RAPID: Reconfigurable and Scalable All-Photonic Interconnect for Distributed Shared Memory Multiprocessors

Avinash Karanth Kodi, *Student Member, IEEE,* and Ahmed Louri, *Senior Member, IEEE*

***Abstract*—In this paper, we describe the design and analysis of a scalable architecture suitable for large-scale distributed shared memory (DSM) systems. The approach is based on an interconnect technology which combines optical components and a novel architecture design. In DSM systems, numerous shared memory transactions such as requests, responses and acknowledgment messages propagate simultaneously in the network. As the network size increases, network contention results in increasing the critical remote memory access latency, which significantly penalizes the performance of DSM systems. In our proposed architecture called reconfigurable and scalable all-photonic interconnect for distributed-shared memory (RAPID), we provide high connectivity by maximizing the channel availability for remote communication to reduce the critical remote latency. RAPID provides fast and efficient unicast, multicast and broadcast capabilities using a combination of aggressively designed wavelength division multiplexing (WDM), time division multiplexing (TDM), and space division multiplexing (SDM) techniques. RAPID is wavelength-routed, permitting the same limited set of wavelength to be reused among all processors. We evaluated RAPID based on network characteristics, power budget criteria, 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 still provides good performance as the network is scaled to very large numbers.**

***Index Terms*—Cache coherence, distributed shared memory, optical interconnects, scalable optical networks.**

I. Introduction

LARGE-SCALE distributed shared-memory (DSM) multiprocessors, also called cache-coherent nonuniform memory access (cc-NUMA) systems, 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. Each processor has its own hierarchy of caches to retain recently accessed data that can be quickly reused, thereby avoiding contention for further memory accesses. In DSMs, the use of private cache, poses the complex problem of cache coherence; namely how data, coherence is maintained among various copies of data which can reside in multiple caches and main memory. Snooping cache coherence protocol broadcasts every transaction to all nodes in contrast to directory-based protocols which depend on maintaining the identity of sharers (at the directory) to avoid the need for broadcasts, and are much better suited for larger designs. The requesting nodes transmit coherence transaction over an arbitrary point-to-point network to the directory entry (home node), which either replies with the data (if the block is clean) or forward the request to the owner node that is caching the block (if the block is dirty) [1], [2]. An example of the state-of-the-art DSM machine is the SGI Origin 2000 [3] which can scale up to 512 nodes. 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 nodes in the system increases. (The remote memory latency is the latency in accessing a memory location in a processor other than the one originating the request and includes both the communication latency and data access latency from remote memory.) 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. Latency reducing techniques (reduces average time between when the processor issues a reference and when the memory responds) and latency hiding techniques [4] (overlaps memory access with useful processor computation) 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 [5], [6] for various shared resources such as virtual channels, network buffers, network interface accesses, etc. Moreover, synchronization operations, that are required to implement critical sections or to exploit parallelism across loop iterations in a parallel program can lead to highly contended and concentrated accesses of shared data objects (hot spots) for a short duration [7]. 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.

Additionally, the International Technology Road Map for semiconductors projects that by 2010 the off-chip clock speeds

Manuscript received December 25, 2003; revised May 11, 2004. This work was supported by NSF Grant CCR-0000518.

The authors are with the Department of Electrical and Computer Engineering, University of Arizona, Tucson, AZ 85721 USA (e-mail: louri@ece.arizona.edu).

Digital Object Identifier 10.1109/JLT.2004.833249

to reach 1.8 GHz, the width of the off-chip bus to reach 3000 high speed lines and a total off-chip I/O capacity of 4–40 Tb/s, which will be a major challenge to achieve using conventional electrical wire technologies. With growth in the system size of DSMs, increase in overhead costs of the interconnect and coherence mechanism along with lack of sufficient memory and communication bandwidths cumulatively result in a significant increase in the critical remote memory access latency and makes it challenging to scale DSMs to a large number of nodes while maintaining reasonable performance levels across a wide variety of applications at a reasonable cost.

A. Optical Interconnects for Distributed Shared Memory Multiprocessors

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 [8], [9]. The use of optics has been recognized widely as a solution to overcome many fundamental problems in high-speed and parallel data communications. Recently, there have been significant developments in optical and optoelectronic devices (vertical cavity surface emitting laser and photodetector arrays [10]–[12], arrayed waveguide grating [13], [14], microoptical components [15], etc.) and packaging technologies (OE-VLSI heterogeneous integration [16], smart pixel technology [17]) which make optical interconnects a viable and cost-effective option for building high bandwidth, low latency, and scalable optical interconnection networks.

This paper proposes the application of optical interconnect to the design of scalable interconnection networks for distributed shared-memory parallel computing systems. The proposed architecture, called "reconfigurable and scalable all-photonic interconnect for distributed-shared memory" (RAPID), dramatically reduces the critical remote memory latency in high-performance DSMs

1) by *increasing the connectivity and maximizing the channel availability* using wavelength division multiplexing (WDM), time division multiplexing (TDM), and space division multiplexing (SDM) techniques that result in further increasing the memory bandwidth;
2) by using a *decentralized wavelength allocation and wavelength reuse schemes* such that any node can reach any other node with a maximum of 1–2 hops for very large network sizes and thereby provide sufficient bandwidth per processor for the cost per processor;
3) by *implementing an innovative media access protocol* that lowers the waiting/queueing time for packet transmission and by implementing efficient multicast and broadcast functionality, which will help reduce the part of memory latency associated with the implementation of synchronization operations.

The objective of this paper is to design DSM systems using optical interconnects that supports scalable bandwidth and low latency without large overhead in hardware costs and results in a significant reduction in remote memory latency while allowing the system to scale to a large number of processors.

B. Related Work

In the SPEED architecture [18], write requests are broadcast using the snooping protocol and read requests are unicast using the directory protocol. The I-SPEED coherence protocol used for this architecture implement a single owner for dirty blocks to preserve the consistency of caches. SPEED uses a star coupler that can result in significant losses in the system as the number of nodes increases. Lightning network [19] uses directory cache coherence protocols in which all transactions are completed in a single hop and is constructed as a tree configuration with a wavelength partitioned at each level of the tree. The media access protocol in Lightning, called FatMac [20], requires all nodes to broadcast for channel allocation. We have adopted the token-based allocation, which is decentralized without requiring broadcast mechanism for channel allocation.

II. Architectural Overview

In this section, we describe and explain the design of RAPID architecture. A RAPID network is defined by a 3-tuple (P, D, G) where G is the total number of groups, D is the total number of node per group, and P is the number of processors per node. In this paper, we assume $P = 1$ for all network sizes, therefore we drop P; each node is identified as R(d, g) where $0 \leq g \leq G-1$; $0 \leq d \leq D-1$ such that $G \leq D$. This condition enables every group to communicate to every other group. The total number of processors in RAPID is the multiplicative factor $N = P \times D \times G$.

Fig. 1(a) and 1(b) shows the RAPID architecture. In Fig. 1(a), each node in RAPID network, contains the processor and its caches, a portion of the machine's physically distributed main memory, and a node controller (shown as a bus) which manages communication within nodes. Few nodes (0 up to $D-1$) are connected together to form a group. All nodes are connected to two subnetworks; a scalable intragroup interconnection (IGI) and a scalable intergroup remote interconnection (SIRI) via the intergroup passive couplers (IGPC). We have separated intragroup (local) and intergroup (remote) communications from one another to provide a more efficient implementation for both communications. Fig. 1(b) shows the conceptual diagram of RAPID network. Each group containing a few nodes on a system board is connected to SIRI using IGPC. All interconnections on the board are implemented using waveguide optics for shorter distances and the interconnections from the board to SIRI are implemented using fiber optics for longer distances. Fibers are chosen for longer distances as they can be extended to different lengths, when more nodes are added, as opposed to waveguides. All details regarding the RAPID network implementation are explained in Section III.

Fig. 2 shows the functional diagram of RAPID. As seen, the figure shows $D = 4$ (nodes) and $G = 4$ (groups). Each node is identified by R(d, g), with d as the node number and g as the group number. Within a group, all nodes are connected to multiplexers and demultiplexers for intra- and intergroup communication. For intergroup communication, all nodes are connected to SIRI via IGPC, the subscript indicates IGPC associated with

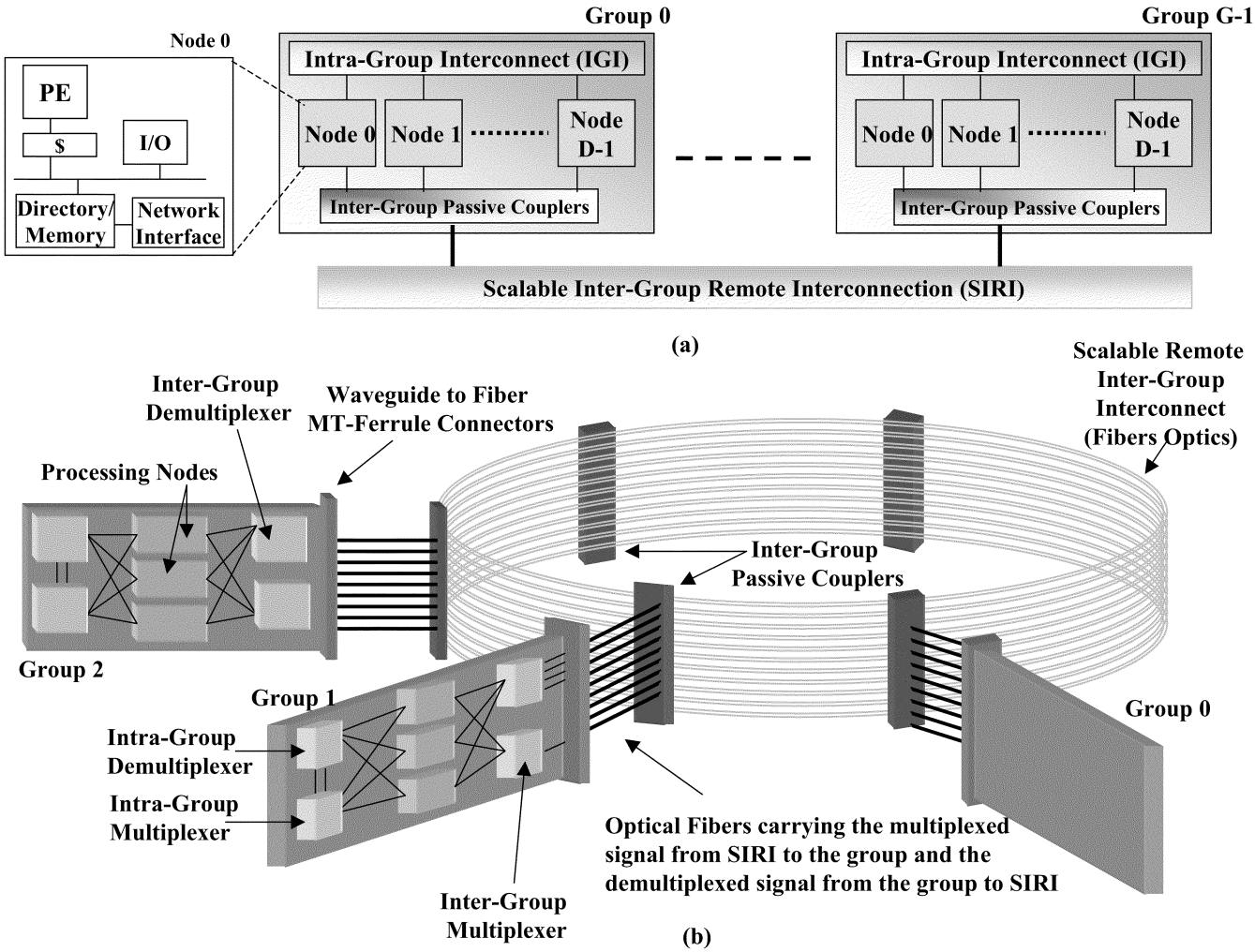

Fig. 1. (a) Architectural overview of RAPID. Every node is connected to two scalable subnetworks; a scalable local intragroup subnetwork and a scalable remote intergroup subnetwork. (b) Conceptual diagram of RAPID network. Several processing nodes are connected to intragroup and intergroup multiplexers and demultiplexers. All groups are connected via optical fibers and IGPC to the SIRI.

the group. 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.

A. Wavelength Assignment in RAPID

We propose a novel method based on wavelength reuse and spatial division multiplexing (SDM) techniques to design an efficient wavelength assignment strategy. The proposed methodology allows wavelengths to be reused when they are spatially separated, that is, when they are used at the local (intragroup) level or remote (intergroup) level. 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.

1) Wavelength Assignment for Intragroup Communication: The number of wavelengths employed for local communication equals the maximum number of nodes, D located in each group of the system. Fig. 3(a) shows an example of the intragroup wavelength assignment (of RAPID system shown in Fig. 2) and shows group 0. The wavelengths located next to each node correspond to the wavelength that each node receives on. This same wavelength assignment applies to all groups shown in Fig. 2. For example, R(1,0) to transmit to R(3,0) in group 0, R(1,0) would simply transmit on the wavelength assigned to node R(3,0) (e.g., λ_3). Similarly from Fig. 2, for node R(0,1) to transmit to R(3,1) in group 1, node R(0,1) would transmit on the wavelength assigned to node R(3,1), i.e., λ_3. Therefore, distinct wavelength allocation in different groups is possible by assigning an unique wavelength to every node at which it can receive optical packet from other intragroup nodes.

2) Wavelength Assignment for Intergroup Communication: In our remote wavelength assignment scheme shown in Fig. 2, all nodes within the source group is assigned a unique wavelength at which it can transmit to communicate with any destination group. We consider anti-clockwise as the direction of propagation on the scalable intergroup interconnect. Remote wavelengths are indicated by $\lambda_j^{(i)}$, where j is the wavelength and i is the group number from which the wavelength originates. In Fig. 2, any node in group 2 can communicate with group 3 on $\lambda_3^{(2)}$, any node in group 2 can communicate with group 0 on $\lambda_2^{(2)}$ and any node in group 2 can communicate with group 1 on $\lambda_1^{(2)}$. A cyclic wavelength allocation scheme is used and is

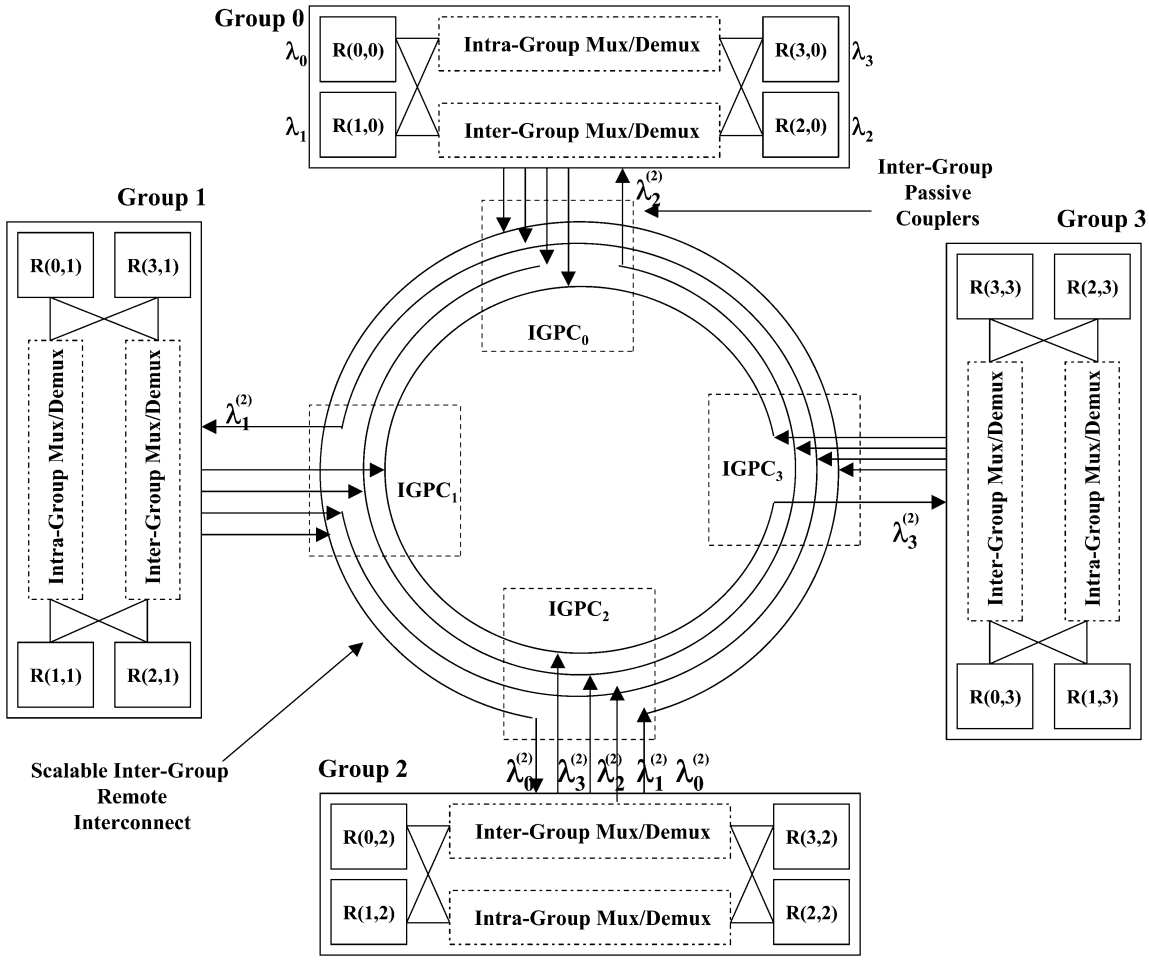

Fig. 2. Functional diagram of RAPID network. The figure shows $D = 4$ (nodes) and $G = 4$ (groups). Each node is identified by R(d, g), d as the node number and g as the group number. In addition, each node is shown with actual number for clarity. For example, node 4 in group 1 is identified as R(0,1). Within a group, all nodes are connected to multiplexers and demultiplexers for intra- and intergroup communication.

TABLE I
WAVELENGTH PRE-ALLOCATED FOR DIFFERENT SOURCE GROUPS (SG) AND DESTINATION GROUPS (DG)

	DG 0	DG 1	DG 2	DG 3	...	DG (G-2)	DG (G-1)
SG 0	λ_0	λ_{G-1}	λ_{G-2}	λ_{G-3}	...	λ_2	λ_1
SG 1	λ_1	λ_0	λ_{G-1}	λ_{G-2}	...	λ_3	λ_2
SG 2	λ_2	λ_1	λ_0	λ_{G-1}	...	λ_4	λ_3
SG 3	λ_3	λ_2	λ_1	λ_0	...	λ_5	λ_4
..	..	..	..	..	..	..	..
SG (G-2)	λ_{G-2}	λ_{G-3}	λ_{G-4}	λ_{G-5}	...	λ_0	λ_{G-1}
SG (G-1)	λ_{G-1}	λ_{G-2}	λ_{G-3}	λ_{G-4}	...	λ_1	λ_0

shown in Table I. The SG are the source groups and DG are the destination groups. Every destination group receives the same set of wavelengths $(\lambda_0 \ldots \lambda_{(G-1)})$ from various source groups. However, different source groups transmit wavelengths that are shifted by 1 wavelength to different destination groups. Note here that, the wavelength λ_0 is the wavelength at which every group communicates with itself. This wavelength is used to multicast transaction requests to all nodes within a group. 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. The destination nodes are fixed for every intergroup communication, i.e., for remote communication with group 2 as the destination, node R(0,2) always receives data on λ_1, R(1,2) always receives data on λ_2 and node R(2,2) always receives data on λ_3 and so on. Generalizing, $\forall$ g, the destination node within g, for wavelengths λ_i, is node (i,g). This gives us the criteria, that the there should exist at least D nodes within a group to receive data from G groups, i.e., $D \leq G$. In RAPID, it is possible to have $D > G$, as there will always be a destination node to receive signals from other groups. In RAPID, the wavelengths that are used for local communication are completely reused for remote communication which enables scalability.

B. Message Routing in RAPID

1) One-to-One Intragroup Communication: Local communication takes place when both the source and destination nodes are in the same group, $\mathrm{R}(j,g)_{\mathrm{source}} = \mathrm{R}(k,g)_{\mathrm{destination}}$. The source node tunes its transmitter to the preassigned wavelength of the destination node and transmits using waveguides.

2) One-to-One Intergroup Communication: Remote (intergroup) communication takes place when both the source and destination nodes are on different groups, $\mathrm{R}(j,g)_{\mathrm{source}}$ and $\mathrm{R}(k,m)_{\mathrm{destination}}$. Now, node $\mathrm{R}(j,g)$ can transmit the packet on a specific wavelength to group m. The destination node in group m which can receive the packet from group g may not be node k (the intended destination). To illustrate this, consider Fig. 2. Let the source node be R(1,1) and the destination node be R(0,3). The source node can transmit to group 3 on wavelength λ_2. The destination node which receives packets for remote communication in group 3 on wavelength λ_2 is R(2,3). So, node R(1,1) transmits on λ_3 and the packet is received by node R(2,3). R(2,3) then uses the local group interconnection to forward the packet to node R(0,3) on wavelength λ_0. Therefore, a single optoelectronic (O/E) conversion takes place at node R(2,3). In some cases source node $\mathrm{R}(j,g)$ may directly transmit to destination node $\mathrm{R}(k,m)$. RAPID requires a maximum of two optoelectronic conversion, one at the intermediate node and another at the final destination, to implement complete connectivity for any network size. This is possible as the wavelength assignment algorithm designed for remote group permits high connectivity.

3) Multicast and Broadcast Communications: We discuss how multicast communication on a given group is possible in RAPID. There are two cases, 1) when the source node is located within group, $\mathrm{R}(d,g)_{\mathrm{source}} = \mathrm{R}(g)_{\mathrm{destination}}$ and 2) when the source node is located outside the group, $\mathrm{R}(d,g)_{\mathrm{source}} \neq R(g)_{\mathrm{destination}}$. We use wavelength λ_0 for multicast communication. Considering the first case, the source nodes transmits the packet on wavelength λ_0. This packet is routed back to the same group and is broadcast to all nodes within the group. To illustrate with an example, consider node R(0,0) wants to send a multicast message to group 0. It transmits the packet on λ_0 which is routed back to group 0 and is received by all nodes in group 0. Considering the second case, the source node uses the previously mentioned remote group communication pattern and transmits the multicast packet to a particular destination node within the group. Now, the destination node within the group transmits the packet on λ_0 which is routed back to all nodes within the group. To illustrate with an example, consider node R(1,1) that sends a multicast packet to group 3. It transmits on λ_2 and the destination is node R(2,3) on group 3. R(2,3) then retransmits the packet on λ_0 which reaches all nodes within group 3. Similarly, broadcast communication is possible in RAPID by extending the multicast routing algorithm. The source node will transmit the multicast messages to all $(G-1)$ destination groups. The specific destination nodes will then retransmit the request on λ_0 such that all nodes within its group receive the message. The source node also transmits on λ_0 to send the multicast message to all nodes within its own group. More details regarding the media access protocol in RAPID can be found in [21].

III. Optical Implementation

In this section, we explain the optical components needed for the implementation and integration of the proposed network architecture using current CMOS technology. The key components of the proposed architecture are multiwavelength vertical cavity surface emitting lasers (VCSELs), photodetectors, waveguides/fibers, directional couplers, multiplexers, and demultiplexers.

A. Optical Component Specifications

Laser Sources: VCSELs are a natural candidate as laser sources in the proposed architecture, owing to their ease of fabrication in one- and two-dimension arrays, high power, good optical coupling to fibers, and low cost. Multiwavelength transmitters require either a tunable laser or an array of fixed wavelength lasers. Multiple wavelength VCSEL arrays could be the design choice for laser sources both for inter- and intragroup communications [22]. Multiwavelength VCSEL array consisting up to 16 channels having a maximum wavelength span of 46 nm, emitting at 1.1–1.2 μm and a wavelength spacing of 0.7 nm have been reported [12]. We will consider a similar VCSEL array for our proposed architecture.

Waveguides/Fiber Ribbons/Couplers: Optical polymers are increasingly considered as highly versatile elements that can be readily transformed into single-mode, multimode, and microoptical waveguide/fiber structures [23]. The multiplexer in our proposed architecture is designed as a tree of 1×2 couplers constructed using optical polymers to comine the various signals from different modes [24].

Demultiplexers: The key components of the WDM systems are demultiplexers, for separating the various signals. Integrated optic demultiplexers have been either grating-based or phased array based devices (also called arrayed waveguide gratings) [13], [14], [25]. In phased-array based devices the focusing and dispersive properties required for demultiplexing are provided by an array of waveguides, the length of which has been chosen such as to obtain the required imaging and dispersive properties. The length of the array waveguides is chosen such that the optical path length difference $\bigtriangledown L$ between adjacent waveguides equals an integer multiple of the central wavelength of the demultiplexer [14], thus attaining a phase difference at the waveguide exit and is given by

$$n_c \bigtriangledown L = m\lambda_0 \tag{1}$$

where n_c is the effective refractive index of the arrayed waveguide, m is the diffraction order and λ_0 is the center wavelength. The most important parameter of the wavelength multiplexer is the channel spacing $\bigtriangledown\lambda_0$ and is obtained by

$$\bigtriangledown\lambda_0 = \frac{\bigtriangledown x}{f}\frac{n_s d}{m}\left(\frac{n_g}{n_c}\right)^{-1} \tag{2}$$

where $\bigtriangledown x$ is the spacing of the output waveguide, n_s is effective refractive index of the slab waveguide, d is the pitch of the arrayed waveguide, n_g is the group index. $\bigtriangledown x$ and d values should be small to realize a narrow wavelength spacing and at the same time it is necessary that the waveguides be sufficiently separated

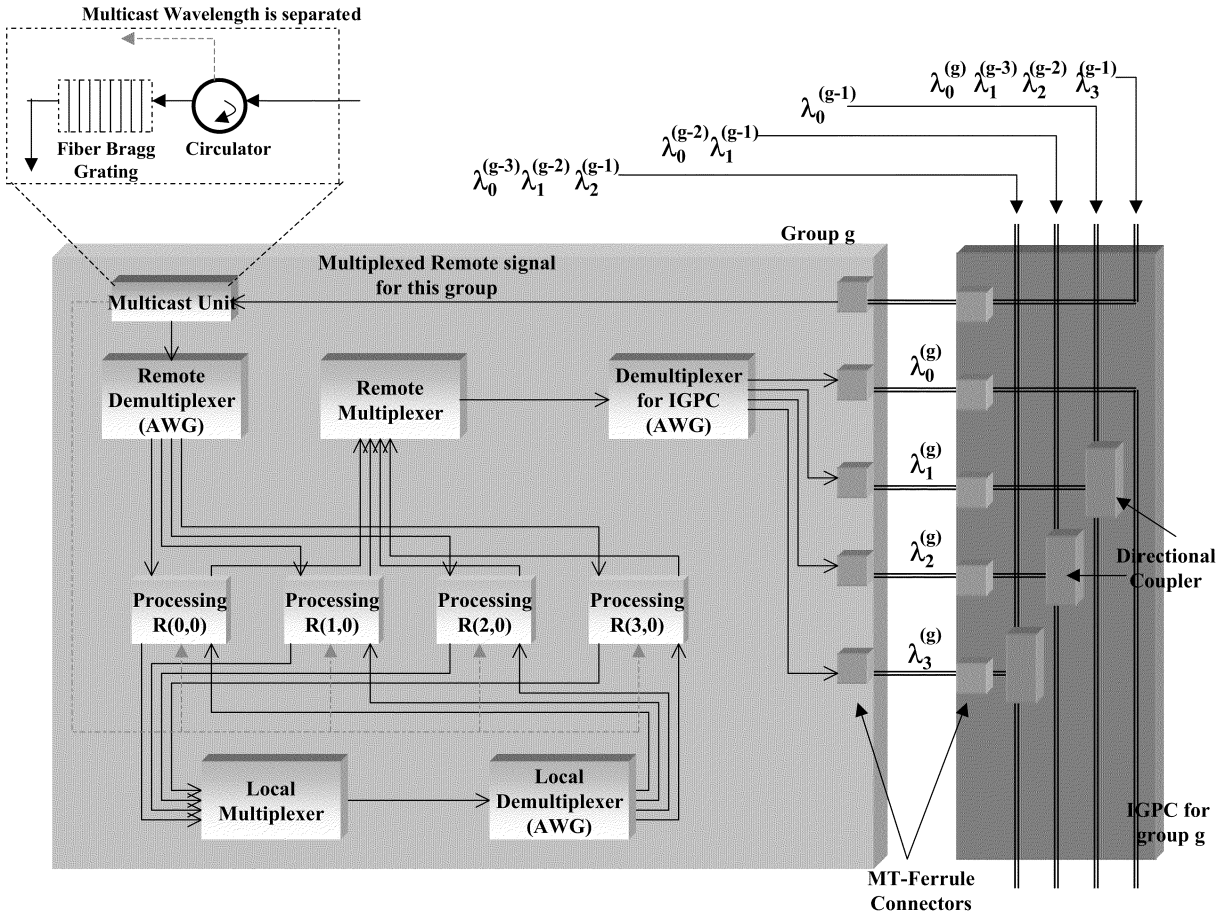

Fig. 3. Possible implementation of group g. Intragroup communication is possible by a combination of local multiplexer, directional coupler array and the local demultiplexer. For intergroup communication, the processing nodes are connected to the remote multiplexer, directional couplers, and IGPC demultiplexers to the IGPC as shown. The demultiplexed signal is first passed through a multicast unit and then to the remote demultiplexer.

for no interaction. Free spectral range (FSR) is the spacing between the center wavelengths and orders from (1) and is given by

$$\mathrm{FSR} = \frac{\lambda_0}{m}\left(\frac{n_g}{n_c}\right)^{-1}. \tag{3}$$

The maximum number of wavelength channels M depends on the FSR. The bandwidth of the multiplexed light, that is $M \bigtriangledown \lambda_0$ must be narrower than the FSR to prevent the overlapping of orders in the spectral region. Accordingly

$$M < \frac{\lambda_0}{m \bigtriangledown \lambda}\left(\frac{n_g}{n_c}\right)^{-1}. \tag{4}$$

The VCSEL array chosen for the proposed architecture has a wavelength span of over 46 nm and a channel spacing $\bigtriangledown \lambda_0$ of 0.7 nm [12]. The center wavelength, λ_0 is 1119 nm, choosing $\bigtriangledown x$ and $n_c = 1.47$, $n_g = 1.497$, and $n_s = 1.539$, and from (2) we can calculate $f \times m$ to be 0.092. Choosing suitable values for $f = 2.1$ mm and $m = 42$, and from (3), we can determine the FSR to be 25.44 nm. From (4), the maximum number of channels, M should be eight channels. The length $\bigtriangledown L$ is determined from (1) to be 31.97 μm. The number of waveguides N is not a dominant parameter and we assume that 201 waveguides are required [14].

B. Optical Integration Methodology

Optical interconnects based on complimentary-metal-oxide-semiconductor CMOS/VCSEL technologies have been widely proposed for high-performance computing applications [16], [17]. The approach followed in our design is the most widely used hybrid integration using flip-chip bonding of OE-VLSI components. The VCSEL/PD arrays can be fabricated on a GaAs substrate such that the devices are designed to be back-side emitting because of the desire to flip-chip bond them to CMOS driver circuits. The n-contact and p-contact should then be on the top surface of the wafer to facilitate electrical connectivity with CMOS circuits. GaAs substrate can then be selectively etched leaving the VCSEL/PD contact pad on the backside of the wafer and all optical sources/detectors on the other side of the wafer. The VCSELs and PDs can now be integrated onto the CMOS driver using flip-chip bonding and substrate removal techniques [11], [16].

In Fig. 3, we propose a possible optical implementation of the DSM architecture, which could be constructed directly onto the PC board. Each PC board is a group, containing a few processing nodes as shown. The nodes are connected to the intragroup (local) multiplexer and the multiplexed output is coupled using 1 $\times$ 2 couplers to the local demultiplexer. The demultiplexer used in our proposed architecture is the low-loss arrayed waveguide

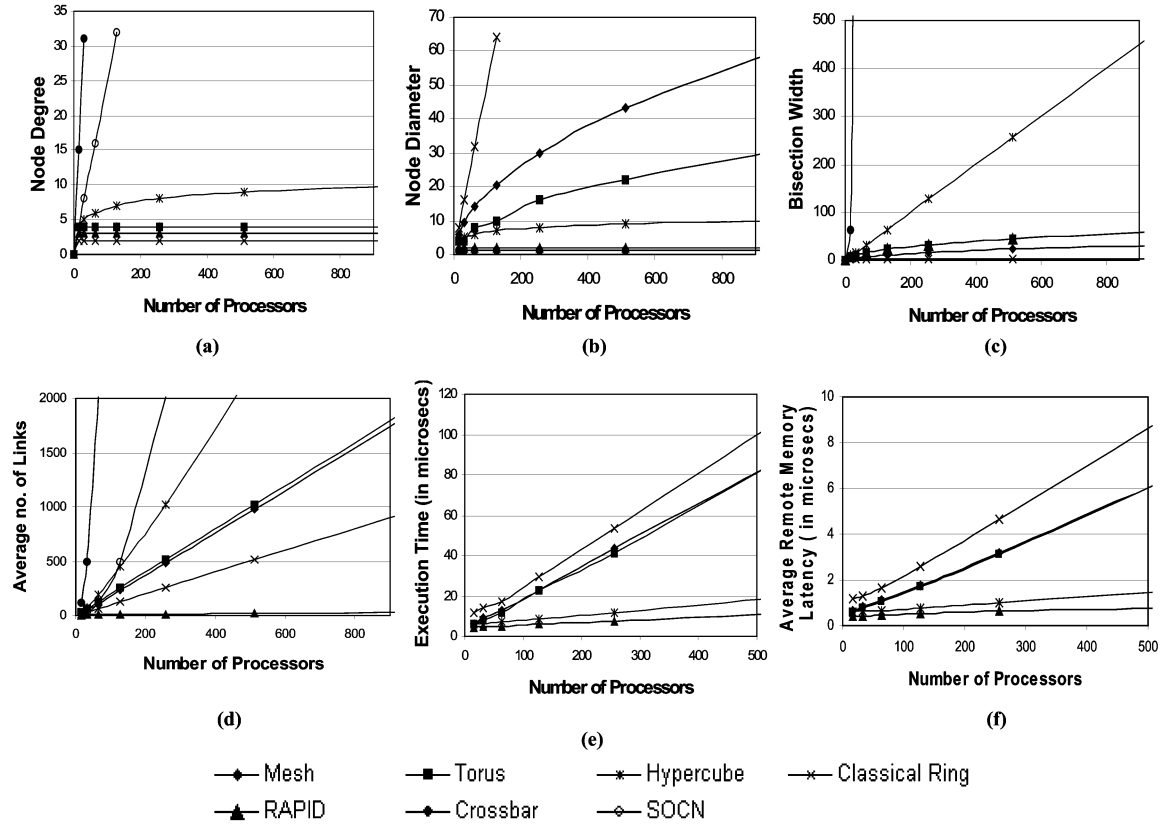

Fig. 4. (a) Degree comparison. (b) Diameter comparison. (c) Bisection width comparison for varying number of processors. (d) Number of links comparison. (e) Execution time for different simulated electrical and RAPID networks for varying number of processors. (f) Average remote memory latency for varying number of processors for RAPID and other electrical networks.

grating (AWG) that can be integrated using planar waveguide technology. As each node will receive the signal specified on a given wavelength, the demultiplexed output from the local demultiplexer is sent to the respective node using waveguides. For intergroup (remote) 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 intergroup interconnects. The merging of signals should ensure that different wavelengths are combined to separate fibers. 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 $\lambda_1^{(g-1)}$. These signals originated from previous groups $(g-2)$ and $(g-1)$, respectively. The wavelength $\lambda_0^{(g)}$ does not merge with any existing fiber, but creates a new fiber channel to which successive groups will merge different wavelengths. This 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 and forms the input to the multicast unit, shown in the inset in Fig. 3. This unit consists of an optical circulator and fiber Bragg grating and is used exclusively for separating the broadcast wavelength. By tuning the Bragg grating to wavelength λ_0 we can drop this multicast wavelength from the multiplexed signal. λ_0 is then broadcast to all the nodes within the group. The remaining signals can then be demultiplexed and returned to the respective node.

IV. Performance Evaluation

In this section, we evaluate the performance of RAPID for DSMs by analyzing the network characteristics, scalability based on power budget, and BER criteria and performance based on simulation.

A. Comparison With Other Popular Networks Based on Network Characteristics

The scalability of RAPID architecture with respect to several network parameters is discussed. RAPID R(d, g) is compared with several well-known network topologies such as a traditional crossbar network (CB), the Binary Hypercube, the Ring network, the Torus, two–dimensional (2-D) Mesh, and Scalable Optical Crossbar Network (SOCN) [24]. Each of these networks will be compared with respect to degree, diameter, number of links and bisection width. Fig. 4(a) shows a comparison of the node degree of various networks with respect to system size (number of nodes). For RAPID network, the node degree remains constant for any network size, i.e., even for 1000-node network, and each node needs to be connected only to IGI (local communication), to IGPC (remote communication) and to the multicast channel. Fig. 4(b) shows a comparison of the diameter of various networks with respect to system size. In RAPID, to support better connectivity using limited wavelength, a diameter of 2 is achieved for any network size. This is comparable to other less scalable networks such as the crossbar and better than other scalable networks such as the Torus and the Hypercube. Fig. 4(c) shows the plot of the bisection width of various

network architectures with respect to the number of processors in the system. The crossbar and the hypercube networks provide much better bisection width than RAPID network. Yet, the bisection width of RAPID network is very comparable to the best of the remaining networks. Fig. 4(d) shows the plot of the number of network links with respect to the number of processors in the system. RAPID shows the least cost for intergroup communication, thereby showing a much better scalability in the number of links for very large-scale systems.

B. Power Budget and BER Estimation

Calculation of a power budget and the signal-to-noise ratio (SNR) at the receiver is important for confirming the realizability and scalability of any optical interconnect implementation. The SNR at the receiver gives an indication of the expected bit-error rate (BER) of the digital data stream. For a parallel computing interconnect, the required BER may be as low as 10^{-15}. For such a BER, we computed that the received power should be 9.487 μW or -50.2284 dB or -20 dBm [26]. High-powered VCSEL arrays delivering output power as high as 2 mW or -26.989 dB have been reported [12]. The total optical loss in the system is the sum total of the losses (in decibels) of all optical components that a beam must pass through from the transmitter (VCSEL array) to the receiver (photodetector).

We first calculate the losses in the system for intragroup interconnections. The various losses are VCSEL-waveguide coupling (-0.2 db), propagation in the waveguide/fiber (-0.5 dB), arrayed waveguide grating (AWG) (-2.1 db $\times$ 2) for demultiplexing, 1×2 coupler array (-3 dB $\times \log_2(d)$). The loss in an AWG up to 32 channels with 1 nm (100 GHz) channel spacing can be as low as 2.1 dB [25]. The local loss has a strong dependence on the fixed coupler loss of -3 dB times the number of local intragroup nodes D. Based on the total loss, we can have a maximum of 128 nodes. However, a practical system can only support four nodes because a typical board is 1031 cm^2 area [3]. For remote intergroup one-to-one communication, all the losses for intragroup interconnection will also be present. The other losses are an additional AWG (-2.1 db) G directional couplers at the IGPC $-0.225 \times G$, circulator (-0.5 dB), fiber Bragg grating (-0.5 db), waveguide-to-fiber connector (-0.5 dB), and additional propagation loss in the fiber/waveguide (-0.5 dB). Note, that the total remote losses are added to the total local losses to give -10.1 dB -0.225 dB $\times \log_2(G) - 3$ dB $\times G$. Now, with $G = D$, the number of groups that can be connected is approximately 16. This implies that RAPID can scale up to $256 (= 16 \times 16)$ nodes, though current board sizes limit the number of nodes per board to 4, and the system size to 16 nodes. RAPID uses only passive technology for routing between sources and detectors such as gratings (AWG, fiber Bragg, 1×2 couplers, and waveguide/fiber optics. The use of passive technology are twofold: 1) the optical signal transfer is much faster since there is no optical switching or conversion, and 2) the cost of constructing the architecture reduces considerably.

C. Simulation Methodology and Assumptions

In this section, we describe the simulation methodology and the results obtained by comparing RAPID with few scalable electrical networks such as the 2-D Mesh, 2-D Torus, Hypercube, and the classical ring. We use CSIM [27], a process-oriented, discrete-event model simulator to evaluate the performance of RAPID network using synthetic traffic workloads. In this simulation, we model accurately contention at all resources for both electrical and optical networks. In this simulation, we model accurately contention at all resources for both electrical and optical networks. Each node of the simulated network contains 1 GHz processor and 8 MSHRs (miss status holding registers). L1 cache hit time is 1 cycle, L2 cache hit time is 15 cycles, cache-to-cache transfer is 25 cycles, and memory access time is 70 cycles.

In our simulated model, a processor generates a maximum of N_{requests} memory requests at an average rate of P_{traffic} (Poisson distributed) requests per cycle. The caches in our model use miss status holding registers (MSHRs) [1] to track the status of all outstanding requests. If no MSHR is available when the processor generates a request, then the processor is blocked from sending requests until the next clock cycle after a reply arrives that frees the MSHR. The generated request is satisfied at the caches with a probability of P_{L1} (at L1) and a probability of P_{L2}) (at L2). This request reaches the directory and memory module of the concerned node with a probability of $[1 - (P_{L1} + P_{L2})]$. With a probability of P_{nohop}, the request is locally satisfied and with a probability of $(1 - P_{\text{nohop}})$, this request is considered to be a remote memory request. In case of a clean block, for load/store miss, with a probability of P_{2hop}, the request is satisfied at the remote memory. In case of a dirty block for load miss, with a probability of P_{3hop}, the request is forwarded to the owner. Cache to cache transfer of the requested block takes place and the home node replies with the acknowledgment message to the requestor. In case of a store miss for a dirty block, the home node is responsible for collecting invalidations from N_{sharers} before acknowledging the request for exclusive permission. All the above simulation parameters were chosen from different technical manuscripts [28] and these parameters were consistent for both optical and electrical networks. Contention is modeled at all system resources; MSHRs, directory/memory modules, network interfaces, virtual channels (in case of electrical networks) and optical tokens (in case of RAPID).

For the electrical network, wormhole routing is modeled with a fit size of 8 B and up to four virtual channels per link. Nondata size message size is 16 B, data size messages are 64 B, router speed is 500 MHz, router's internal bus width is 64 B, and channel speed is 10 GHz. Various routing, switching and propagation times [28] 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 [12]. We model O/E (optical to electrical) and E/O (electrical to optical) delays of 12.8 ns ($=$ 16 b $\times$ 8/10 GHz). The optical packet can be processed as soon as the header is received, thereby reducing the latency. The optical token consists of 16 b, under the assumption that the maximum $4 * 4$ (16 nodes) are present. The first 4 b are used as local token bits, one per node and the remaining 4 b are used for global tokens, one per group. The token passing latency is completely overlapped with the packet transmission latency, i.e., a

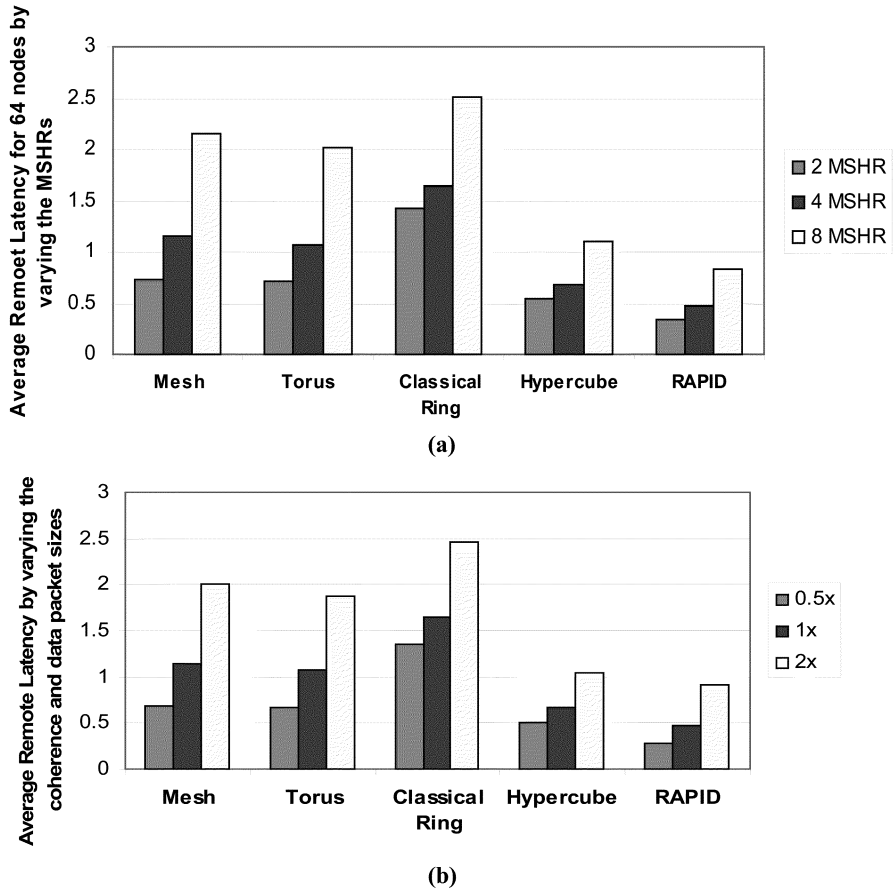

Fig. 5. (a) Average remote latency when the number of MSHRs is equal to 2, 4, and 8 for RAPID and other electrical networks for 64 nodes. (b) Average remote memory latency when the coherence and the data block size is doubled for 64 nodes for various networks.

node that begins transmission on a specific wavelength can immediately transmit the token to the next node.

D. Simulation Results

Execution Time: Fig. 4(e) shows the execution time for varying number of nodes for both the simulated electrical and optical networks. RAPID outperforms all networks by maximizing the channel availability and maintaining a low diameter for large number of nodes. RAPID outperforms the classical ring by almost 89% for 512 nodes. This can be attributed to the large increase in network diameter for the ring network $(N/2)$. The mesh and torus have similar latencies, with RAPID performing them by almost 86% for 512 nodes. The hypercube performs reasonably well, though RAPID outperforms hypercube by almost 38%. All electrical networks showed different latencies depending on how many switches needed to be traversed.

Average Memory Latency: Fig. 4(f) shows the average remote memory access latency. RAPID performed the best as compared to all other networks. RAPID outperformed hypercube by 46%, the mesh torus by 87%, and the classical ring by 91%. RAPID, even though undergoes a single optoelectronic conversion, provides better utilization of the bandwidth by distributing the load on various wavelengths. Therefore, the average remote memory latency increases less than linearly for RAPID.

Effects of Varying MSHRs: Fig. 5(a) shows the average remote memory access latency for various network topologies consisting of 64 nodes. The MSHRs handle all the outstanding requests from the node. If all MSHR are occupied, no new request will be injected into the network and this prevents the nodes from flooding the network. When the MSHR is equal to 2 or 4, all networks perform reasonably well. When the number of MSHR is increased to 8, we see that the average latency almost doubles for mesh and the torus. This is due to the increased contention as more requests are being injected into the network. RAPID performs reasonably well, the increase is around 57% and is comparable to hypercube (60%).

Effects of Varying the Coherence and Data Packet Sizes: Fig. 5(b) shows the effects of varying the coherence and data packet sizes on the average remote memory latency for 64 nodes. 1x refers to 16 B coherence packet size and 64 B data packet. 0.5x refers to 8 B and 32 B coherence and data packet, respectively. RAPID is sensitive to packet size changes, as seen from the plot.

V. Conclusion

In this paper, we developed an optically interconnected architecture called RAPID to reduce the remote memory access latency in distributed shared memory multiprocessors. RAPID was completely conceived and developed using passive optical technology for routing between sources and detectors making

the proposed architecture much faster and inexpensive as compared to other optical and electrical architectures. RAPID, not only maximizes the channel availability for intergroup communication, but at the same time wavelengths are completely reused for both intragroup and intergroup 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. This network architecture provides distinct performance and cost advantages over traditional electrical interconnects and even over other optical networks.

REFERENCES

[1] D. E. Culler, J. P. Singh, and A. Gupta, *Parallel Computer Architecture: A Hardware/Software Approach*. San Fransisco, CA: Morgan Kaufmann, 1999.

[2] D. E. Lenoski and W.-D. Weber, *Scalable Shared-Memory Multiprocessing*. San Fransisco, CA: Morgan Kaufmann, 1995.

[3] J. Laudon and D. Lenoski, "SGI origin: A ccNUMA highly scalable server," in *Proc. 24th Annual Int. Symp. Computer Architecture*, June 1997, pp. 241–251.

[4] K. Gharachorloo, D. Lenoski, J. Laudon, P. Gibbons, A. Gupta, and J. Hennessy, "Memory consistency and event ordering in scalable shared-memory multiprocessors," in *Proc. 17th Ann. Int. Symp. Computer Architecture*, May 1990, pp. 15–26.

[5] D. Dai and D. K. Panda, "How much does network contention affect distributed shared memory performance," in *Int. Conf. Parallel Processing (ICPP '97)*, 1997, pp. 454–461.

[6] ——, "How can we design better networks for DSM systems?," *Lecture Notes in Comp. Sci.*, vol. 1417, pp. 171–184, 1998.

[7] S. P. Dandamudi, "Reducing hot-spot contention in shared-memory multiprocessors," *IEEE Concurrency*, vol. 7, pp. 48–59, 1999.

[8] D. A. B. Miller, "Rationale and challenges for optical interconnects to electronic chips," *Proc. IEEE*, vol. 88, pp. 728–749, June 2000.

[9] 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," *Appl. Opt., Special Issue on Optics in Computing*, vol. 39, pp. 671–682, 2000.

[10] R. Pu, E. M. Hayes, C. W. Wilmsen, K. D. Ohoquette, H. Q. Hou, and K. M. Geib, "Comparison of techniques for bonding VCSEL's directly to ICs," *J. Opt. Soc. Amer.*, vol. 1, pp. 324–329, 1999.

[11] H. J. J. Yeh and J. S. Smith, "Integration of GaAs vertical cavity surface emitting laser on Si by substrate removal," *Appl. Phys. Lett.*, vol. 64, pp. 1466–1468, 1994.

[12] M. Arai, T. Kondo, A. Matsutani, T. Miyamoto, and F. Koyama, "Growth of highly strained GaInAs-GaAs quantum wells on patterned substrate and its application for multiple-wavelength vertical-cavity surface-emitting laser array," *IEEE J. Select. Topics Quantum Electron.*, vol. 8, pp. 811–816, July/Aug. 2002.

[13] M. K. Smit and C. Van Dam, "PHASAR-based WDM-devices: Principles, design and applications," *IEEE J. Select. Topics Quantum Electron.*, vol. 2, pp. 236–250, June 1996.

[14] H. Takahashi, S. Suzuki, and I. Nishi, "Wavelength multiplexer based on SiO2-Ta2O5 arrayed-waveguide grating," *J. Lightwave Technol.*, vol. 12, pp. 989–995, June 1994.

[15] R. Ramaswami and K. Sivarajan, *Optical Networks: A Practical Perspective*. San Fransisco, CA: Morgan Kaufmann, 2001.

[16] A. V. Krishnamoorthy, K. W. Goossen, L. M. F. Chirovsky, R. G. Rozier, P. Chandramani, S. P. Hui, J. Lopata, J. A. Walker, and L. A. D'Asaro, "16 × 16 VCSEL array flip-chip bonded to CMOS VLSI circuit," *IEEE Photon. Technol. Lett.*, vol. 12, pp. 1073–1075, Aug. 2000.

[17] Y. Liu, "Heterogeneous integration of OE arrays with Si electronics and micro-optics," in *Proc. Electron. Comp. Technol. Conf.*, 2001, pp. 864–869.

[18] J.-H. Ha and T. M. Pinkston, "The speed cache coherence for an optical multi-access interconnect architecture," in *Proc. 2nd Int. Conf. Massively Parallel Processing Using Optical Interconnections*, 1995, pp. 98–107.

[19] P. Dowd, J. Perreault, J. Chu, D. C. Hoffmeister, R. Minnich, D. Burns, F. Hady, Y. J. Chen, and M. Dagenais, "Lightning network and systems architecture," *J. Lightwave Technol.*, vol. 14, pp. 1371–1387, 1996.

[20] K. Bogineni and P. W. Dowd, "A collisionless multiple access protocol for wavelength division multiplexed star-coupled configuration: Architecture and performance analysis," *J. Lightwave Technol.*, vol. 10, pp. 1688–1699, 1992.

[21] A. Kodi and A. Louri, "A scalable architecture for distributed shared memory multiprocessors using optical interconnects," in *Proc. 18th Int. Parallel Distributed Processing Symp. (IPDPS'04)*, Sante Fe, NM, Apr. 2004, pp. 11–20.

[22] R. R. Patel, S. W. Bond, M. D. Pocha, M. C. Larson, H. E. Garrett, F. Drayton, H. E. Petersen, D. M. Krol, R. J. Deri, and M. E. Lowry, "Multiwavelength parallel optical interconnects for massively parallel processing," *IEEE J. Select. Topics Quantum Electron.*, vol. 9, pp. 657–666, Mar./Apr. 2003.

[23] L. Eldada and L. W. Shacklette, "Advances in polymer integrated optics," *IEEE J. Select. Topics Quantum Electron.*, vol. 6, pp. 54–68, 2000.

[24] B. Webb and A. Louri, "A class of highly scalable optical crossbar-connected interconnection networks (SOCN's) for parallel computing systems," *IEEE Trans. Parallel Distrib. Syst.*, vol. 11, no. 1, pp. 444–458, May 2000.

[25] Y. Hibino, "An array of photonic filtering advantages: Arrayed waveguide-grating multi/demultiplexers for photonic networks," *IEEE LEOS Newslett.*, Aug. 2001.

[26] T. V. Moui, "Receiver design for high-speed optical fiber systems," *J. Lightwave Technol.*, vol. LT-2, pp. 234–267, 1984.

[27] H. Schwetman, "CSIM19: a powerful tool for building system models," in *Proc. 2001 Winter Simulation Conf.*, 2001, pp. 250–255.

[28] M. E. Acacio, J. Gonzalez, J. M. Garcia, and J. Duato, "The use of prediction for accelerating upgrade misses in cc-NUMA multiprocessors," in *Proc. Int. Conf. Parallel Architectures and Compilation Techniques*, 2002, pp. 155–164.

Avinash Karanth Kodi (S'04), photograph and biography not available at the time of publication.

Ahmed Louri (S'86–M'88–SM'95), photograph and biography not available at the time of publication.