RAPID for high-performance computing systems: architecture and performance evaluation

Avinash Karanth Kodi and Ahmed Louri

The limited bandwidth and the increase in power dissipation at longer communication distances and higher bit rates will create a major communication bottleneck in high-performance computing systems (HPCS), affecting not only their performance, but also their scalability. As a solution, we propose an optical-interconnect-based architecture for HPCS called reconfigurable all-photonic interconnect for parallel and distributed systems (RAPID) that alleviates the bandwidth density, optimizes power consumption, and enhances scalability. We also present two cost-effective design alternatives of the architecture, a modified version called M-RAPID and an extended version called E-RAPID that minimizes the cost of the interconnect based on the number of transmitters required. We perform a detailed simulation of the proposed RAPID architecture and compare it to several electrical HPCS interconnects. Based on the performance study, RAPID architecture shows 30%–50% increased throughput and 50%–75% reduced network latency as compared to HPCS electrical networks.

OCIS codes: 200.0200, 200.4650.

1. Introduction

In high-performance computing systems (HPCS), fundamental electrical signaling limitations result in reduced interprocessor communication bandwidth and increased power dissipation at higher bit rates and longer communication distances.[1,2] This limited bandwidth and connectivity of electrical interconnects will have negative effects on several key performance measures of HPCS. This includes execution time or processor latency, processor utilization, and network latency. The limited bandwidth will cause the processor to stall for a longer time, while waiting for the required data, and consequently will lower its utilization. For the network, the limited connectivity and bandwidth will cause longer queuing and routing delays throughout the network switches, due to extensive multiplexing of numerous signals onto limited serial input/output (I/O) links.[3] It is estimated that future HPCS utilizing off-the-shelf processors will be limited to approximately 20 Gbits/s even with aggressive signaling techniques (modulation schemes, equalization, low-loss board materials),[4,5] and architectural innovations (chip multiprocessors, multicore, simultaneous multithreading).[3] However, future bandwidth predictions for HPCS are estimated to be approximately 40 Tbits/s.[1,2] Clearly, this will create a major performance bottleneck at the board level, and, if not dealt with, will become the fundamental impediment to future scalable HPCS.

It is generally accepted that to continue the design of HPCS into the next decade, it is necessary to integrate new interconnect technologies with complementary metal-oxide semiconductor (CMOS) processing.[2,5–8] One such technology is optical interconnects. Optical interconnects offer several well-known advantages such as higher spatial and temporal bandwidths, lower cross talk independent of data rates, higher interconnect densities, better signal integrity at high frequencies, lower signal attenuation, and lower power requirements at higher bit rates; all of which could potentially achieve the much desired high bit rates data communication at a much reduced power level at the board-to-board distances (0.1–1 m).

The proposed optical-interconnect-based architecture called reconfigurable all-photonic interconnect for parallel and distributed systems (RAPID) maximizes the channel bandwidth, increases the connectivity, reduces the network latency, minimizes cost, and provides easy scalability for board-to-board and backplane HPCS. It should be noted that while the baseline RAPID architecture is an all-photonic net-

The authors are with the Department of Electrical and Computer Engineering, University of Arizona, 1230 E. Speedway Boulevard, Tucson, Arizona 85721. A. Louri's e-mail address is louri@ece.arizona.edu.

Received 1 December 2005; revised 16 May 2006; accepted 7 June 2006; posted 7 June 2006 (Doc. ID 66411).

0003-6935/06/256326-09$15.00/0

work, the modified and extended versions are optoelectronic networks that are regulated by electrical flow control mechanisms. RAPID exploits several optical technology features that enable: (1) ample communication bandwidths by aggressively utilizing wavelength-division multiplexing (WDM) and space-division multiplexing (SDM) techniques and combining them into a multiple WDM (M-WDM) technique, (2) high connectivity thereby reducing the network diameter resulting in lower queueing–routing delays for packet transmission, and (3) scalable bandwidth and cost that grow linearly with the number of nodes, while providing low latency. In addition, RAPID relies solely on passive components for the design of the transfer medium without any active optical switches, thereby reducing power dissipation, minimizing cost, and accelerating communication. The baseline RAPID architecture, the routing and wavelength assignment in RAPID, and media access protocol for RAPID are described in the next section.

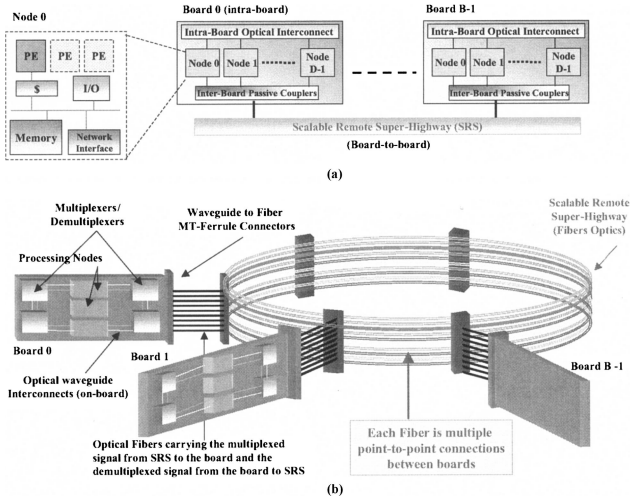

Fig. 1. Architectural overview of RAPID. Every node is connected to two scalable interconnects: an optical intraboard interconnect and a SRS.

2. Baseline RAPID Architecture

A baseline RAPID network is defined by a triple: (C, B, D) where C is the total number of clusters, B is the total number of boards per cluster, and D is the total number of nodes per board. The total number of nodes in RAPID is the multiplicative factor, $N = C \times D \times B$. Figure 1 shows the conceptual RAPID architecture for a single cluster. In Fig. 1(a), 0 up to $D - 1$ nodes are connected together to form a board. Boards, 0 up to $B - 1$, are connected to form a single cluster. All nodes are connected to two subnetworks, a scalable intraboard optical interconnection (IBI) and a scalable remote superhighway (SRS) via passive couplers. We have separated intraboard and interboard (remote) communications from one another to provide a more efficient implementation for both communications. It should be noted that the term all-photonic used in RAPID is applicable only to interboard communication, whereas intraboard communication can be both optical as well as electrical. RAPID is designed such that every node has two sets of fixed-array transmitters and fixed receivers for intraboard and interboard communication. Figure 1 (b) shows the conceptual diagram of a RAPID network. All interconnections on the board are implemented using optical waveguides and the interconnections from the board to SRS are implemented using optical fiber using multiplexers and demultiplexers. Although the architecture is shown as a ring system, this is only done for the clarity of the illustration. RAPID is actually implemented as a point-to-point topology as explained next.

A. Wavelength Allocation and Routing for Baseline RAPID

In the baseline RAPID design, for intraboard communication, the number of wavelengths equals the number of nodes D, such that every node is assigned a wavelength on which it can receive signals. Contention to the same wavelength is resolved through the token-based approach explained later. Figure 2 shows the remote wavelength assignment scheme in a $R(1, 4, 4)$ system, i.e., $C = 1$, $D = 4$, $B = 4$. For remote communication, different wavelengths from various boards are selectively merged to separate channels to provide high connectivity. Remote wavelengths are indicated by $\lambda_i^{(s,c)}$, where i is the wavelength, s is the source board number, and c is the cluster number from which the wavelength originates. To clarify, c is dropped since only single cluster working is explained. The wavelength assigned for a given source board s and destination board d is given by $\lambda_{B-(d-s)}{}^{(s)}$ if $d > s$ and $\lambda_{(d-s)}{}^{(s)}$ if $s > d$, where B is the total number of boards in the system, the superscript indicates the source board (in parentheses), and the subscript indicates the wavelength to be transmitted on. For example, if any node on board 1 needs to communicate with any node in board 2, the wavelength to be used is $\lambda_3{}^{(1)}$, and for reverse communication, the wavelength required is $\lambda_1{}^{(2)}$. To illustrate with an example, consider the board 0 transmitter set. All nodes on board 0 have an array of transmitters such that they can transmit on any wavelength $\lambda_i{}^{(0)}$, $i = 0, 1, 2, 3$. Any node in board 0 can communicate with itself on $\lambda_0{}^{(0)}$, with board 1 on $\lambda_3{}^{(0)}$, with board 2 on $\lambda_2{}^{(0)}$, and with board 3 on $\lambda_3{}^{(0)}$. The physical fiber channel on which λ_0 is transmitted is called the home channel for that particular board (shown as a dotted line for board 0). All signals originating from a particular board are demultiplexed and then selectively multiplexed with different home board channels. For board 0, the multiplexed signal on the home channel, $(\lambda_0{}^{(0)} + \lambda_1{}^{(1)} + \lambda_2{}^{(2)} + \lambda_3{}^{(3)})$ is then demultiplexed at the board 0 receiver. As the receivers are fixed, λ_i, $i = 1, 2, 3$ are received by node $i - 1$. The wavelength λ_0 is used for multicast and broadcast communication.[9] For remote traffic, the number of wavelengths required to obtain the connectivity mentioned above is B, i.e. $(B - 1)$ wavelengths are required to communicate with every other board and one more wavelength for multicast

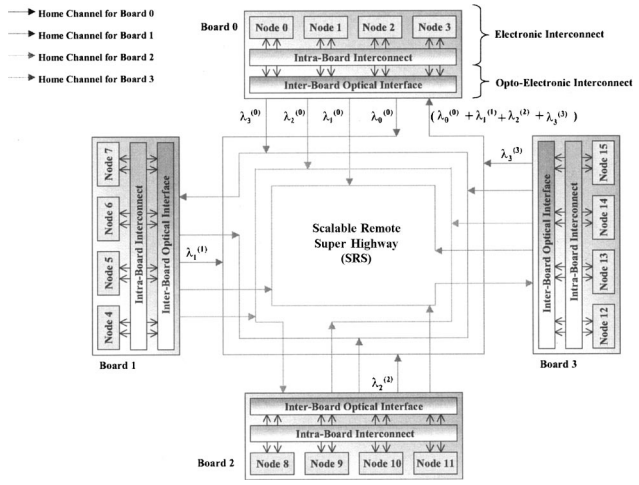

Fig. 2. Routing and wavelength assignment for four nodes/board, four boards, and four wavelengths.

communication. The wavelengths used in each home channel are the same, thereby reusing the same set of wavelengths for both intraboard and interboard communications.

B. Media Access Control: Parallel Token-Based Approach

The media access control (MAC) protocol designed to achieve mutual exclusive access to the remote communication channels, is an enhanced time-division multiplexing (TDM) technique called parallel-token time-division multiplexing (PT-TDM). In PT-TDM, token sharing for off-board communications is confined to the locally connected nodes (within a board), and not among all the nodes on various system boards, thereby enabling a distributed control policy that enhances scalability. Moreover, as opposed to a traditional TDM in which preassigned slots are available irrespective of communication needs, PT-TDM forwards tokens immediately to its successor, reducing communication latency. In RAPID, under a high network load, a node waits only for $D - 1$ transmissions of the packet to a particular destination board before it can transmit its request, thereby significantly reducing the interboard communication latency. Two sets of tokens are generated for every board b; one set of D tokens are shared for intraboard communications, and the other set of D tokens are shared for intergroup communications. To prevent a collision of requests, a node can transmit a packet on wavelength d depending on the token received. The token is held by the concerned node until it completes the transmission of the packet, after which it forwards the token to the next node.

3. Cost-Effective Design Alternatives

Wavelength reuse minimizes the number of different wavelengths needed for the architecture. In the baseline RAPID architecture, the number of transmitters required per node is proportional to $(2D)$, where D is the number of nodes per board. This is because D wavelengths are required for intraboard communication and another D for interboard communication. For example, a moderate-sized RAPID network designed with $N = 16$ ($B = 4$, $D = 4$, $C = 1$) will require at least eight transmitters working at different wavelengths per node for both intraboard and interboard communications. While tunable transmitters and active switching elements can potentially reduce the number of transmitters required per node, the downside is the increase in the cost of the interconnect. To reduce the number of the transmitters required per node in a cost-effective manner, we propose a modified version called M-RAPID and an extended version called E-RAPID that attempt to reduce the number of multiple transmitters required, by replacing parts of the interconnect with an electrical interconnect while retaining the optical interconnect for interboard communication (SRS).

A. M-RAPID: Modified RAPID

In the M-RAPID design, the optical interconnect for intraboard communication is replaced with an electronic switching configuration. Conventional peripheral component interface (PCI) and PCI-X-based[10] electrical solutions are performance limited due to difficulties with electrical signaling at multigigahertz rates over electrical interconnects.[5] Newer serial, point-to-point, I/O technologies such as PCI-Express[10] and HyperTransport[11] deliver much higher bandwidths (tens of Gbps) for distances <0.1 m. These interconnects can be utilized on-board in M-RAPID design. This achieves two distinct advantages: (1) The wavelengths required for the SRS implementation now reduce to (D) per node as opposed to $(2D)$ for the baseline RAPID design. (2) The token processing overhead for the intraboard communication is eliminated. Consequently, electrical credit-based flow control mechanism is implemented for buffer management. The detailed implementation of the architecture is discussed in the next subsection.

B. E-RAPID: Extended RAPID

In the E-RAPID design, the optical transmitters and receivers are pushed to the board edges while retaining the same functionality of SRS design for interboard communication as explained before. This achieves several distinct advantages: (1) The token processing overhead is completely removed and replaced by the usual electrical flow control mechanisms for all communications occurring on the board. (2) Unlike in RAPID, in E-RAPID, if there are any packets in the output buffers of the transmitters, these packets are transmitted immediately without waiting for TDM slots as in PT-TDM. This decouples optical interfaces from the processor, requiring minimal redesign of current system boards and provides a practical perspective for the insertion of optics at the board level. (3) The wavelengths required for SRS implementation now reduces to approximately one per node with the assumption that the number of wavelengths required per board equals the total number of boards B connected. In fact, the number of

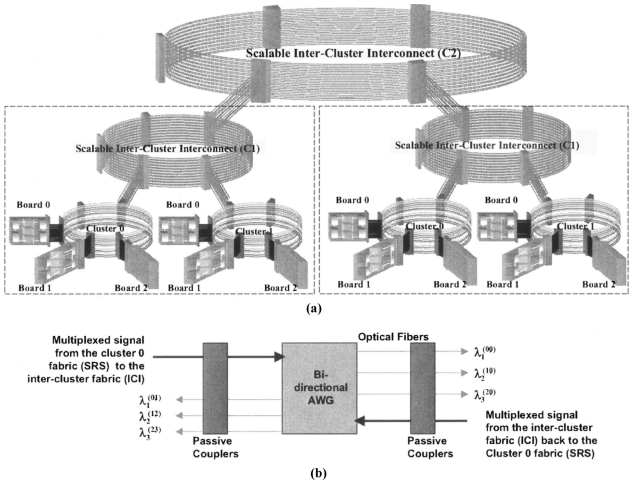

Fig. 3. (a) Scalable RAPID architecture, multiple levels of clusters connected to SICI. (b) Bidirectional demultiplexers are used to connect different levels of clusters.

wavelengths has no dependence on the number of nodes D connected within a board. The electrical routing–switching component is responsible for routing the packets to the appropriate wavelengths (for transmitting) and to the appropriate receiver nodes (for receiving).

C. Scalability of RAPID

RAPID provides system scaling in several ways[9]: these include adding nodes (D), boards (B), or even clusters (C). In this paper, we analyze the more interesting scaling by adding clusters (C) as shown in Fig. 3(a). This scaling mechanism applicable to all RAPID configurations is as follows: The original cluster [cluster 0 from Fig. 1(b)] will be replicated to obtain a new cluster (cluster 1). The original SRS will be replicated and named scalable intercluster interconnect (SICI), level C1. These clusters will be connected to the SICI, level C1 using bidirectional arrayed waveguide gratings (AWGs) as shown in Fig. 3(b). The multiplexed signal $(\lambda_1^{(0,0)} + \lambda_2^{(1,0)} + \lambda_3^{(2,0)})$ arriving from the cluster 0 SRS is demultiplexed and then selectively merged into different fibers at the SICI. Similarly, the multiplexed signal from the SICI $(\lambda_1^{(0,1)} + \lambda_2^{(1,2)} + \lambda_3^{(2,3)})$ is demultiplexed and then selectively merged into different fibers in SRS at cluster 0. From Fig. 3(a), it should be noted that the resulting architecture is symmetric, and resembling the original cluster with the boards being replaced by clusters. This scaling method would neither require new active switching elements, nor require any new wavelengths. Additionally, the routing algorithm will remain the same as for a single cluster. For example, with $D = 16$, $B = 16$, and $C = 16$, we can potentially have 4 K nodes using the E-RAPID architecture. This also provides the basis for designing hierarchical scalable architectures. For example, similar to adding intercluster interconnect, C1, another level of scalable intercluster interconnect, C2, can be added as shown in Fig. 3(a). Again, this design will not require any additional wavelengths and a similar bidirectional AWG used previously can be utilized. For example, with two levels of intercluster interconnect, eight wavelengths, eight clusters, eight boards, and eight nodes, we can achieve $8^5 = 32{,}768$ nodes. As current HPC systems consist of dual or quad processing cores, we can further scale the system to 65K or 131K processing elements.

4. Optical Implementation

In this section, we explain the optical components needed for the implementation and the integration methodology of the proposed network architecture using current CMOS technology.

A. Optical Components

The key components in designing RAPID are multiwavelength vertical-cavity surface-emitting lasers (VCSELs), photodetectors, couplers, multiplexers, and demultiplexers. We briefly describe some of the required components and their desired characteristics along with the methodology for their integration to build a scalable parallel architecture.

Laser Sources: VCSELs are a natural candidate as laser sources in the proposed architecture, owing to their ease of fabrication in one- and two-dimensional (2D) arrays, high power, good optical coupling to fibers and low cost. High-performance GaAs- and InGaAs-based selectively oxidized or proton implanted top-emitting VCSELs emitting at 780 to 980 nm have been reported in the literature.[5,12] While tunable transmitters could provide a range of wavelengths needed to implement RAPID, the need for fast wavelength switching (nanoseconds) over the entire spectral range make multiple wavelength VCSEL arrays as the design choice for RAPID interconnect. A multiwavelength VCSEL array[13] consisting of 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 can be used for RAPID architecture.

Waveguides and Fibers: Optical polymers are increasingly considered as highly versatile elements that can be readily transformed into single-mode, multimode, and micro-optical waveguide–fiber structures.[14] Acrylate-based polymers, developed by Allied Signals, have shown optical loss less than 0.1 dB/cm at 0.8 μm. The low loss in these polymers makes them an attractive material for constructing the 2×1 couplers, 1×2 splitters, and directional couplers for routing optical pulses from VCSELs to photodetectors.

Demultiplexers: Wavelength multiplexers–demultiplexers are fabricated using diffraction gratings or phased-array-based devices (also called AWGs).[15,16] Both are imaging devices; i.e., they image the field of an input waveguide onto an array of output waveguides in a dispersive way. 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

Table 1. Design Parameters for AWG Demultiplexer in RAPID

Parameter	Notation	Value
Wavelength spacing	$\nabla\lambda_0$	0.7 nm
Output waveguide spacing	∇x	8 mm
Path difference of arrayed waveguides	∇L	31.97 μm
Diffraction order	m	42
Pitch of the arrayed waveguide	d	8 μm
Focal length of the slab	f	2.1 mm
Free spectral range	FSR	22.45 nm
Number of arrayed waveguides	N	201
Effective refractive index of channel	n_c	1.470
Effective refractive index of the slab	n_s	1.497
Group refractive index	n_g	1.539

such as to obtain the required imaging and dispersive properties. In such a grating, both multiplexing and demultiplexing operations are the same except that the direction of light propagation is reversed. This feature is used in designing bidirectional AWG for ICI. The detailed parameters of the AWG for RAPID[9] for the specified VCSEL array is shown in Table 1.

B. Optical Integration

Optical interconnects based on CMOS–VCSEL technologies have been widely proposed for high-performance computing applications.[5,12] The approach followed in our design is the most widely used hybrid integration using flip-chip bonding of optoelectronic (QE) VLSI components. The VCSEL–PD arrays can be fabricated on a GaAs substrate such that the devices are designed to be backside emitting because of the desire to flip-chip bond them to CMOS driver circuits. The n and p contacts 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.[2,17]

The packet from the node is first routed through an electronic crossbar switch to the optical transmitter. The optical transmitter (VCSEL) then emits a signal on a particular wavelength. This signal is coupled to the waveguide using 45° micromirrors.[18,19] The optical signal from other nodes is combined using the coupler and is input to the demultiplexer, which separates all the wavelengths into different fibers of the SRS.

Figure 4(a) shows the cluster configuration. Figure 4(b) shows the cluster 0 implementation. Boards 0–2 each consisting of D nodes are interconnected using passive couplers and demultiplexers. Another board can be connected in the free slot or can be used to connect bidirectional AWG for the SICI. These boards are then plugged into the passive optical backplane using waveguides–fibers. Different wavelengths originating from each system board are coupled using

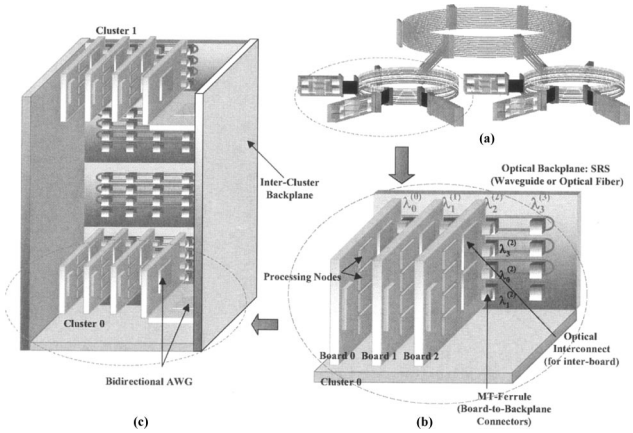

Fig. 4. Possible RAPID implementation. (a) Scalable cluster interconnect, (b) single cluster board-to-board interconnect implementation on a passive optical backplane, and (c) intercluster implementation using bidirectional AWGs.

MT-Ferrule connectors into a passive coupler onto the backplane. The wavelengths indicated along the horizontal direction ($\lambda_0^{(0)}$, $\lambda_1^{(1)}$, $\lambda_2^{(2)}$, and $\lambda_3^{(3)}$) are the wavelengths that board 0 receives the signals from other system boards (from Fig. 2). The wavelengths indicated along the vertical direction [$\lambda_2^{(2)}$, $\lambda_3^{(2)}$, $\lambda_0^{(2)}$, and $\lambda_1^{(2)}$] are the wavelengths that board 2 inserts into different fibers for communicating with other system boards. Figure 4(c) shows the complete system implementation of RAPID architecture as seen in Fig. 4(a). As explained before, the SRS backplane [from Fig. 4(b)] is replicated to form cluster 1. The free slot is used here to connect the bidirectional AWG to both the clusters. SICI is designed by replicating the SRS backplane and interconnecting the two clusters.

5. Performance Evaluation

The performance of RAPID is evaluated using YACSIM, a discrete event simulation, an electrical network component library and NETSIM,[20] discrete-event simulators and is compared to various electrical interconnects for both uniform and nonuniform traffic traces.[21] The electrical networks chosen for comparison were 2D torus, hypercube, and fat-tree topologies. These topologies are the most common clustering interconnects: the 2D torus is used in the Alpha 21364 network,[22] the hypercube is used in the SGI Spider chip used for SGI Origin machines,[23] and the fat-tree topology is the basis of most Mellanox switches used in Infiniband architectures,[24] as well as in Elan 2 used in QsNet.[21]

A. Simulation Methodology

YACSIM and NETSIM can be combined to construct a wide range of direct and indirect electrical interconnects. We modified the baseline-wormhole-routed NETSIM with virtual channels to decouple the allocation of channel bandwidth from the channel state, to achieve substantially higher throughput. Due to the lack of optical simulators at the system level, we

augmented the NETSIM component library by adding several optical components such as couplers, fibers, waveguides, demultiplexers, and splitters. The functional modeling of each of these components at the system level was implemented to determine three parameters of interest: (1) length, to determine the propagation latency; (2) attenuation, to determine the signal loss due to component; and (3) wavelength, to determine the routing within a component (demultiplexer). The components were then connected to design various WDM-routed RAPID configurations. For M-RAPID and E-RAPID, we designed the electrical intraboard interconnect using a crossbar switch.

We use cycle accurate simulations to evaluate the performance of RAPID and other electrical interconnects. Packets were injected according to the Bernoulli process based on the network load for a given simulation run. The network load is varied from 0.1 to 0.9 of the network capacity. The network capacity was determined from the expression N_c (packets–node–cycle), which is defined as the maximum sustainable throughput when a network is loaded with uniform random traffic.[25] The simulator was warmed up under load without taking measurements until steady state was reached (up to 1000 cycles). Then a sample of injected packets were labeled during a measurement interval (1000–10,000). The simulation was allowed to run until all the labeled packets reached their destinations.

The electrical network router model parameters reflect the design from an SGI Spider routing chip. These parameters reflect the design from the SGI Spider routing chip.[23] For the router model designed, the channel width is 16 bits and the speed is 400 MHz, resulting in a unidirectional bandwidth of 6.4 Gbits/s and a per-port bidirectional bandwidth of 12.8 Gbits/s. Credit-based flow control is implemented for a single flit buffer with credits incurring a single cycle channel delay. For the optical network, we assume a channel speed of 10 GHz, based on current optical technology. At 10 Gbits/s data rates, the transmission of an 8 byte flit takes around 6.4 ns $[= (8 \times 8)/(10 \times 10^9)]$. For most of the runs, we maintained a constant packet size of 64 bytes, resulting in an 8-flit packet size. The number of nodes simulated for various networks was 64 and 256.

B. Simulated Traffic Patterns

Network workloads that accurately reflect the high temporal and spatial traffic variance of many parallel numerical algorithms usually employed by scientific applications are most useful for evaluating the performance of HPCS. We present three sets of traces:

Uniform Traffic: In this pattern, each node randomly selects its destinations with equal probability.

Nonuniform Traffic: In this pattern, 75% of the traffic is directed to 25% of the nodes connected; the remaining 25% are uniformly distributed.

Permutation Patterns: In these static communication patterns, each node selects a fixed destination for all its transactions.[21] The permutation patterns tested were:

- Bit reversal: The node with binary coordinates $a_{n-1}, a_{n-2}, \ldots, a_1, a_0$ communicates with node $a_0, a_1, \ldots, a_{n-2}, a_{n-1}$.
- Matrix transpose: The node with binary coordinates $a_{n-1}, a_{n-2}, \ldots, a_1, a_0$ communicates with node $a_{n/2-1}, \ldots, a_0, a_{n-1}, a_{n/2}$.
- Complement: The node with binary coordinates $a_{n-1}, a_{n-2}, \ldots, a_1, a_0$ communicates with node $\overline{a_{n-1}, a_{n-2}, \ldots, a_1, a_0}$.

C. Simulation Results

The performance of an interconnection network under dynamic load is usually assessed by two quantitative parameters, the accepted bandwidth or throughput and the latency.[25] Accepted bandwidth is defined as the sustained data delivery rate given some offered bandwidth at the network input. Two important characteristics are the saturation point and the sustained rate after saturation. Saturation is defined as the minimum offered bandwidth where the accepted bandwidth is lower than the global packet creation rate at the source nodes. Before saturation, the offered bandwidth and accepted bandwidth remain the same. It is usually expected that the accepted bandwidth will remain stable after saturation, both in the presence of bursty applications that require peak performance for a short period of time and applications that operate after saturation in normal conditions. The network latency is the average delay spent by a packet in the network, from the insertion of the head flit into the input buffer till the reception of the tail flit at the destination. It includes the source queueing delay to ensure that the traffic pattern being applied at the measurement points is the intended pattern.

Figure 5 shows the throughput and network latency for uniform and nonuniform traffic traces for 64 and 256 nodes. For 64 nodes uniform traffic, RAPID configurations outperform all electrical networks, with all RAPID configurations showing an almost 30% improvement in throughput: due to the ample bandwidth provided by optics. From the latency plot for 64 nodes uniform traffic, it can be seen that the latency for RAPID and M-RAPID saturates at 40% of the network load, while E-RAPID shows a better performance and saturates at almost 60% of the network load. For nonuniform traffic, RAPID and M-RAPID almost double the throughput, while E-RAPID shows an almost 60% improvement in performance as compared to electrical interconnects. E-RAPID has a lower throughput due to the overhead of electrical flow control mechanisms implemented for both intraboard and interboard communications. RAPID and M-RAPID show almost identical throughput, as all nodes are directly connected to the SRS network. For 256 nodes, it can be seen that the delivered throughput is much higher for RAPID configurations, though

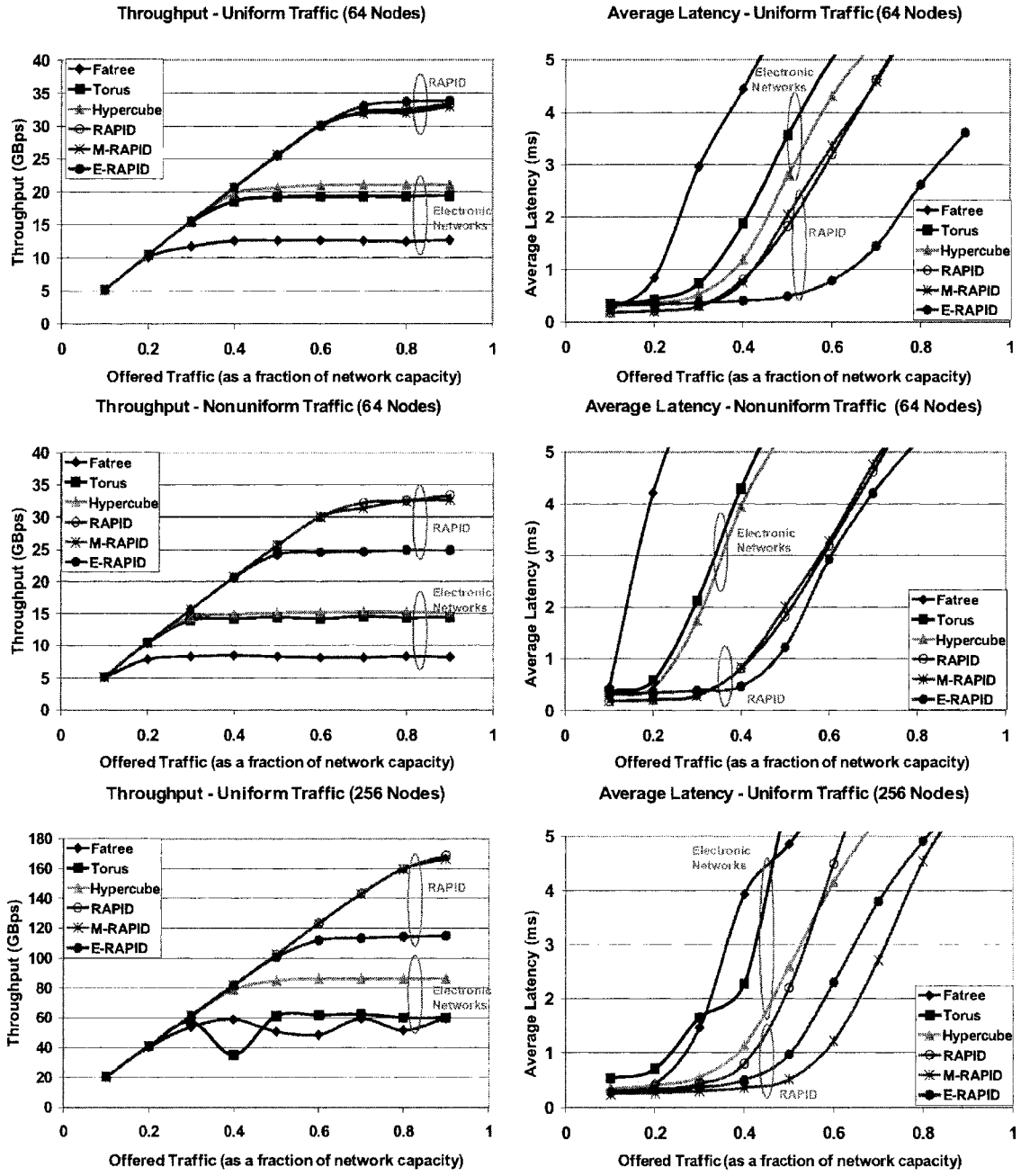

Fig. 5. Throughput and latency estimations for uniform and nonuniform traffic traces for 64 and 256 nodes.

the latency for baseline RAPID saturates much earlier and increases dramatically. As the size of the network increases, contention for tokens at both the intraboard and the interboard results in increased average latency for the baseline RAPID architecture. M-RAPID and E-RAPID provide a better performance as compared to the baseline RAPID for large network sizes, while for nonuniform traffic, baseline RAPID and M-RAPID provide better performance than E-RAPID.

Figure 6 shows the throughput and network latency for various permutation traffic for 64 nodes. RAPID configurations show better performance for Bit-Reversal and Matrix-Transpose traffic patterns. For the Complement traffic pattern, electronic networks outperform RAPID configurations. This is due to the design of RAPID architecture where all nodes within a board communicate with a particular destination board on a single wavelength. For example, nodes 0, 1, . . . , 7 on board 0 communicate with nodes 63, 62, . . . , 56 on board 7 using wavelength $\lambda_1^{(0)}$. This results in highly contented access for the same wavelength by all the nodes within the board, leading to low throughput and high average latency. To improve the throughput for these communication patterns, an increased number of wavelengths–channels should be allocated. Such reconfiguration mechanisms

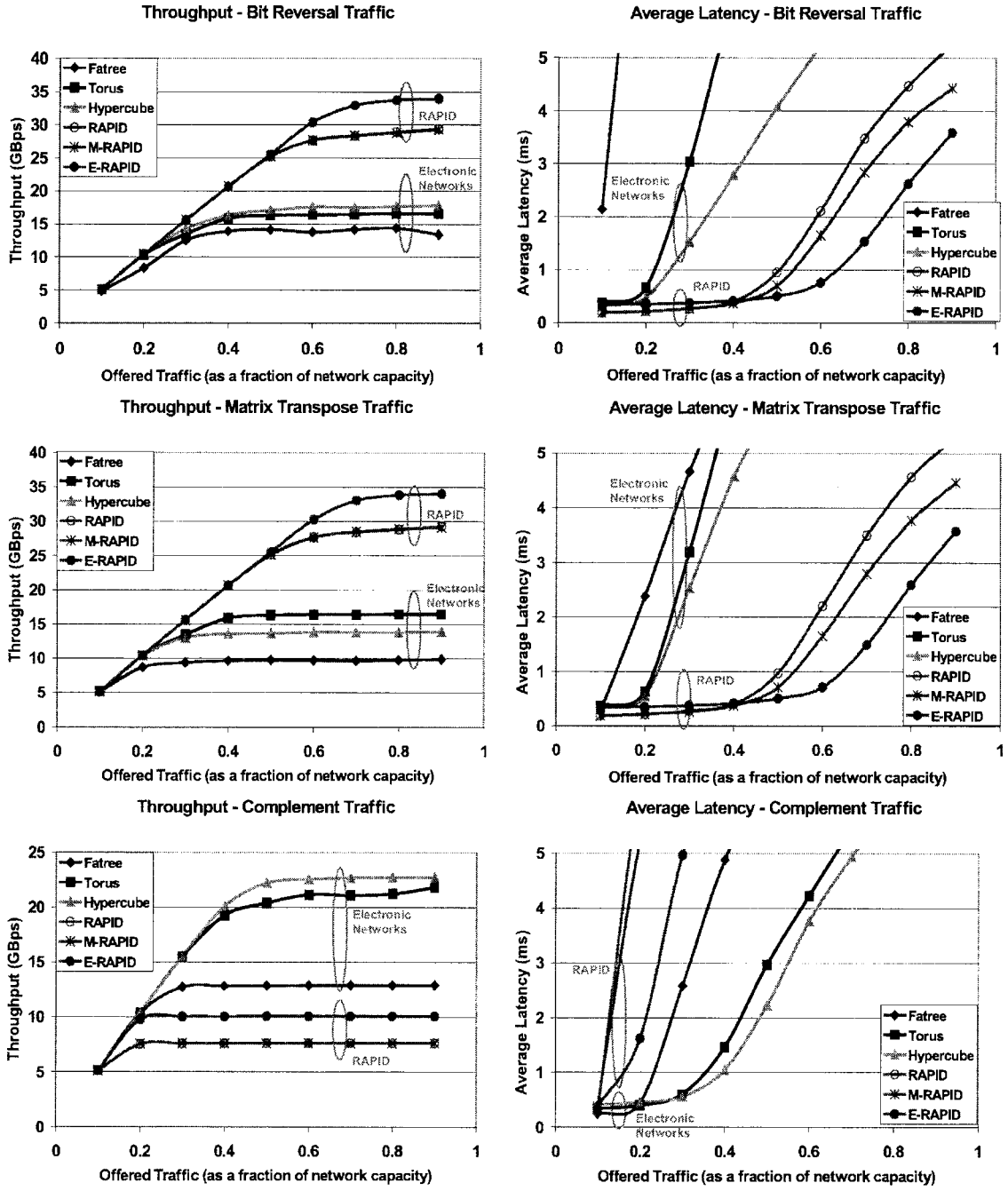

Fig. 6. Throughput and latency estimations for permutation traffic traces for 64 nodes.

should dynamically adapt to traffic patterns. In the future, we will evaluate the dynamic reconfiguration of RAPID.

6. Conclusion

We have presented two design alternatives of the baseline RAPID architecture: a modified version called M-RAPID and an extended version called E-RAPID that showed a reduction in the cost of the interconnect based on the number of transmitters required. We also presented cluster scaling that enabled building large-scale optical networks for HPCS using far fewer wavelengths and minimum system redesign. In addition, we developed an end-to-end system modeling and simulation framework to evaluate the performance of various RAPID configurations for uniform and nonuniform communication traffic patterns. A simulation result showed that E-RAPID could deliver a better performance than electrical interconnects using fewer components. Results also indicated that M-RAPID could perform as well as the baseline RAPID architecture using more optical components than E-RAPID, but less than baseline RAPID. The trade-offs based on the results indicate that E-RAPID architecture could optimize the cost while delivering better performance and

could form the basis of building cost-effective OE networks for HPCS.

This research was supported by NSF grants CCR-0000518 and CCR-0309537 and a grant from Intel Corporation.

References

1. A. F. Benner, M. Ignatowski, J. A. Kash, D. M. Kuchta, and M. B. Ritter, "Exploitation of optical interconnects in future server architectures," IBM J. Res. Dev. **49,** 755–775 (2005).
2. E. Mohammed, A. Alduino, T. Thomas, H. Braunisch, D. Lu, J. Heck, A. Liu, I. Young, B. Barnett, G. Vandentop, and R. Mooney, "Optical interconnect system integration for ultra-short-reach applications," Intel Technol. J. **8,** 114–127 (2004).
3. D. E. Culler, J. P. Singh, and A. Gupta, *Parallel Computer Architecture: a Hardware/Software Approach* (Morgan Kaufmann, 1999).
4. T. S. D. Huang, A. Landin, R. Lytel, and H. L. Davidson, "Optical interconnects: out of the box forever?" IEEE J. Sel. Top. Quantum Electron. **9,** 614–623 (2003).
5. B. E. Lemoff, M. E. Ali, G. Panotopoulos, G. M. Flower, B. Madhavan, A. F. J. Levi, and D. W. Dolfi, "MAUI: enabling fiber-to-the-processor with parallel multiwavelength optical interconnects," J. Lightwave Technol. **22,** 2043–2054 (2004).
6. D. A. B. Miller, "Rationale and challenges for optical interconnects to electronic chips," Proc. IEEE **88,** 728–749 (2000).
7. J. 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., **39,** 671–682 (2000).
8. J. Higgins, "Building secure Sun fire link interconnect networks using midframe servers," Sun MicroSystems, Sun BluePrints Online, 2003, http://www.sun.com/blueprints/0203/817-1656.pdf 817-1656.
9. A. K. Kodi and A. Louri, "RAPID: reconfigurable and scalable all-photonic interconnect for distributed shared memory multiprocessors," J. Lightwave Technol. **22,** 2101–2110 (2004).
10. *Intel Corporation White Paper: PCI Express Ethernet Networking,* Technical Rep. 254108-002 (Intel Corporation, 2003).
11. *Hyper Transport Technology I/O Link*, Technical Rep. 25012A (Advanced Micro Devices, 2001).
12. A. V. Krishnamoorthy and K. W. Goossen, "Optoelectronic-VLSI: photonic integrated with VLSI circuits," IEEE J. Sel. Top. Quantum Electron. **4,** 899–912 (1998).
13. 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. Sel. Top. Quantum Electron. **8,** 811–816 (2002).
14. L. Eldada and L. W. Shacklette, "Advances in polymer integrated optics," IEEE J. Sel. Top. Quantum Electron. **6,** 54–68 (2000).
15. M. K. Smit, "PHASAR-based WDM devices: principles, design and applications," IEEE J. Sel. Top. Quantum Electron. **2,** 236–250 (1996).
16. C. Dragone, "An $N \times N$ optical multiplexer using a planar arrangement of two star couplers," IEEE Photon. Technol. Lett. **3,** 1073–1075 (1991).
17. A. V. Krishnamoorthy, K. Gossen, L. Chirovsky, R. Rozier, P. Chandramani, S. Hui, J. Lopata, J. Walker, and L. A. D'Asaro, "16×16 VCSEL array flip-chip bonded to CMOS VLSI circuit," IEEE Photon. Technol. Lett. **12,** 1073–1075 (2000).
18. Y. Liu, L. Lin, C. Choi, B. Bihari, and T. Chen, "Optoelectronic integration of polymer waveguide array and metal-semiconductor-metal photodetector through micromirror couplers," IEEE Photon. Technol. Lett. **13,** 355–357 (2001).
19. M. Kagami, A. Kawasaki, and H. Ito, "A polymer optical waveguide with out-of-plane branching mirrors for surface-normal optical interconnections," J. Lightwave Technol. **19,** 1949–1955 (2001).
20. J. R. Jump, *YACSIM Reference Manual* (Rice University, 1993), http://wwwece.rice.edu/rppt.html.
21. F. Petrini, E. Frachtenberg, and A. Hoisie, "Performance evaluation of the quadries interconnection network," Cluster Comput. **6,** 125–142 (2003).
22. S. S. Mukherjee, P. Bannon, S. Lang, A. Spink, and D. Webb, "The Alpha 21364 network architecture," IEEE Micro **22,** 26–35 (2002).
23. M. Galles, "Spider: a high-speed network interconnect," IEEE Micro **17,** 34–39 (1997).
24. Mellanox Technologies, http://www.mellanox.com/.
25. W. J. Dally and B. Towles, *Principles and P. of Interconnection Networks* (Morgan Kaufmann, 2004).