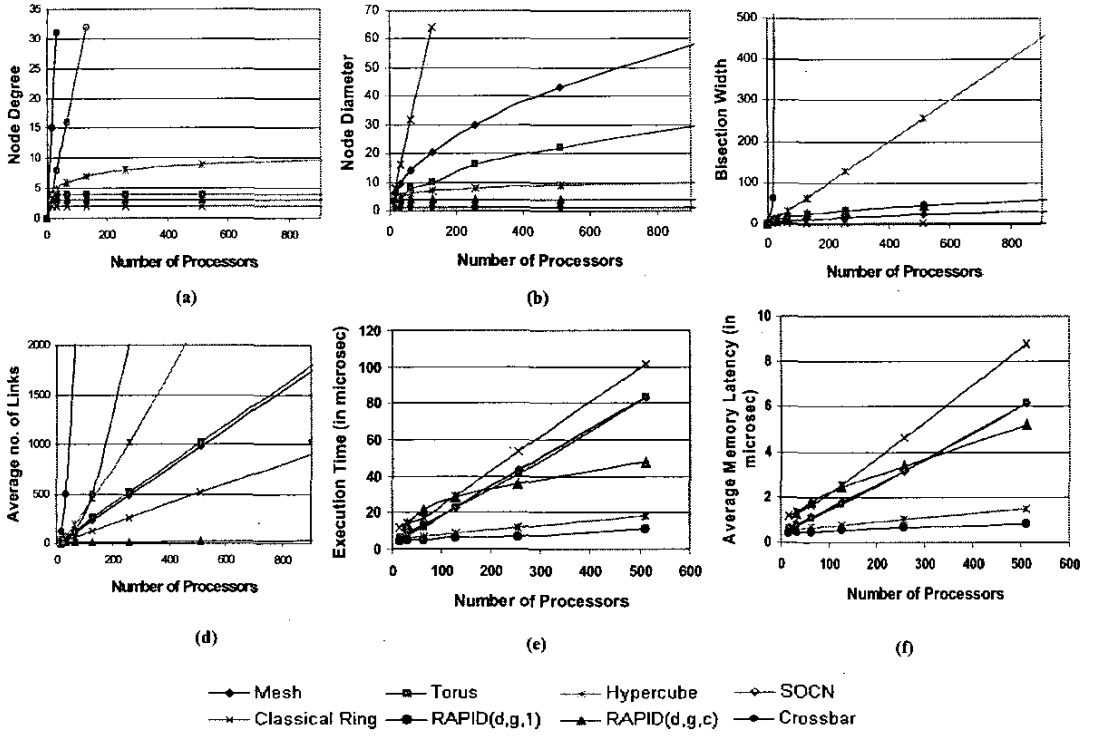

Figure 4. Fig (a) shows the degree comparison, (b) shows the diameter comparison, (c) shows the bisection width comparison for varying number of processors, (d) shows the no. of links comparison, (e) shows the execution time for various topologies simulated and (f) shows the average remote memory access latency for various topologies.

3 Performance Analysis

In this section, we evaluate the performance of RAPID for DSMs by analyzing the network characteristics and performance based on simulation. For clarity, only RAPID R(d,g,c) configuration is compared to several well-known network topologies such as a traditional crossbar network (CB), the Binary Hypercube, the Ring network, the Torus, 2-D Mesh and Scalable Optical Crossbar Network (SOCN)[10] as shown in Figures 4(a-d). Each of these networks are compared with respect to degree, diameter, number of links and bisection width. RAPID network maintains a constant degree for any system size as each node is connected to only intra- and inter-group/cluster interconnects as seen in Figure 4(a). RAPID supports better connectivity at a reasonable cost as the comparison of the diameter is seen in Figure 4(b). The bisection width of RAPID network is very comparable to the best of the scalable networks as seen in Figure 4(c). RAPID shows the least cost for inter-cluster communication, thereby showing a much better scalability in the number of links for very large-scale systems as seen in Figure 4(d).

Simulation Assumptions and Methodology In this section, we describe the simulation methodology and the preliminary results obtained by comparing both R(d,g,1) and R(d,g,c) with few scalable electrical networks such as the 2-D Mesh, 2-D Torus, Hypercube and the classical ring. We use CSIM[11], a process-oriented, discrete-event model simulator to evaluate the performance of RAPID network using synthetic traffic workloads. Due to the complexities of a full system simulation and the difficulty in tuning the simulator for large number of nodes, we currently present data for as many as 512 nodes. For the electrical network, wormhole routing is modelled with a flit size of 8 bytes and up to 4 virtual channels per link. Various routing, switching and propagation times[12] are chosen such that they reflect future high performance electrical interconnect technology. For the optical network, we

assume a channel speed of 10 Ghz, based on current optical technology. We model O/E (optical to electrical) and E/O (electrical to optical) for both configurations. In this simulation, we model accurately contention at all resources for both electrical and optical networks, and is not presented here due to page constraints[9].

Simulation Results: We evaluated RAPID network with other electrical topologies such as the classical ring, the hypercube, the 2-D mesh and the 2-D torus based on execution time and average remote memory latency. Figure 4(e) shows the execution time for varying number of processors for both the simulated electrical and optical networks. RAPID R(d,g,1) outperforms all networks by maximizing the the channel availability and maintaining a low diameter for large number of processors. RAPID R(d,g,1) outperforms the classical ring by almost 89% for 512 nodes. The mesh and torus have similar latencies, with R(d,g,1) configuration outperforming them by almost 86% for 512 nodes. The hypercube performs reasonably well, though R(d,g,1) outperforms hypercube by almost 38%. R(d,g,c) actually has a higher latency than most networks for small system configurations. But, as the system size increases, the curve for R(d,g,c) starts to flatten showing a reasonable performance as the diameter doesn't change with increase in number of processors. For system configurations greater than 512, we expect the latency for R(d,g,c) configuration to further stabilize and perform better than other networks. All electrical networks showed different latencies depending on how many switches needed to be traversed. Figure 4(f) shows the average remote memory access latency. RAPID R(d,g,1) performed the best as compared to all other networks. RAPID R(d,g,1) outperformed hypercube by 46%, the mesh, torus by 87% and the classical ring by 91%. These results show that RAPID R(d,g,1) can reduce the latency for smaller system configurations by using more wavelengths and maintaining low diameter. Additionally, RAPID R(d,g,c) can scale to very large configurations, yet provide low latency by using minimal wavelengths.

4 Conclusion

In this paper, we proposed an optically interconnected architecture called RAPID to reduce the remote memory access latency in distributed shared memory multiprocessors. RAPID was completely designed using passive optical technology making the proposed architecture much faster and inexpensive as compared to other optical and electrical architectures. RAPID, not only maximizes the channel availability for inter-group communication, but at the same time wavelengths are completely re-used for both intra-group and inter-group communications. This novel architecture fully utilizes the benefits of wavelength division multiplexing along with space division multiplexing to produce a highly scalable, high bandwidth network with low overall latency that could be very cost effective to produce.

Acknowledgement This research is sponsored by NSF grant no. CCR-0000518.

References

[1] J.Laudon and D.Lenoski, "Sgi origin: A ccnuma highly scalable server," in *Proceedings of the 24th Annual International Symposium on Computer Architecture*, June 1997, pp. 241–251.

[2] David E. Culler, Jaswinder Pal Singh, and Anoop Gupta, *Parallel Computer Architecture: A Hardware/Software Approach*, Morgan Kaufmann, San Fransisco, 1999.

[3] Jose Duato, Sudhakar Yalmanchili, and Lionel Li, *Interconnection Networks: An Engineering Approach*, IEEE Computer Society Press, New Jersey, 1997.

[4] Donglai Dai and Dhabaleswar K. Panda, "How much does network contention affect distributed shared memory performance," in *International Conference on Parallel Processing (ICPP '97)*, 1997, pp. 454–461.

[5] David A.B.Miller, "Rationale and challenges for optical interconnects to electronic chips," *Proceedings of the IEEE*, vol. 88, pp. 728–749, June 2000.

[6] J.H. Collet, D. Litaize, J. V. Campenhut, C. Jesshope, M. Desmulliez, H. Thienpont, J. Goodman, and A. Louri, "Architectural approaches to the role of optics in mono and multiprocessor machines," *Applied Optics, Special issue on Optics in Computing*, vol. 39, pp. 671–682, 2000.

[7] Patrick Dowd, James Perreault, John Chu, David C. Hoffmeister, Ron Minnich, Dan Burns, Frank Hady, Y. J. Chen, and M. Dagenais, "Lighnting network and systems architecture," *Journal of Lightwave Technology*, vol. 14, pp. 1371–1387, 1996.

[8] Joon-Ho Ha and T.M.Pinkston, "The speed cache coherence for an optical multi-access interconnect architecture," in *Proceedings of the 2nd International Conference on Massively Parallel Processing Using Optical Interconnections*, 1995, pp. 98–107.

[9] Avinash Karanth Kodi and Ahmed Louri, "A scalable architecture for distributed shared memory multiprocessors using optical interconnects," in *to appear in 18th International Parallel and Distributed Processing Symposium*, April 2004.

[10] Brian Webb and Ahmed Louri, "A class of highly scalable optical crossbar-connected interconnection networks (socns) for parallel computing systems," *IEEE Transactions on Parallel and Distributed Systems*, vol. 11, no. 1, pp. 444–458, May 2000.

[11] Herb Schwetman, "Csim19: A powerful tool for building system models," in *Proceedings of the 2001 Winter Simulation Conference*, 2001, pp. 250–255.

[12] Mauael E. Acacio, Jose Gonzalez, Jose M. Garcia, and Jose Duato, "The use of prediction for accelerating upgrade misses in cc-numa multiprocessors," in *Proceedings of the International Conference on Parallel Architectures and Compilation Techniques*, 2002, pp. 155–164.