A Scalable Architecture for Distributed Shared Memory Multiprocessors using Optical Interconnects

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

Abstract

In this paper, we describe the design and analysis of a scalable architecture suitable for large-scale DSMs (Distributed Shared Memory) systems. The approach is based on an interconnect technology which combines optical components and a novel architecture design. In DSM systems, 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 RAPID (Reconfigurable and scalable All-Photonic Interconnect for Distributed-shared memory), we provide high connectivity by maximizing the channel availability for remote communication to reduce the remote memory access latency. RAPID also provides fast and efficient unicast, multicast and broadcast capabilities using a combination of aggressively designed wavelength, time and space-division multiplexing techniques. We evaluated RAPID based on network characteristics, power budget criteria and simulation using synthetic traffic workloads and compared it against other scalable electrical networks. We found that RAPID, not only outperforms other networks, but also, satisfies most of the requirements of shared memory multiprocessor 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. The two most common types of shared-memory models are snooping and directory based multiprocessors. Snooping cache coherence protocols are successful because they obtain data quickly (without indirection) by broadcasting coherence transactions to all processors and memory in the system[3]. Nevertheless, snoop bandwidth limitations, and the need to act upon all transactions at every processor, make snooping designs challenging especially in the light of aggressive processors and limits the snooping-based multiprocessors to smaller system configurations. In contrast, directory-based shared memory multiprocessors which depend on maintaining the identity of sharers (at the directory) to avoid the need for broadcasts, are much better suited for larger designs. A state of the art example is the SGI Origin 2000[1] that implements sequential consistency and can scale to several hundreds of nodes (512).

The increasing performance gap between processors and memory systems imposes a memory bottleneck that is intensified in distributed shared-memory multiprocessors by contention and cache coherency. Directory protocols transmit a coherence transaction over an arbitrary point-to-point network to a directory entry which, in turn re-directs the transaction to a superset of processors caching the block. Several processor cycles maybe required in order to complete a remote memory transaction in DSM systems. To address this issue, modern microprocessors are capable of issuing many instructions per cycle, and are able to tolerate large memory access latencies. Techniques such as lock-free caches[4], hardware and software prefetching[5], speculative loads[6] and multiple outstanding requests[2] are significantly reducing cache misses, and consequently memory access latency. However, these successful and efficient latency-tolerating techniques require much more bandwidth, and create much more memory traffic and contention in the network system. To minimize network contention and to avoid deadlock or starvation, several innovative switching techniques such as cut-through routing, wormhole routing, adaptive routing and use of virtual chan-

Proceedings of the 18th International Parallel and Distributed Processing Symposium (IPDPS'04)

0-7695-2132-0/04/$17.00 (C) 2004 IEEE

nels have been implemented[7]. While, in wormhole routing, the communication latency is almost *distance insensitive* in the absence of contention, the amount of traffic generated by sequencing various requests, acknowledgements and data responses by multiple outstanding requests still causes contention in the network[7, 8, 9]. Due to the above considerations, it is hard to scale DSMs to a large number of nodes (1000 nodes) while maintaining reasonable performance levels across a wide variety of applications at a reasonable cost.

Commerical cc-NUMA systems improve the interconnection network design by using crossbars to minimize the network latency and maximize the usable bandwidth[10, 11]. Traditional electronic crossbars require O(N^2) switches and wires to implement a crossbar, which tend to limit the scalability. Yet, they are ideal for small to medium scale multiprocessors ranging from 32-64 processors. The SGI Origin 2000[1] employs a scalable approach by using local crossbars and SPIDER routers to create a bristled fat hypercube interconnection topology. In SGI Origin 2000, the latency in accessing remote memory increases substantially with size of the network. Origin 2000 also achieves a lower latency to close memories but a higher latency to memories which are far away. The restart latency for read transaction for an unowned cache block is $338nsec$ to local memory and $554nsec$ to the closer remote memory, and it increases by about $100nsec$ for each additional hop to the memory[1]. While several innovative advances are made to improve the range and extent of high-speed electronic channels[1, 3, 11], it is increasingly difficult to keep pace with the bandwidth, latency, connectivity and scalability requirements of modern distributed shared memory multiprocessor systems.

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[12, 13, 14, 15, 16, 17]. 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, arrayed waveguide grating, micro-optical components, etc) and packaging technologies (OE-VLSI heterogeneous integration, smart pixel technology) which make optical interconnects a viable and cost-effective option for building high bandwidth, low latency, and scalable optical interconnection networks.

1.1 Related Work

SPEED[16] and Lightning[14] are two distributed shared memory multiprocessor architectures proposed using optical interconnects. In the SPEED[16] architecture, write requests are broadcast using the snooping protocol and read requests are unicast using the directory protocol. SPEED uses a star coupler which can result in significant losses in the system. Lightning network[14] 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 partitioner at each level of the tree. The media access protocol in Lightning, called FatMac[14], requires all processors to broadcast for channel allocation. We have adopted the token based allocation[16] which is decentralized without requiring broadcast mechanism for channel allocation.

This paper proposes an integrated solution to reduce the remote memory access latency in DSMs and still be able to scale the network significantly using low-latency, high-bandwidth optical technology. For this purpose, an interconnect technology is used that combines optical components and a novel architecture design. Wavelength allocation scheme is designed so as to reduce the queuing time for remote requests significantly. Some of the benefits of RAPID are:
(1) RAPID provides high connectivity by *maximizing channel availability for remote communication* and this results in achieving low latency for remote memory transactions.
(2) RAPID is designed using passive optical components as opposed to active components, thereby making the network much faster and less expensive.
(3) RAPID uses *a decentralized wavelength allocation* strategy so as to maximize the channel availability. This is implemented by sharing wavelengths for remote communication by only locally connected processors, thereby reducing the queueing delays for channel availability.
(4) RAPID supports *unicast, multicast and broadcast communications* using a combination of aggressively designed WDM (wavelength division multiplexing), TDM (time division multiplexing) and SDM (space division multiplexing) techniques.
(5) RAPID is easily scalable by either adding processors to the local groups or by adding additional local groups to the system with minimum reconfiguration as explained below. Therefore, RAPID can be easily scaled at low cost to very large number of nodes.

2 Architecture Details

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) $\forall$ 1$\leq$ g $\leq$ G; 1 $\leq$ d $\leq$ D such that G $\leq$

0-7695-2132-0/04/$17.00 (C) 2004 IEEE

D. This condition enables every group to communicate to every other group.

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 (shown as a bus) which manages communication within nodes. Few nodes (0 up to D) 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 Remote Interconnection (SIRI) via the Inter-Group Passive Couplers (IGPC). We have separated intra-group (local) and inter-group (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. 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 and the interconnections from the board to SIRI are implemented using fiber optics.

Figure 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), d as the node number and g as the group number. For example, node 4 in group 1 is identified as R(0,1). Each node is identified by the node number (0-15) and also by the notation of node number and group number. Within a group, all nodes are connected to multiplexers and demultiplexers for intra- and inter-group communication. For inter-group communication, all nodes are connected to SIRI via IGPC, the subscript indicates IGPC associated with 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.

2.1 Wavelength Assignment in RAPID

We propose a novel method based on wavelength re-use and spatial division multiplexing (SDM) techniques to design an efficient wavelength assignment strategy. 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 or remote (inter-group) 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.

Wavelength Assignment for Intra-Group Communication: The number of wavelengths employed for local communication equals the maximum number of nodes, D located in each group of the system. Figure 2 shows the intra-group wavelength assignment for 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 figure 2. For example, for node 1, R(1,0) to transmit to node 3 in group 0, node 1 would simply transmit on the wavelength assigned to node 3 (e.g. λ_3). Similarly from figure 2, for node 4 to transmit to node 7 in group 1, node 4 would transmit on the wavelength assigned to node 7 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 intra-group nodes.

Wavelength Assignment for Inter-Group Communication: In our remote wavelength assignment scheme shown in figure 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 inter-group 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 figure 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 $\lambda_1^{(2)}$. Similarly, group 3 can communicate with group 0 on $\lambda_3^{(3)}$, group 1 on $\lambda_2^{(3)}$ and with group 2 with $\lambda_1^{(3)}$. For clarity, only the wavelengths received by all groups from group 2 is shown in fig 2. A cyclic wavelength allocation scheme is used and is shown in Table 1. The SG are the source groups and DG are the destination groups. Note here that, the wavelength λ_0 is the wavelength at which every group communicates with itself. 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 inter-group 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. In our example, shown in figure 2, there are 4 groups, so there should exist at least 4 nodes per group. The maximum number of wavelengths then required for either local (inter-group) or remote (intra-group) communication 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.

0-7695-2132-0/04/$17.00 (C) 2004 IEEE

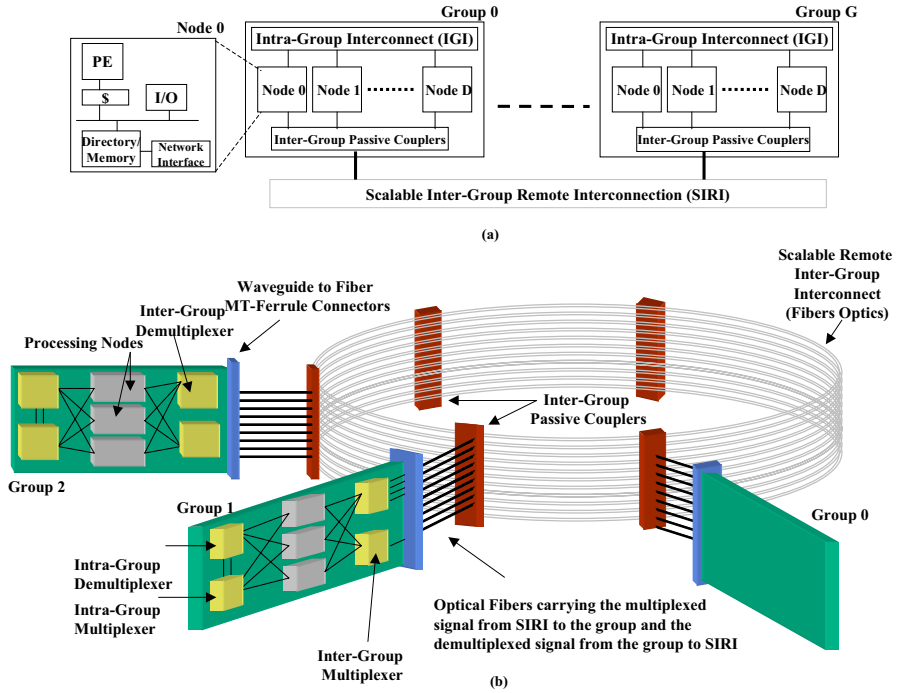

Figure 1. (a) shows the architectural overview of RAPID. Figure 1(b) shows the conceptual diagram of RAPID network.

Table 1. Wavelength Pre-Allocated for different source groups (SG) and destination groups (DG)

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

2.2 Message Routing in RAPID

One-to-one Intra-Group Communication: Local communication takes place when both the source and destination nodes are in the same group, R(j,g)_{source} = $\text{R(k,g)}_{destination}$. The source node tunes its transmitter to the pre-assigned wavelength of the destination node and transmits. A logical channel is established and mapped onto the physical fiber and a diameter of one is achieved for local communication.

One-to-one Inter-Group Communication: Remote (inter-group) communication takes place when both the source and destination nodes are on different groups, $\text{R(j,g)}_{source} \neq R(k,m)_{destination}$. Now, node $R(j,g)$ 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 figure 2. Let the source node be R(1,1) (node 5) and the destination node be R(0,3) (node 12). 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) (node 14). So, node 5 transmits on λ_3 and the packet is received by node 14. Node 14 then uses the local group interconnection to forward the packet to node 12 on wavelength λ_0. So, a single opto-electronic (O/E) conversion takes place at node 14. In some cases source node R(j,g) may directly transmit to destination node R(k,m). As in the previous example, if the source node was again node R(1,1) (node 5) and if the destination was node R(2,3), then node 5 could directly transmit on λ_2 which is received by node 14, the intended destination. In RAPID, only a single opto-electronic conversion is needed to implement complete connectivity for any network size.

0-7695-2132-0/04/$17.00 (C) 2004 IEEE

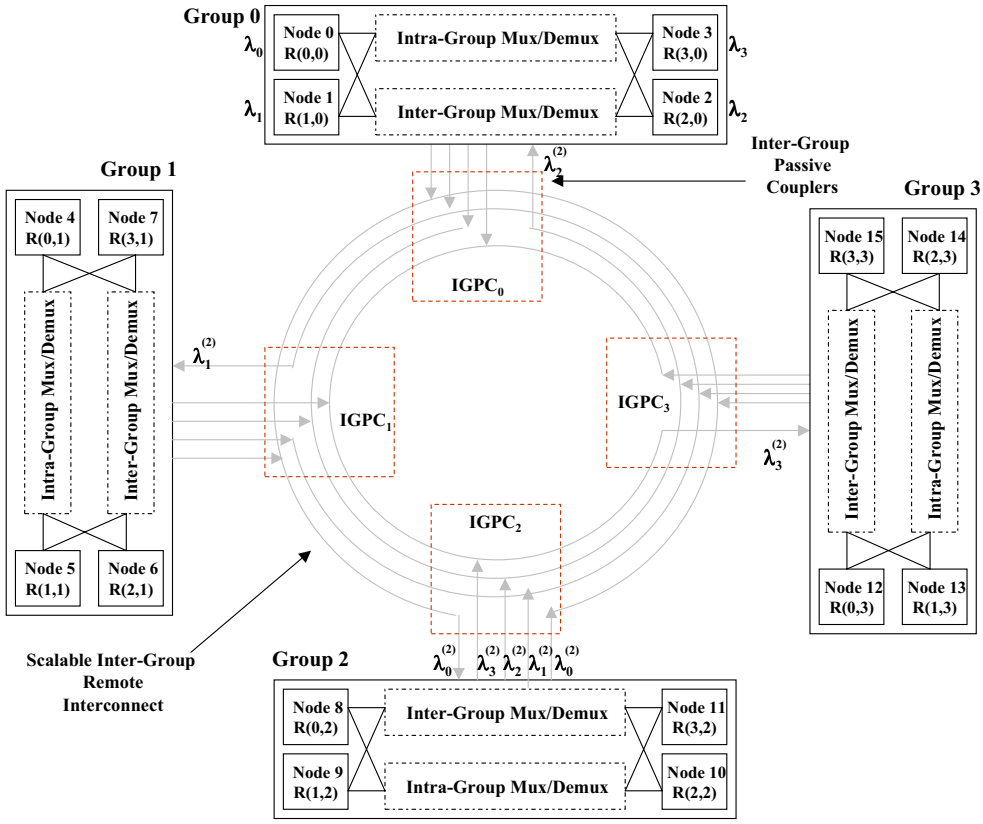

Figure 2. This figure shows the functional diagram of RAPID network along with local and remote wavelength allocation.

This is possible as the wavelength assignment algorithm designed for remote group permits high connectivity.

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, $\mathrm{R(d,g)}_{source}$ = $\mathrm{R(g)}_{destination}$ and 2) when the source node is located outside the group, $\mathrm{R(d,g)}_{source} \neq R(g)_{destination}$. 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. 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 5 that sends a multicast packet to group 3. It transmits on λ_2 and the destination is node 14 on group 3. Node 14 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.

2.3 Media Access Protocol for RAPID

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[14, 16]. 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 communications are shared among the nodes locally connected and not among all nodes. This is a significant feature

0-7695-2132-0/04/$17.00 (C) 2004 IEEE

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 before it can transmit its request, thereby significantly reducing the remote group latency.

We generate two sets of token for every local group g; one set of D tokens are shared for intra-group 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. If a processor doesn't have any request to transmit, it forwards the token to the next processor within the group, thereby reducing the waiting time for the next processor. The optical token will be held by the concerned processor until all communications has been completed that uses the particular wavelength. The processor transfers the token when it is transmitting the last request such that the token transfer completely overlaps with the transmission of the request/response[18].

2.4 Optical Implementation

Optical interconnects based on complimentary metal oxide semiconductor CMOS/VCSEL technology have been widely proposed for high-performance computing applications[19, 20]. The approach followed in our design is the most widely used hybrid integration using flip-chip bonding of OE-VLSI components[20]. The key components of the proposed architecture are multiwavelength vertical cavity surface emitting lasers (VCSELs)[21], photodetectors, directional couplers, multiplexers and demultiplexers that can be integrated with compatible optical technology. All wavelengths from the different nodes are combined using the Arrayed Waveguide Grating[22] and coupler array for intra-group interconnections. Low loss AWG can be used in combination with an array of low loss directional couplers constructed as a tree for multiplexing the signals. Depending on the wavelengths combined by the AWG, the multiplexed output from the AWG may appear at any output port. In order to combine this multiplexed output to the specific output port which will then feed the signal to the next AWG for demultiplexing, directional coupler array is used. Low loss directional couplers can be designed for specific lengths, such that all wavelengths can couple to the next waveguide with low loss. The multiplexed output forms the input to the demultiplexer. The signal is then demultiplexed to the appropriate destination nodes using integrated optical waveguides.

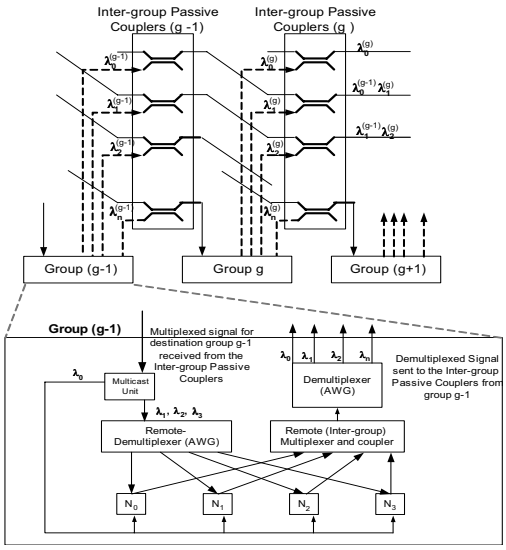

Figure 3. The figure shows the group $g-1$ inter-group interconnections.

For remote communication shown in figure 3, again all packets on different wavelengths are multiplexed at group $g-1$ (shown in the inset) similar to the intra-group multiplexer. The output is then sent to the AWG demultiplexer, which demultiplexes the signal. The demultiplexed signal forms the input to the Inter-Group Passive Couplers (IGPC (g)). The IGPC is comprised of directional couplers to couple the demultiplexed signal from groups, [...$(g-1)$, g, $(g+1)$,...]. Each IGPC, consists of $(G-1)$ directional couplers. Consider the output of the demultiplexer from group $(g-1)$ that feeds the signals to IGPC (g-1). The wavelength λ_0 is coupled with λ_1 from the next IGPC (g) and this combined signal will be added to λ_2 at the next IGPC $(g+1)$. In this way, different wavelengths from each group are added to this signal using directional couplers with low loss. The multiplexed signal then is returned back to group $(g-1)$. Going back to the inset in figure 3, the multiplexed signal reaches the multicast unit. The multicast unit comprises of a optical circulator and a fiber bragg grating is used to remove the wavelength λ_0. The multicast link then distributes this signal to all the processors in the group, thereby implementing multicast functionality. The multiplexed signal without λ_0 is then locally demultiplexed using AWG. For both local and remote communication implementations, RAPID uses only passive technology such as gratings (AWG, fiber bragg), directional couplers and waveguide/fiber optics. The use of passive technology are two fold; (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.

0-7695-2132-0/04/$17.00 (C) 2004 IEEE

3 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.

3.1 Comparison with other popular networks based on network characteristics

In this section, we present an analysis of the scalability of RAPID architecture with respect to several parameters such as node degree, diameter, bisection width and the number of links. 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, 2-D Mesh and Scalable Optical Crossbar Network (SOCN) [23]. Each of these networks will be compared with respect to degree, diameter, number of links and bisection width. The notation $RAPID(d = g)$ implies that both the number of nodes per group d and number of groups g are varied in order to vary the number of processors. For both ring and the crossbar, number of nodes, N is the only variable. The notation, $Mesh(w, d = 2)$ and $Torus(w, d = 2)$ implies that the dimensionality is fixed and the size of the network varies with number of nodes/ring/bus w. $SOCN(d = 4, g)$ implies that the number of groups g is changed in order to vary the number of nodes N.

Figure 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 a 1000 node network, each node needs to be connected only to IGI (local communication), to IGPC (remote communication) and to the multicast channel. Figure 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. Figure 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. Figure 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 inter-group communication, thereby showing a much better scalability in the number of links for very large-scale systems.

3.2 Power Budget Analysis

Calculation of a power budget and the signal-to-noise ratio at the receiver is important for confirming the realizability and scalability of any optical interconnect implementation. The signal-to-noise ratio 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 maybe as low as 10^{-15}. For such a BER, we computed that the received power should be 6.5 $mWatt$ or 49.694 dB or 8.129 dBm[18, 24]. High-powered VCSEL arrays delivering output power as high as 6.4 mW or -21.938 dB have been reported[25]. 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 intra-group interconnections. The various losses are VCSEL-waveguide coupling (-0.2dB), propagation in the waveguide/fiber (-0.5dB), arrayed waveguide grating (-2.1 $\times$ 2dB) for multiplexing and demultiplexing, coupler array ($-0.225 \times log_2(D)$) and receiver coupling (-0.2dB). The total loss amounts to $-5.1 - 0.225 \times log_2(D)$. The loss in a AWG up to 32 channels with 0.8 nm (100 (GHz)) channel spacing can be as low as 2.1 dB[26]. The loss in a directional coupler (5%) has a logarithmic dependence on the number of intra-nodes connected, D. For such values, we can have thousands of nodes connected. This implies that locally, we will be limited only by the number of wavelengths available.

For remote inter-group one-to-one communication, all the losses for intra-group interconnection will also be present. The other losses are an additional AWG (-2.1dB), G directional couplers at the IGPC (-0.225$dB \times G$), circulator (-0.5dB), fiber bragg grating (-0.5dB), waveguide-to-fiber MT Ferrule connector (-1dB) and additional propagation loss in the fiber/waveguide (-2dB). Note, that the total remote losses are added to the total local losses to give $-13.2dB - 0.225dB \times log_2(D) - 0.225 \times G$. Now, with G = D, the number of groups that can be connected is approximately 60. This implies that RAPID can scale up to 3600 (=60 $\times$ 60) nodes, which represents a factor of 3-5 times greater scalability as compared to current electrical DSMs. By using passive optical components in the design of RAPID, we have achieved greater scalability that conventional optical or electrical interconnects.

3.3 Simulation Assumptions and Methodology

In this section, we describe the simulation methodology and the preliminary 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

0-7695-2132-0/04/$17.00 (C) 2004 IEEE

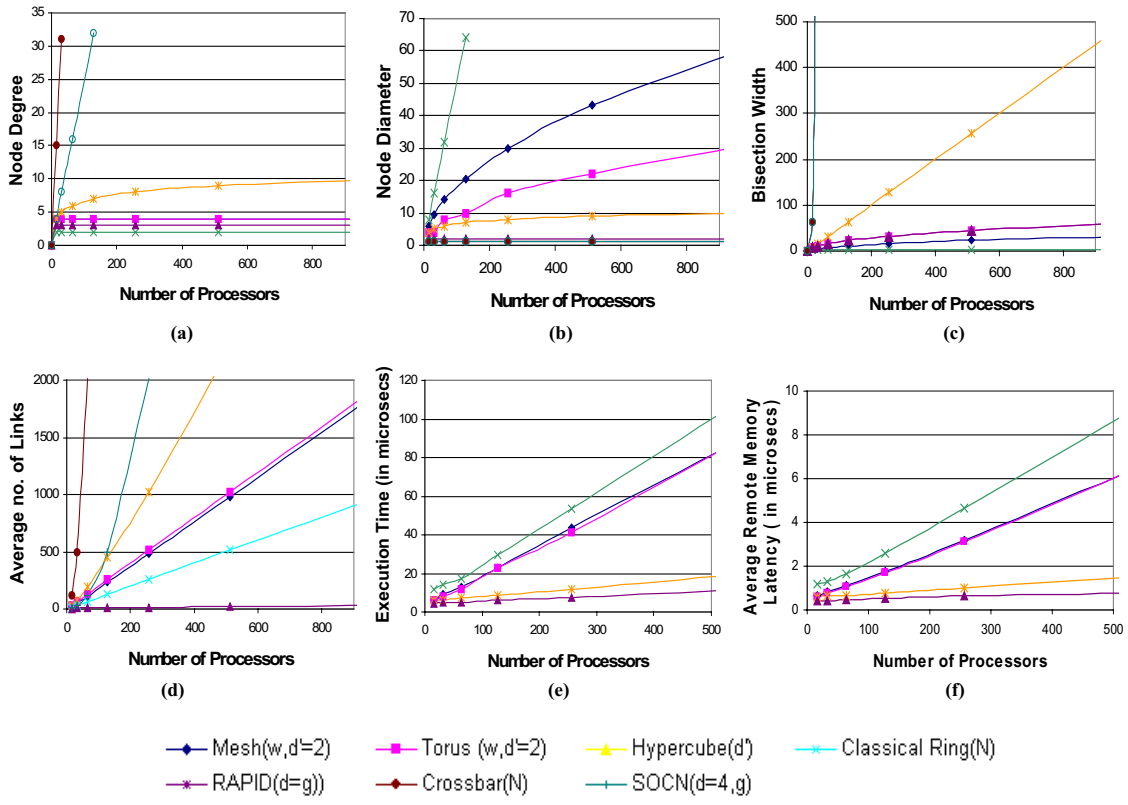

Figure 4. Fig (a) shows the degree comparison, (b) shows the diameter comparison, (c) shows the bisection width comparison for varying number of processors, (d) shows the no. of links comparison, (e) shows the execution time for various topologies simulated and (f) shows the average remote memory access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

CSIM[27], 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 only preliminary data and future simulation will include results based on shared memory benchmarks. This simulation was carried out to give us confidence that RAPID can reduce the remote memory access latency. In this simulation, we model accurately contention at all resources for both electrical and optical networks.

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) 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 $(1 - P_{nohop})$, 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 acknowledgement 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. Write backs are modelled for every transac-

0-7695-2132-0/04/$17.00 (C) 2004 IEEE

Table 2. Electrical and Optical Simulated System Parameters

Processor Parameters	
Processor Speed	1 GHz
Number of MSHRs	4
L1 hit Time	1 cycle (1nsec)
L2 hit Time	15 cycle (15nsec)
Cache to Cache Transfer Time	25 cycles (25nsec)
Memory Access Time	70 cycles (70nsec)
Electrical Network Parameters	
Flit Size	8 bytes
Non-data message size	16 bytes
Data-size	64 bytes
Router Speed	500 Mhz
Router's Internal bus width	64 bits
Channel Speed	10 Ghz
Virtual Channels	4
Header Routing Time	12.8nsec
Data Switching Time	6.4nsec
Data Propagation Time	6.4nsec
Optical Network Parameters	
Channel Speed	10 Ghz
Non-data Transmission Time	12.8nsec
Data Transmission Time	51.2nsec
O/E and E/O Delay (non-data)	(12.8+12.8)nsec
O/E and E/O Delay (data)	(12.8+51.2)nsec
Token Passing Latency	6.4nsec

tion that originates at the requestor with a probability of $P_{writeback}$. For 2-hop transactions, the requestor chooses the home node and for 3-hop transactions, the owner and N_{sharer} are chosen by the the home node, from the maximum number of simulated nodes using a uniform distribution. All the above simulation parameters were chosen from different technical manuscripts[28, 29] and these parameters were consistent for both optical and electrical networks. Contention is modelled at all system resources; MSHRs, directory/memory modules, network interfaces, virtual channels (in case of electrical networks) and optical tokens (in case of RAPID).

Table 2 summarizes the parameters of the simulated system. 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[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. We model O/E (optical to electrical) and E/O (electrical to optical) delays of $12.8nsec$. The optical packet can be processed as soon as the header is received, thereby reducing the latency. The token passing latency is completely overlapped with the packet transmission latency i.e. a node that begins transmission on a specific wavelength can immediately transmit the token to the next processor.

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 outperforms all networks by maximizing the the channel availability and maintaining a low diameter for large number of processors. 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. Figure 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%. These results show that RAPID can improve the performance of DSMs by reducing the remote memory access latency.

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. This network architecture provides distinct performance and cost advantages over traditional electrical interconnects and even over other optical networks.

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

0-7695-2132-0/04/$17.00 (C) 2004 IEEE

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] Alan Charlesworth, "Starfire: Extending the smp envelope," *IEEE Micro*, vol. 18, no. 1, pp. 39–49, Jan-Feb 1998.

[4] D.Kroft, "Lock-free instruction fetch/prefetch cache organization," in *Proceedings of the 8th Annual International Symposium on Computer Architecture*, New York, 1981, pp. 81–87.

[5] Tien-Fu and Jean-Loup Baer, "A performance study of software and hardware data prefetching schemes," in *Proceedings of the 21st Annual International Symposium on Computer Architecture*, Chicago, 1994, pp. 223–232.

[6] A.Rogers and K.Li, "Software support for speculative loads," in *Proceedings of the Fifth Symposium on Architectural Support for Programming Languages and Operating Systems*, 1992, pp. 38–50.

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

[8] 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.

[9] Donglai Dai and Dhabaleswar K. Panda, "How can we design better networks for DSM systems?," *Lecture Notes in Computer Science*, vol. 1417, pp. 171–184, 1998.

[10] Alan Charlesworth, "The sun fireplane smp interconnect in the sunfire 3800-6800," in *Hot Interconnects 9*, August 2001, pp. 37–42.

[11] Kourosh Gharachorloo, Madhu Sharma, Simon Steely, and Stephen Van Doren, "Architecture and design of alphaserver gs320," in *Proceedings of the 9th International Conference on Architectural Support for Programming Languages and Operating Systems*, 2000, pp. 13–24.

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

[13] 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.

[14] 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.

[15] Val N. Morozov, Peter S. Guilfoyle, and Zephyr Cove, "A new paradigm for parallel optical interconnects," in *Frontiers in Optics OSA Annual Meeting 2003*, October 2003, p. 95.

[16] 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.

[17] Ahmed Louri and Rajdeep Gupta, "Hierarchical optical ring interconnection (horn): A scalable interconnection-network for multiprocessors and multicomputers," *Applied Optics*, vol. 36, pp. 430–442, January 1997.

[18] Ahmed Louri and Avinash Karanth Kodi, "Symnet: An optical interconnection network for large-scale, high-performance symmetric multiprocessors," *Applied Optics*, vol. 42, pp. 3407–3417, 2003.

[19] Yue Liu, "Heterogeneous integration of oe arrays with si electronics and micro-optics," in *Proceedings of the Electronic Components and Technology Conference*, 2001, pp. 864–869.

[20] Ashok V. Krishnamoorthy and Keith W. Goossen, "Optoelectronic-vlsi: Photonic integrated with vlsi circuits," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 4, pp. 899–912, 1998.

[21] Masakazu Arai, Takashi Kondo, Akihiro Matsutani, Tomoyuki Miyamoto, and Fumio 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 Journal of Selected Topics in Quantum Electronics*, vol. 8, pp. 811–816, July/August 2002.

[22] Meint K. Smit, "Phasar-based wdm-devices: Principles, design and applications," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 2, pp. 236–250, June 1996.

[23] 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.

[24] T. V. Moui, "Receiver design for high-speed optical fiber systems," *IEEE Journal of Lightwave Technology*, vol. LT-2, pp. 234–267, 1984.

[25] R.Pu, E.M.Hayes, C.W.Wilmsen, K.D.Ohoquette, H.Q.Hou, and K.M.Geib, "Comparison of techniques for bonding vcsels directly to ics," *JOSA*, vol. 1, pp. 324–329, 1999.

[26] Yoshinori Hibino, "An array of photonic filtering advantages: Arrayed waveguide-grating multi/demultiplexers for photonic networks," IEEE LEOS Newsletter, August 2001.

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

[28] 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.

[29] Milo M. Martin, Pacia J. Harper, Daniel J. Sorin, Mark D. Hill, and David A. Wood, "Using destination-set prediction to improve the latency/bandwidth tradeoff in shared memory multiprocessors," in *Proceedings of the 30th Annual International Symposium on Computer Architecture*, June 2003, pp. 25–36.