Scalable Optical Interconnection Network for Parallel and Distributed Computing

Avinash Karanth Kodi and Ahmed Louri
Electrical and Computer Engineering Department, University of Arizona, 1230 E Speedway Blvd, Tucson, AZ – 85721.
E-mail: louri@ece.arizona.edu

Abstract: In this paper, a high-performance, scalable, parallel computing system called RAPID is designed using switchless, passive optical interconnect technology. RAPID outperforms current electrical multiprocessor systems by significantly decreasing the remote memory access latency.

OCIS codes: (200.0200) Optical Computing; (200.4650) Optical Interconnects

1. Introduction

Market demands and the explosive growth in internet applications are accelerating the need for scalable parallel computer systems. Architectural models, such as the distributed shared memory multiprocessor (DSM) and the message passing (MP) models are emerging as the design choices for scalable high-performance computers. Large-scale distributed shared-memory (DSM) architectures provide a shared address space by physically distributing the memory among different processors[1]. Message passing multi-computers and clusters are popular as they can be built from commercially off-the-shelf components for applications such as distributed computing and storage area networks. Parallel and distributed computing are big business, over a 60 billion-a year market[2]. One of the fundamental communication problem in parallel computers that significantly affects scalability, is the increase in remote memory access latency with increase in the number of processors in the system. A remote memory access takes 1-2 orders of magnitude longer than the local memory access, with most of the time consumed in communication over the interconnection network of the machine. Though latency tolerating/hiding techniques are frequently used in parallel systems to reduce remote latency, these techniques require more bandwidth and create much more memory traffic by fetching more data than is needed[1]. Moreover, the power dissipation in electrical interconnects increases with length and higher bit rates due to larger attenuation, and a greater impact of unattenuated as well as fixed noise sources[3], thereby limiting parallel computers from reaching their full potential.

One technology that has the potential for providing higher bandwidths, lower crosstalk, zero EMI, and lower latencies at lower power requirements than current electronics-based interconnect is the optical interconnect[4, 5]. 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 optical technology for both board-to-board and backplane communications. As a solution, we proposed an optical interconnect called **RAPID** (Reconfigurable All-Photonic Interconnect for DSMs) in [6] which included preliminary theoretical simulation data. This paper extends the architecture by adding clusters, discusses possible implementation and includes execution-driven simulation results. RAPID reduces remote memory access latency by (1) increasing the connectivity, maximizing the channel availability and providing scalable bandwidth using a combination of WDM, TDM and SDM techniques; (2) using a decentralized wavelength allocation scheme along with efficient re-use of the available wavelengths and (3) using a switchless topology that is based on passive optical interconnect technology that reduces the cost and improves performance significantly.

2. RAPID Architecture

A RAPID network is defined by 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 $0 \le d \le D-1$; $0 \le g \le G-1$; $0 \le c \le C-1$ such that $G \le D-1$ and $C \le D$. Figure 1 shows the RAPID architecture. In Figure 1, 0 up to D-1 nodes are connected together to form a group. 0 up to G-1 groups are connected to form a single cluster. All nodes are connected to two sub-networks; a scalable Intra-Group interconnection (IGI) and a Scalable Inter-group Interconnection (SIGI) via passive couplers. 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. RAPID is designed such that every node has two sets of tunable transmitters and fixed receivers for intra- and inter-group communication. All interconnections on the board are implemented using optical waveguides and the interconnections from the board to SIGI are implemented using optical fiber using multiplexers

and demultiplexers. RAPID can be scaled in three ways; by adding more nodes D within a group G, by adding more groups G within a cluster, or by replicating the existing network such as adding another cluster to an existing cluster to form a new inter-cluster interconnect. This inter-cluster interconnect is similar to Figure 1, except, now every group is a cluster by itself.

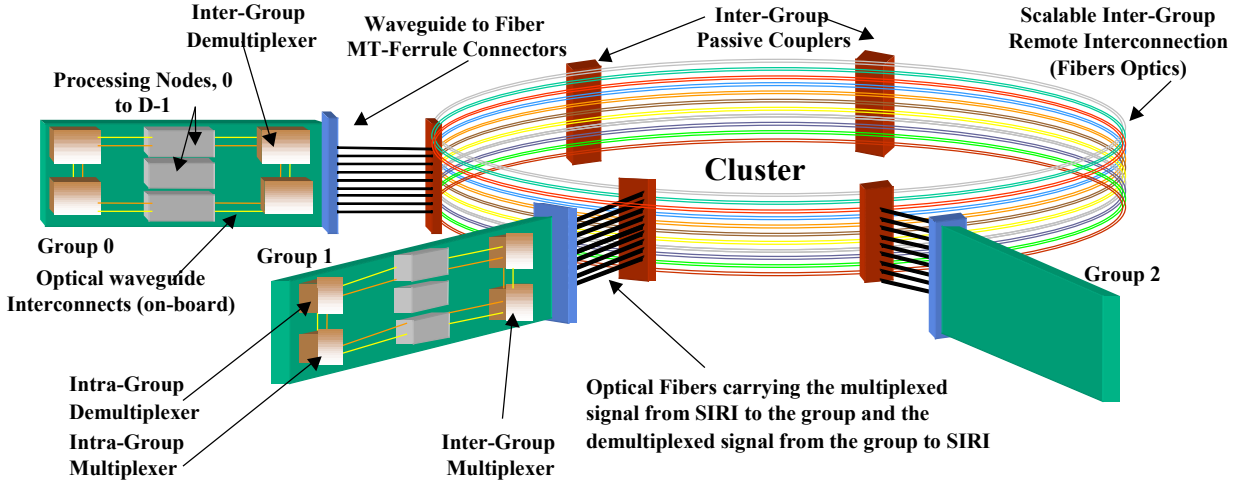

Figure 1: The architectural overview of RAPID.

We proposed a novel method based on wavelength re-use and spatial division multiplexing (SDM) techniques to design an efficient wavelength assignment and routing strategy[6]. The proposed methodology allows wavelengths to be re-used when they are spatially separated and by doing so, we can have a much greater number of nodes while requiring only a small number of distinct wavelengths to implement the entire system. We propose a possible optical implementation of the proposed architecture, which could be constructed directly onto the PC board as shown in Figure 2(a). Each PC board is a group, containing a few processing nodes as shown. The nodes are connected to the intra-group and inter-group multiplexer/demultiplexer. The demultiplexer used in our proposed architecture is the low loss arrayed waveguide grating (AWG) that can be integrated using planar waveguide technology.

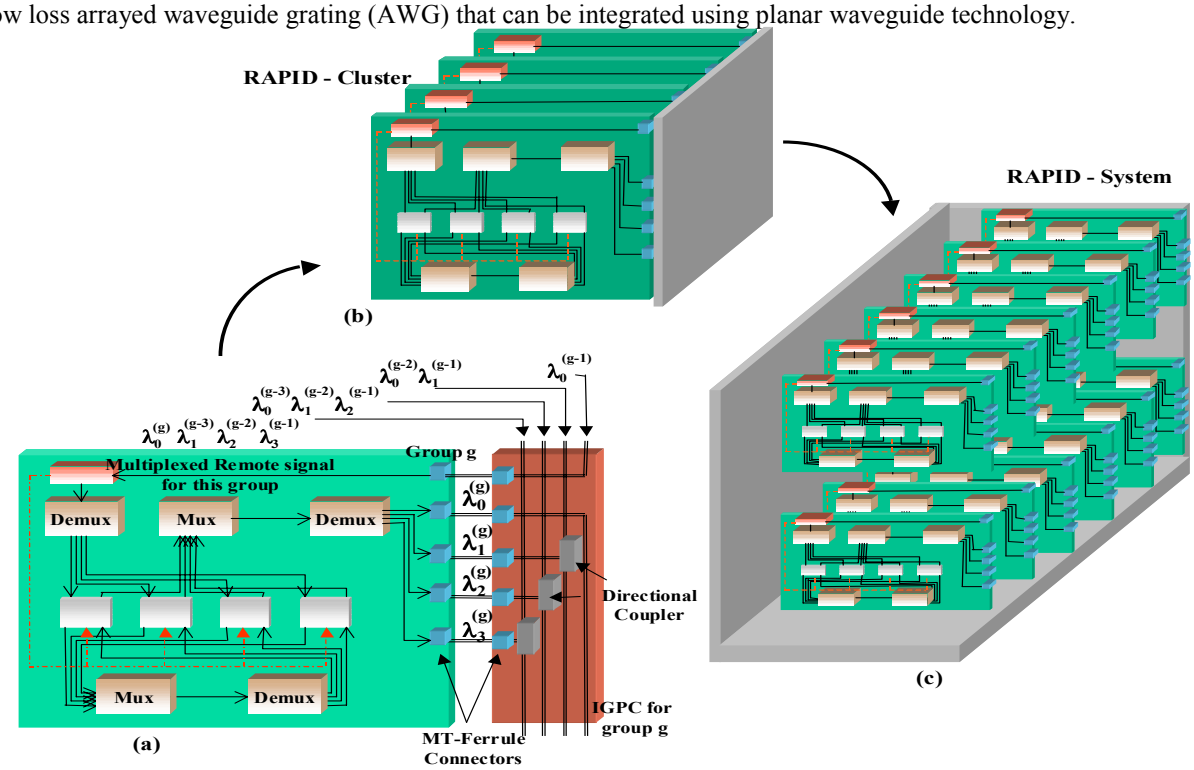

Figure 2: The proposed implementation of RAPID at (a) group-level (intra-board communication) (b) cluster-level (inter-board communication) and (c) system-level (inter-cluster communication)

For inter-group communication, signals on different wavelengths from the nodes are combined using remote multiplexer, directional couplers and IGPC demultiplexer. Each demultiplexed signal is then selectively merged with the traffic on the scalable remote inter-group interconnects using low loss couplers. As shown, wavelength $\lambda_1^{(g)}$ from this group g, is coupled with the fiber containing the signal on wavelength $\lambda_0^{(g-1)}$ originating from the previous group (g-1). Similarly, wavelength $\lambda_2^{(g)}$ from this group g, is merged with the fiber containing signals on wavelengths $\lambda_0^{(g-2)}$ and $\lambda_1^{(g-1)}$. The multiplexed signal containing wavelengths $\lambda_0^{(g)}$ $\lambda_3^{(g-1)}$ $\lambda_2^{(g-2)}$ and $\lambda_1^{(g-3)}$ will be received by the group g, which are then demultiplexed to appropriate destination nodes. G groups could be combined to form a cluster interconnect for board-to-board communication as shown in Figure 2(b). Similarly, C clusters could be combined to form the RAPID system for inter-cluster communication as shown in Figure 2(c).

3. Simulation & Results

The performance of RAPID configuration is evaluated using RSIM simulator[7] and is compared with electrical mesh interconnection for Splash-2 suite benchmarks from 4-64 nodes. RSIM models an electrical mesh based multiprocessor interconnection network subsystem, including contention at all resources. On RSIM, we designed the RAPID network with WDM, and modified the network interface. In this study, we used several benchmarks, covering a spectrum of memory sharing and access patterns from the SPLASH-2 suite, only a few results are shown due to page limitations. These include FFT with input data set 64K points; Radix with 1M integers, 1024 radix and Water-nsquared with 512 molecules. Figure 3 shows the normalized execution time for various applications. The simulated time in clock cycles was normalized to the maximum of the two networks for each simulated number of nodes. RAPID outperformed mesh electrical network by almost 40% for all simulation runs as shown in Figure 3.

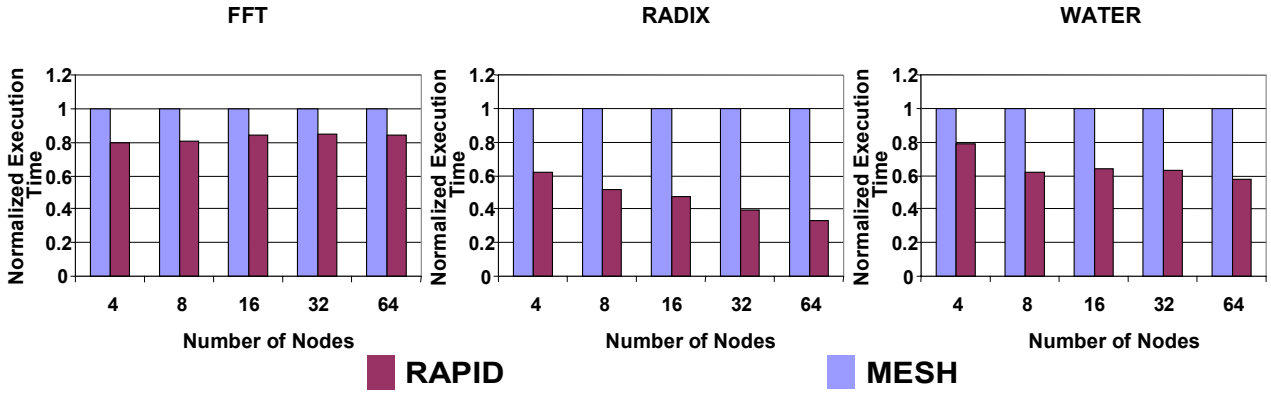

Figure 3. Normalized execution time for RAPID and Mesh networks on RSIM simulator for different workloads ranging from 4-64 nodes.

4. Conclusion

We extended RAPID architecture, proposed the implementation of RAPID at the board, inter-board and inter-cluster levels. We evaluated the performance of RAPID using RSIM and found that RAPID outperforms mesh electrical network by 40%. RAPID fully utilizes the benefits of WDM along with SDM and TDM to produce a highly scalable, high bandwidth network with low overall latency that could be very cost effective to produce.

Acknowledgement: This research is supported by NSF grants CCR-0000518, CCR-0309537 and a grant from Intel Corporation.

5. References

[1] D. E. Lenoski and W.-D. Weber, *Scalable Shared-Memory Multiprocessing*. (Morgan Kaufmann Publishers Inc., 1995)

[2] A. Louri and A. K. Kodi, "An Optical Interconnection Network and a Modified Snooping Protocol for the Design of Large-Scale Symmetric Multiprocessors (SMPs)," IEEE Trans. on Parallel and Distributed Systems, **15**, 1093-1104 (2004).

[3] H.Cho, P.Kapur, and K.C.Saraswat, "Power Comparison Between High-Speed Electrical and Optical Interconnects for Interchip Communication," IEEE/OSA Journal of Lightwave Technology, **22**, 2021-2033 (2004).

[4] D. A.B.Miller, "Rationale and Challenges for Optical Interconnects to Electronic Chips," Proceedings of the IEEE, **88**, 728-749, (2000).

[5] 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, **39**, 671-682, (2000).

[6] A. K. Kodi and A. Louri, "RAPID: Reconfigurable and All-Photonic Interconnect for Distributed Shared Memory Multiprocessors," IEEE/OSA Journal of Lightwave Technology, **22**, 2101-2110, (2004).

[7] V. Pai, P. Ranganathan, and S. Adve, "RSIM: A simulator for shared-memory multiprocessor and uniprocessor systems that exploit ILP," Proceedings of the 3rd Workshop on Computer Architecture Education, 1997.