Multidimensional and Reconfigurable Optical Interconnects for High-Performance Computing (HPC) Systems

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

***Abstract*—The increasing demand for higher communication bandwidth, reduced power consumption, and increased reliability combined with fundamental electrical signalling limitations is leading the drive for optics as an interconnect technology of choice for high-performance computing (HPC) systems. However, failure in any optical link can completely disrupt communication by isolating processing nodes in HPC systems. Moreover, while static allocation of wavelengths (channels) provides every node with equal opportunity for communication, it can also lead to network congestion for nonuniform traffic patterns. In this paper, we propose a multidimensional optoelectronic architecture, called nD-reconfigurable, all-photonic interconnect for distributed and parallel systems ($n$dimensional-RAPID) where n can be 1, 2, or 3. nD-RAPID exploits optical architecture and technology design space that simultaneously tackles both fault-tolerance and dynamic bandwidth reallocation (DBR) of system architecture. Fault-tolerance in nD-RAPID is enabled through a multidimensional architecture. DBR is implemented by the row–column switching matrix using silicon-on-insulator (SOI)-based microring resonators that adapts to changes in communication patterns at runtime. Simulation results indicate that nD-RAPID outperformed other electrical networks for most traffic patterns. Results on DBR show that the proposed row–column switch organization significantly improves throughput and latency with a slight increase in electrical power consumption ($\sim$0.4% for the worst case traffic).**

***Index Terms*—Fault tolerance, optical interconnections, parallel processing, reconfigurable architectures.**

I. Introduction

THE quest for higher performance (high bandwidth at low power) combined with electrical signaling limitations has resulted in optical interconnects being the interconnect technology of choice for chip-to-chip communications and even for on-chip communications. Several recent publications from both academia and industry have reported the advantages of short distance optical interconnects such as higher spatial and temporal bandwidths, lower crosstalk independent of data rates, higher interconnect densities, lower signal attenuation, and lower power requirements at higher bit rates [1]–[7], all of which make optical interconnects a viable technology for board-to-board communications.

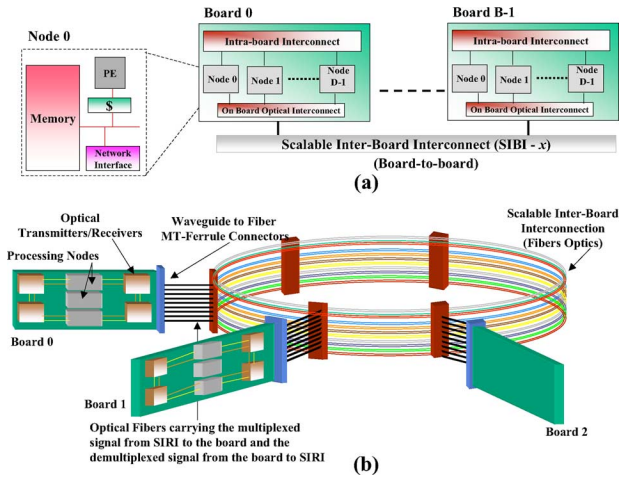

Fig. 1. (a) Architectural overview of nD-RAPID for $n = 1$. (b) Conceptual diagram of 1-D-RAPID.

As the computation and communication capabilities of high-performance computing (HPC) systems have grown at a furious pace as predicted by Moore's law, the probability of component failure either at the computing node [processor, memory, input/output (I/O)] or at the communicating link (channel, transmitter, receiver) has also proportionally increased. Therefore, while some component/link failure is bound to occur, a reliable system should be able to tolerate and recover from those faults with minimum performance degradation. In this paper, our focus is the interconnect or the link failure which causes two major problems: 1) link failure leads to isolating healthy processing node, thereby disrupting communication and its associated computation, and 2) link failure can lead to congestion in the network leading to deadlocks and the more complicated livelocks scenarios. Prior work in fault-tolerant optical interconnects include a hypercube connected rings with a depth-first search-based fault-tolerant routing algorithm (HCRNet, [8]); a three-stage Clos network to overcome link failure [9]; and a torus network that can tolerate a single optical switch/link failure [10].

In our previously proposed optical interconnect called reconfigurable, all-photonic interconnect for distributed and parallel systems (RAPID) [11], the routing and wavelength

Manuscript received September 26, 2008, revised May 11, 2009. First published June 30, 2009; current version published September 10, 2009. This work was supported in part by the National Science Foundation under Grants CCR-0538945 and ECCS-0725765.

A. K. Kodi is with the School of Electrical Engineering and Computer Science, Ohio University, Athens, OH 45701 USA (e-mail: avinashk@eecs.ohio.edu; kodi@ohio.edu).

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

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JLT.2009.2026187

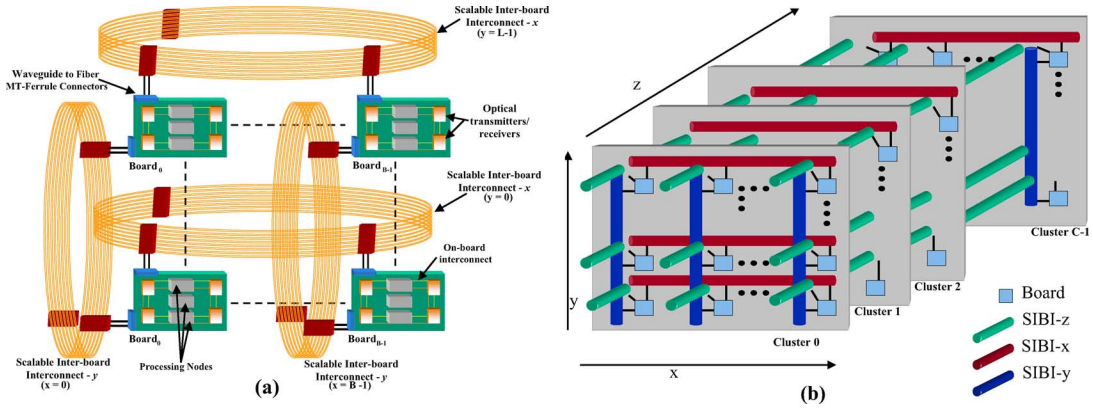

Fig. 2. Conceptual diagram of (a) $n\,D$-RAPID for $n = 2$ (the 1-D-RAPID is duplicated L number of times to create 2-D-RAPID) and (b) $n\,D$-RAPID for $n = 3$ (the 2-D-RAPID is duplicated C number of times to create 3-D-RAPID).

assignment (RWA) allocated bandwidth statically between various communicating boards using different wavelengths, fibers, and time slots. Static allocation of channels offers every node with equal opportunity for communication irrespective of the network loads. Although static allocation ensures fairness and is suitable for uniform traffic pattern, it can lead to network congestion for nonuniform communication patterns. On the other hand, dynamic reallocation of channels in response to actual network load could lead to improved performance for most communication patterns. Prior work on dynamic reconfiguration have used active electrooptic switching elements [12], time-slot-based bandwidth reallocation [13], both time- and space-based bandwidth switching [14], and passive optical devices combined with multiple transmitters [15]. In this paper, we propose an optoelectronic architecture, called nD-RAPID ($n$dimensional–RAPID), where n can be 1, 2, or 3. The term *all-photonic* has been carried on from the baseline architecture [11], while the proposed architecture has on-board electrical interconnect. nD-RAPID solves the combined issues of fault tolerance via architecture design and simultaneously improve performance through dynamic bandwidth reallocation (DBR). The proposed nD-RAPID is scaled in several dimensions to provide multiple paths from any source to any destination, thereby overcoming single or even multiple link failures and improving the reliability of the optical interconnect. In addition, performance-aware row–column switch is proposed for nD-RAPID which allows area-efficient and power-efficient DBR implementation while avoiding additional transmitters. The row–column switch is composed of wavelength-selective 1×2 silicon-on-insulator (SOI)-based microring resonators. These switches are fast ($\sim$10 ns), compact ($\sim$10-μm diameter), low power ($\sim$19 μW), and can be fabricated using standard complementary metal–oxide–semiconductor (CMOS) techniques [16], [17]. To the best of our knowledge, this is the first time an optoelectronic architecture is proposed which provides both fault tolerance and simultaneously improves performance via DBR. In what follows, we explain the nD-RAPID architecture in Section II, DBR implementation in Section III, architecture implementation in Section IV, and performance evaluation in Section V.

II. nD-RAPID: Architecture Details

nD-RAPID is defined by a four-tuple: (C, L, B, D), where C is the total number of clusters in the system, L is the total number of levels in a cluster, B is the total number of boards in a level, and D is the total number of nodes on a board. Each node in the system is defined as $R(c, l, b, d)$ such that $0 \leq c \leq C-1$, $0 \leq l \leq L-1$, $0 \leq b \leq B-1$, and $0 \leq d \leq D-1$. (Upper case indicates total number of cluster, level, board, and node numbers. Lower case identifies the individual node number.) The total number of nodes in the system is a multiplicative factor $N_{\text{total}} = C \times L \times B \times D$.

Fig. 1(a) and (b) shows nD-RAPID architecture for $n = 1$. In Fig. 1(a), each node contains the processing core, L1 and L2 caches, main memory, the network interface, and I/O. Nodes 0 to $D-1$ are connected using an intraboard interconnect to form a board. Boards 0 to $B-1$ are connected to each other via scalable interboard interconnect (SIBI) to form 1-D-RAPID. Fig. 1(b) shows a conceptual diagram of 1-D-RAPID. Though it seems as if all the boards are connected to one another using a ring, the interconnect is actually a point-to-point network. The 1-D-RAPID has a single level and a single cluster (i.e., $L = 1$ and $C = 1$).

As 1-D-RAPID extends in a single dimension [from Fig. 1(b)], the interconnect will be referred to as SIBI-x. Fig. 2(a) shows nD-RAPID for $n = 2$. In Fig. 2(a), 1-D-RAPID system (SIBI-x) is duplicated L times, and connected along the y-dimension to obtain 2-D-RAPID. All boards with the same x coordinate are connected to one another by means of a scalable interboard interconnect (SIBI-y). For example, board 0 of every level is connected to the remaining board 0s through SIBI-y. The 2-D-RAPID has multiple levels, but only a single cluster. Fig. 2(b) shows nD-RAPID for $n = 3$. The 2-D-RAPID structure is duplicated C times to form 3-D-RAPID. The boards are then connected to one another along the z-dimension using a scalable interboard interconnect (SIBI-z). In general, any two boards in 3-D-RAPID are directly connected to one another if exactly two of the three tuples B, L, and C are equal.

A. Intraboard Communication

The network interface at each node consists of send and receive ports, which are connected to optical transmitter and re-

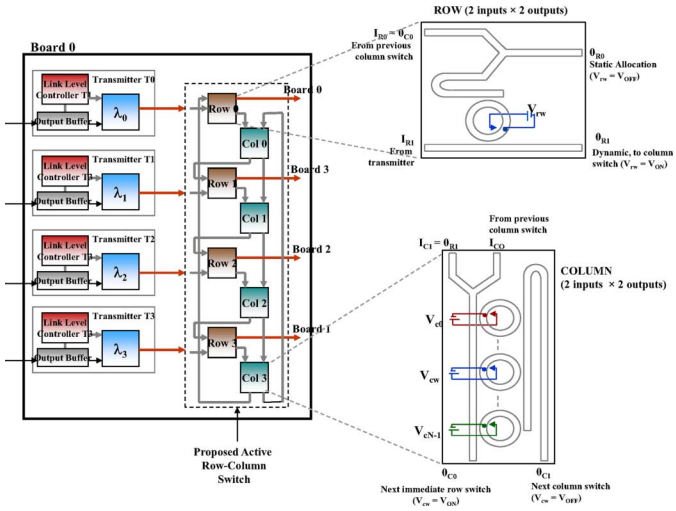

Fig. 3. Active row–column switch with four boards in the x-dimension ($B = 4$).

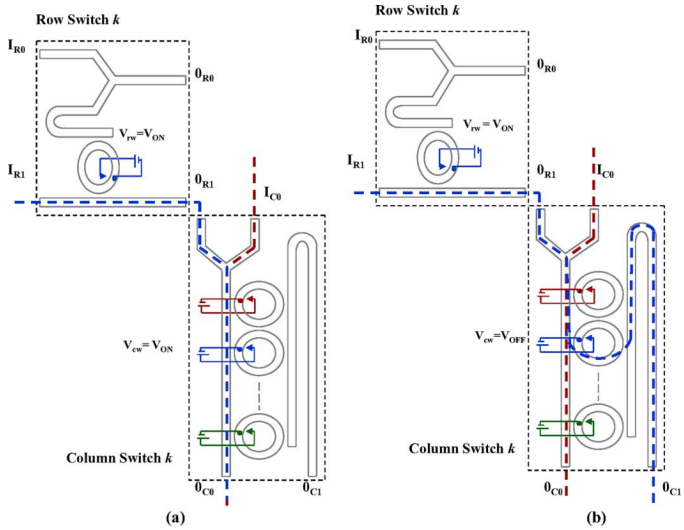

Fig. 4. (a) The output from the kth switch will be the immediate next $(k+1)$ row switch. (b) The output from the kth switch will be the subsequent $(k+i)$ row switch.

ceiver ports by means of a bidirectional crossbar. A separate set of transmitters/receivers exists for x-, y-, and z-directions, respectively, enabling the same set of wavelengths to be used in each direction. When a node needs to communicate it uses the send port to transmit over the crossbar to either the receive ports for intraboard communication or to the optical transmitters for interboard communication. Similarly, an optical receiver after receiving a packet can either send it to a receive port if the packet is at the destination board or send it to an optical transmitter for further interboard transmission.

B. Interboard Communication

We first explain routing in 1-D-RAPID and then extend it to nD-RAPID. In 1-D-RAPID, each board has a fiber associated with it known as the home channel for that board. Home channel is the waveguide on which other boards transmit optical packets to the destination board. A different wavelength from each board is merged into the home channel to provide high connectivity [15]. To illustrate with an example, consider routing along the x-dimension. The wavelength assigned from source board s to destination board d is given by $\lambda^{(s)}_{B-(d-s)}$ if $d > s$ and $\lambda^{(s)}_{(s-d)}$ if $s > d$, the superscript indicates the source board, and the subscript indicates the wavelength to be transmitted on [15]. For nD-RAPID, each board has a home channel in every dimension. A similar approach is followed when communication is required between boards that lie on different levels (y-dimension) or clusters (z-dimension), except in each case the wavelength to be used for communication will be determined based on the relative position of the cluster number c or the level l.

There are two reasons why nD-RAPID $(n > 1)$ is more advantageous than 1-D-RAPID. 1) nD-RAPID offers simpler design complexity that lowers the cost of the network and enables scalable design. In 1-D-RAPID, the number of lasers (wavelengths) increases with the number of boards. For example, a 64-node 1-D-RAPID will require 15 lasers per board (16 boards, four nodes/board) whereas 2-D-RAPID will require six lasers per board and 3-D-RAPID will require five lasers per board. Similarly, for a 256-node architecture, 1-D-RAPID will require 63 lasers (64 boards, four nodes/board), 2-D-RAPID will require 14 lasers per board, and 3-D-RAPID will require nine lasers per board. Moreover, the internal crossbar size which connects nodes to optical transceivers will also increase in size. 2) In 1-D-RAPID, if a single link failed (home channel of a system board), then other system boards will not be able to communicate on that link. This will completely isolate the board whose home channel has failed, leading to a catastrophic failure. However, by providing scalability in multiple dimensions $(n > 1)$, we improve the reliability of the interconnect as every board can be reached by transmitting packets in any of the three interconnects: SIBI-x, SIBI-y, or SIBI-z. Therefore, if a link should fail in one dimension, we can communicate to the system board in another dimension.

III. DBR Using Active Optical Switching

Although static allocation ensures fairness and is suitable for uniform traffic pattern [15], it can lead to network congestion for nonuniform traffic, particularly in case of faults or bursty communications. On the other hand, dynamic allocation/reallocation of channels in response to actual network load can lead to improved performance for most communication patterns. DBR has previously been implemented in E-RAPID (1-D-RAPID) using only passive elements [couplers and arrayed waveguide gratings (AWGs)] and a history-based lock-step (LS) algorithm [15]. Although the LS algorithm showed substantial improvement in throughput and latency, passive implementation required the number of transmitter per board to scale as $O(B^2)$, where B is the number of boards along one dimension, making it cost prohibitive for multiple dimensions proposed here.

In this section, we propose a new method to implement DBR using an active optical switch. The proposed active row–column switch shows similar performance as the passive implementation [15], but reduces the cost considerably (it scales as $O(B)$ along the x-dimension, $O(L)$ along the y-dimension, and $O(C)$ along the z-dimension). The row–column switch and LS algorithm are combined to implement DBR in nD-RAPID. In this

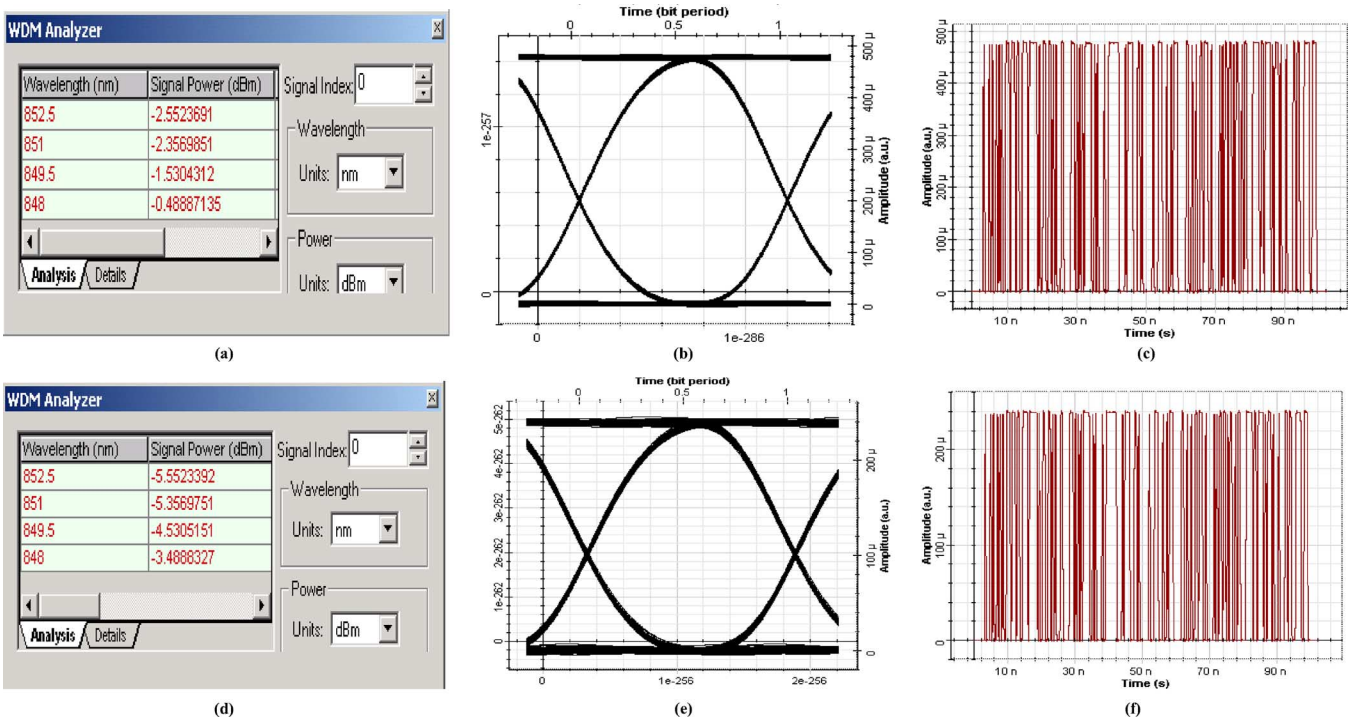

Fig. 5. Signal power for a four-channel system at the after multiplexing, the eye diagram, and received data for (a)–(c) optical backplane implementation and (d)–(f) optical fiber implementation.

section, the switch design is explained with respect to communication along the x-dimension (i.e., indexed with respect to B). The switch design for the other dimensions will be the same, but will be indexed with respect to L and C for the y- and z-dimensions, respectively.

The proposed switch is composed of wavelength-selective 1×2 SOI-based microring resonators. A microring resonator will couple light through it only if it satisfies the relation: $\lambda \times m = n_{\text{eff}} \times 2\pi R$, where R is the radius of the microring resonator, n_{eff} is the effective refractive index, and m is an integer. λ is then known as the resonant wavelength. By changing n_{eff}, the resonant wavelength of the microring resonator can be changed, enabling it to function as an optical switch [17].

The proposed switch consists of row–column switches as shown in Fig. 3. The row and column switches are themselves 2×2 switches. The row switch k has the transmitter and the previous $k-1$ column switch as its input, and has one of its outputs connected to the corresponding column switch, while the other output is connected to the optical SIBI for interboard communication. This is the release route for interboard communication. The column switch has one of its inputs from the corresponding row switch and the second input will be from the previous $k-1$ column switch. The outputs of the column switch are the $k+1$ row and column switch as shown in Fig. 3. The detailed implementation of the proposed row–column switch is shown in the inset for both row and column switches.

From the inset of Fig. 3, the input to the row switch IR_1 is from the transmitter connected to the corresponding row switch. This signal can be switched in two ways by the control voltage V_{rw}, either statically to the destination board as described in the static RWA algorithm ($V_{rw} = V_{\text{OFF}}$), or dynamically to any other destination board ($V_{rw} = V_{\text{ON}}$). Under static control, the signal appears at OR_0 for interboard communication. Setting the dynamic connection ($V_{rw} = V_{\text{ON}}$) causes the signal to be combined with the signal appearing from the previous column switch ($OR_1 = IC_1$). The switching activity at column switch for wavelength w is controlled by the voltage V_{cw}, and depends primarily on whether the signal is dropped at the next row switch or continues further along the column switches. If $V_{cw} = V_{\text{OFF}}$, then the signal couples to the next column switch and appears at OC_1. If $V_{cw} = V_{\text{ON}}$, then the signal couples to the next row switch and appears at OC_0. This signal then exits the row–column switch matrix through the OR_0. By suitably controlling the applied voltages, V_{rw} (row voltage corresponding to wavelength w), and V_{c0}, $V_{c1}, \ldots, V_{cw}$, $V_{cw+1}, \ldots, V_{cB-1}$ (column voltages corresponding to the entire wavelength set, $\Lambda = \lambda_0, \lambda_1, \ldots, \lambda_{B-1}$) to either V_{ON} or V_{OFF}, we can perform any-to-any switching (one-to-one, many-to-one, all-to-one) of all the wavelengths.

Each optical signal enters and exits the row–column switching matrix through the row switch. The column switch is used for routing the optical signal through the row–column switching matrix. To illustrate with an example, consider Fig. 4(a). Fig. 7(a) shows a snapshot of the row–column switching where the desired output for the optical signal from the kth transmitter will be the immediate next $(k+1)$ row switch. The other optical signal (from I_{C0}) being switched by the column switch will be for subsequent $(k+i)$ switches. Fig. 4(b) shows a snapshot of the row–column switching where the desired output for the optical signal will be the subsequent $(k+i)$ row switch. In this way, multiple wavelengths from different transmitters can be routed across the row–column matrix to achieve the desired reconfiguration.

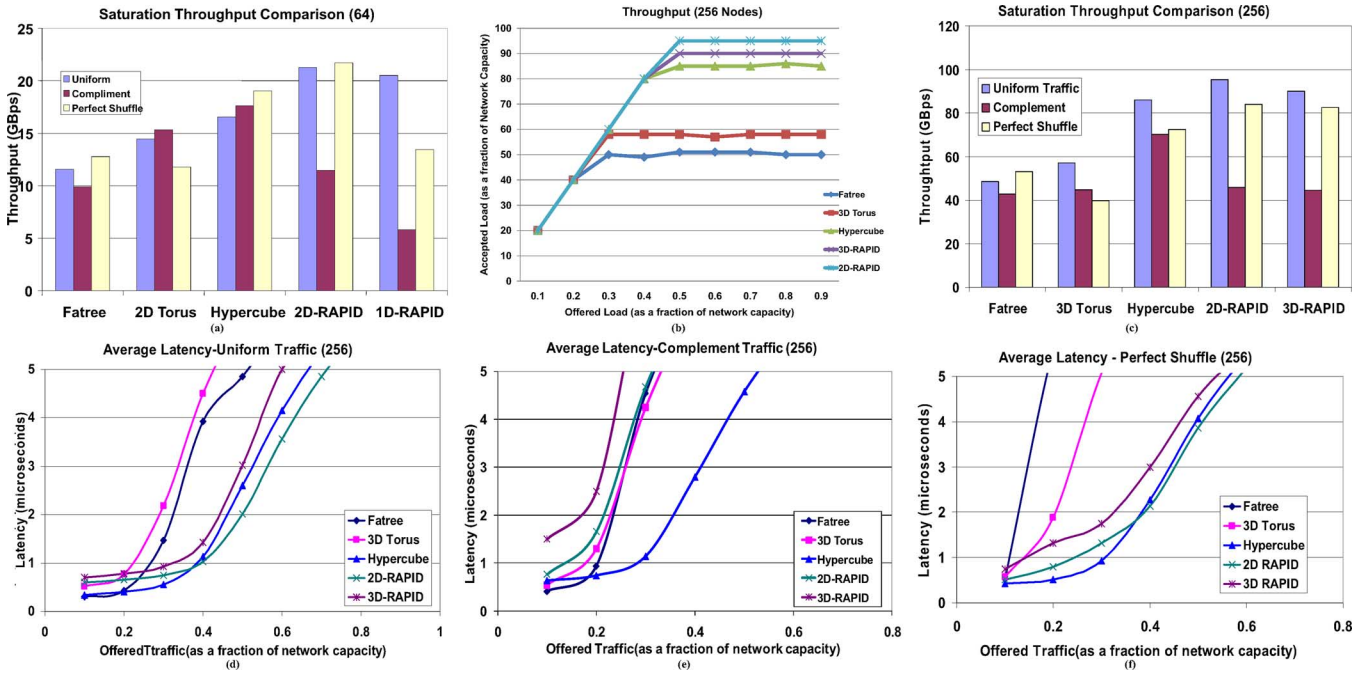

Fig. 6. (a) Throughput for 64 nodes comparing 1-D-RAPID, 2-D-RAPID, and electrical networks. (b) Throughput comparison for 256 nodes (2-D-RAPID, 3-D-RAPID, and electrical networks) for uniform traffic. (c) Saturation throughput comparison for uniform, complement, and perfect shuffle. (d)–(f) Latency comparison for 256 nodes for uniform, complement, and perfect shuffle, respectively.

IV. Optical Implementation

This section discusses optical implementation of baseline nD-RAPID architecture. This section is divided into three subsections: optical components, integration methodology, and system design verification using OptiSystem design tool.[1]

A. Optical Components

1) Laser Source and Receiver: Vertical cavity surface emitting lasers (VCSELs) are a natural candidate owing to their ease of fabrication in 1-D and 2-D arrays, good optical coupling to fibers and low cost [18]. With the increasing availability of commercial high-speed and high-power VCSELs arrays, they are the laser source incorporated for nD-RAPID. P-I-N photodiode arrays are used to convert the optical signals back to electrical signals. Commercially available photodiodes (for example, from Albis Optolectronics, Zürich, Switzerland) have an active area of 70 μm and a sensitivity of 0.5 A/W at 850 nm [18].

2) Waveguides: CMOS compatible waveguides can me made of high-index cores such as Si or low-index cores such as polymers. High-index core offers a smaller waveguide pitch whereas low-index core offers a lower propagation delay. As a result of the above tradeoff, polymer core waveguides allow for comparatively slower transmitters and receivers but require aggressive wavelength division multiplexing (WDM). Silicon waveguides, on the other hand, allow the WDM parameters to be relaxed, but require faster transmitters and receivers [19]. We will follow current trends of using polymer waveguides for integrated optical backplanes [18], [20] and SOI-based waveguides for on-chip applications [21].

[1]http://www.optiwave.com

3) Demultiplexers: Demultiplexers are used to filter out different wavelengths from a WDM signal. In AWGs, 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 waveguides is chosen such that the optical path length difference ∇L between adjacent waveguides equals an integer multiple of the central wavelength of the demultiplexer. The resulting phase difference at the waveguide exit is given by: $n_c \nabla L = m\lambda_0$, where n_c is the effective refractive index of the arrayed waveguide, m is the diffraction order, and λ_0 is the central wavelength. Recent studies have shown integrated AWGs using SOI technology [22], [23] with very small area (< 0.6 mm^2 with crosstalk suppression of over 16 dB) are available.

B. Optical Integration Methodology

1) VCSEL/Photodiode: The most popular integration methodology is to flip-chip bond the VCSEL/photodiode to the driver or receiver circuitry. The devices are designed to be backside emitting because of the desire to flip-chip bond them to the CMOS circuit. The n- and p-contact should then be on the top surface of the wafer to facilitate electrical connectivity to the circuit. The substrate can then be selectively etched leaving the VCSEL/PD contact pads on the backside of the wafer and all optical source/detectors on the other side of the wafer [24], [25].

2) Transmission Medium: The transmission of optical signals can be carried out depending on the distance of communication.

a) *Optical Backplane*: An emerging area of research is to design optical backplane to transmit and multiplex the

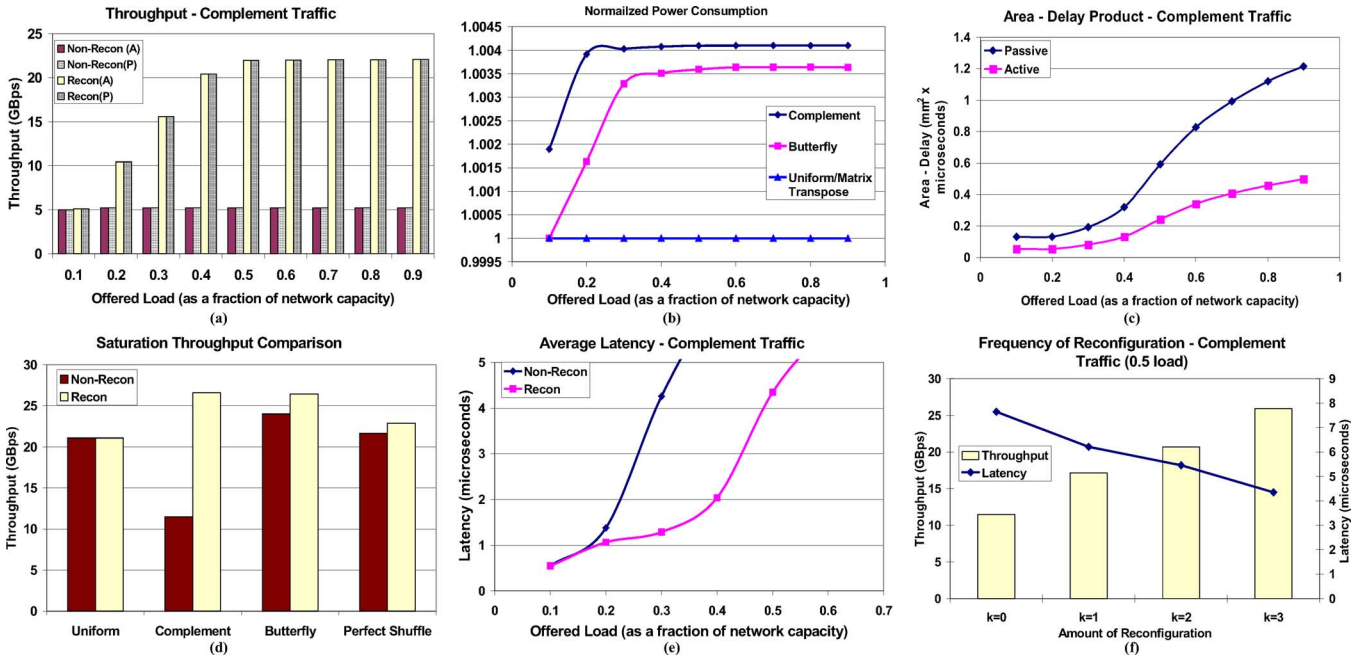

Fig. 7. The 64-node 1-D-RAPID. (a)–(c) Active versus passive implementation of DBR in 64-node 1-D-RAPID, showing (a) throughput, (b) normalized power consumption, and (c) area-delay product (ADP). (d)–(f) Performance evaluation of 2-D-RAPID with DBR showing (d) throughput, (e) latency, and (f) amount of reconfiguration.

signal at short distances. The backplanes use polymer waveguides [18], [20] on standard printed circuit board (PCB) substrates. The polymers generally have losses in the order of 0.05 dB/cm [20] to less than 0.03 dB/cm [18]. The signal can be multiplexed using low-loss directional couplers as explained above. MTP connectors are being used to connect jumpers from the board to the backplane [18], [20]. A common method to incorporate the 90° bends is to use 45° mirrors. The losses in the system stem from MTP connector (0.5 dB), 45° mirror reflection loss (0.5 dB), waveguide loss 0.05 dB/cm, directional coupler loss (1 dB), and AWG loss (3 dB).

b) *Optical Fibers*: Optical fibers have for long been for longer distance and the technology associated with them is quite mature. Standard MT Ferrule connectors are available to couple light from the source into the fiber.

C. Validation and OptiSystem Simulation Results

In this section, OptiSystem was used to validate a four-channel system. We evaluated both the optical backplane and the optical fiber methodologies. The lazing channels were 852.5, 851, 849.5, and 848 nm and the input power was determined to be 2 mW (3 dBm) at 2.5 Gb/s.

1) Optical Backplane: In case of the optical backplane, the worst case power loss can be calculated using

$$\begin{aligned}\text{Loss (dB)} = {} & 0.5\,(\text{MTP loss}) + 0.5\,(\text{reflection loss}) \\ & + 0.05 \times L\ (\text{propagation loss}) \\ & + 1 \times (B-1)\ (\text{directional coupler loss}) \\ & + 0.05\ (\text{reflection loss}) + 0.5\ (\text{MTP loss}) \\ & + 3\,\text{dB}\ (\text{AWG loss})\end{aligned}$$

where L is the maximum length traveled in centimeters. Fig. 5(a) shows the strength of the multiplexed signal at the destination board, Fig. 5(b) shows the eye diagram, and Fig. 5(c) shows the received signal. The eye diagram shows a height of 2.29×10^{-5}, threshold of 2.93×10^{-6}, and a low bit error rate (BER).

2) Optical Fiber: In case of fiber implementation, the worst case loss can be calculated using

$$\begin{aligned}\text{Loss (dB)} = {} & 0.5\ (\text{MTP loss}) + 0.02 \times L'(\text{propagation loss}) \\ & + 3 \times \log_2 B\ (\text{tree coupler loss}) \\ & + 0.5\ (\text{MTP loss}) + 3\,\text{dB}\ (\text{AWG loss})\end{aligned}$$

where L' is the maximum length traveled in kilometers. Fig. 5(e) shows the strength of the multiplexed signal at the destination board, Fig. 5(f) shows the eye diagram, and Fig. 5(g) shows the received signal. The eye diagram shows a height of 2.29×10^{-5}, threshold of 2.93×10^{-6}, and a low BER.

V. Performance Evaluation

We used OPTISIM [26], a discrete event simulator to evaluate the performance of nD-RAPID. Its performance is compared with various electrical networks for uniform and permutation traffic traces. The electrical networks chosen for comparison were 2-D torus, 3-D torus, hypercube, and fat tree. The hypercube network is used in the SGI Spider chip and SGI origin machines [27] and the fat-tree topology is the basis of most Mellanox switches used in Infiniband, Elan 2, and QsNet.

A. Simulation Methodology

Cycle accurate simulations were used to evaluate the performance of nD-RAPID. For the router model designed, the channel width is 16 bits and speed is 400 MHz, resulting in

a unidirectional bandwidth of 6.4 Gb/s and per-port bidirectional bandwidth of 12.8 Gb/s. Routing computation, virtual channel, and switch allocation, each takes one router clock cycle. At 5 Gb/s, the total power consumption of an optical link to transmit and receive a 64-B packet, operating at a supply voltage of 0.9 V, is 43.03 mW [15]. Traffic patterns commonly found in scientific applications were used to measure the performance. We used uniform traffic, butterfly, complement, and perfect shuffle traffic patterns. The performance of the networks was measured in terms of throughput and latency.

B. nD-RAPID Results

This section discusses the performance of nD-RAPID when compared with similarly sized electrical networks. Due to space constraints, only a few results are shown here. For 64-node systems, we compare 1-D and 2-D-RAPID with electrical networks. For 256-node systems, we compared 2-D and 3-D-RAPID with fat-tree, 3-D torus, and hypercube. Fig. 6(a) shows the saturation throughput for 64 nodes; 2-D-RAPID performs better than 1-D-RAPID due to more bandwidth available in two dimensions. Fig. 6(b) shows the throughput (accepted load versus offered load) for 256 cores. Fig. 6(c) shows the throughput comparison for uniform, complement, and perfect shuffle traffic patterns for 256 nodes. As predicted earlier, 2-D-RAPID shows higher throughput at saturation than 3-D-RAPID. This is because the aggregate optical bandwidth per board for 2-D-RAPID (16 lasers per board) is higher than that of 3-D-RAPID (12 lasers per board). Both 2-D and 3-D-RAPID show improved performance over electrical networks for both uniform and perfect shuffle but not for complement traffic. We will show in the next section how DBR improves performance for adverse traffic patterns such as complement; 2-D-RAPID shows a 13.86% improvement over the hypercube for perfect shuffle traffic pattern. Fig. 6(d)-(f) shows the latency comparisons for uniform, complement, and perfect shuffle traffic patterns, respectively. For complement traffic, 2-D and 3-D-RAPID do not perform as well as the electrical networks. In complement traffic, the destination node is determined by complementing all the bits of the source node. For example, if the source node is 5 (000101), the destination node is 58 (111010) for a 64-node system. The contention for a given wavelength is the same for 2-D and 3-D-RAPID as both have four nodes per board. As 3-D-RAPID would require routing in one extra dimension, the latency for 3-D-RAPID is more.

C. DBR Throughput and Latency Results

Fig. 7(a) shows the effect of DBR for complement traffic for a 64-node 1-D-RAPID system. In Fig. 7(a), Recon implies that DBR was implemented, non-Recon implies DBR was not implemented, A stands for active switch implementation (row–column switch proposed in this paper), and P stands for passive implementation [15]. Due to the nature of complement traffic all the nodes on a source board communicate with the same destination board. As a result, without bandwidth reallocation, only one transmitter is active per board resulting in high latency and low throughput. Therefore, reallocation starts at very low loads and the system is fully reconfigured at a load of 0.2. On reaching steady state, throughput is almost four times higher when DBR is implemented. It is important to note that both active and passive implementations show similar performance with and without reconfiguration. In fact, complement traffic ensures maximum possible reconfiguration in the system and thus shows the worst case power consumption and power loss for the active switch design. As the reconfiguration window (2000 cycles) is larger than the switching time (four cycles), the increase in latency is negligible.

D. Power Consumption

The total electrical power (P_T) consumed was calculated using the formula $P_T = \sum_{j=0}^{B} N_{Bj} \times P_{Tx/Rx} + \sum_{j=0}^{B} N_{Rj} \times P_{\text{ring}}$, where B is the total number of boards, N_{Bj} is the total number of optical packets transmitted by board j, $P_{Tx/Rx}$ is the electrical power to transmit and receive a single 64-B optical packet, N_{Rj} is the number of times a switch in the ON state is traversed by packets from board j, and P_{ring} is the electrical power consumed when a ring resonator is on. Theoretical calculations estimate the power consumption of a 5-μm radius microring resonator to be 19 μW [16]. Although current prototypes consume 1 mW, straightforward fabrication advances can reduce this value to 100 μW [17]. Therefore, we assume that each ring consumes 100 μW, i.e., $P_{\text{ring}} = 100\ \mu$W. Fig. 7(b) shows the normalized electrical power consumption for the four traffic traces (total electrical power consumed using active switch divided by total electrical power consumed using passive switch). As expected, for uniform and matrix transpose traffic patterns, there is no extra power consumption since no reconfiguration occurs. Complement traffic results in maximum power consumption, which is about 0.4% more than the passive case. Fig. 7(c) shows the ADP for complement traffic. We see that as the load increases, the advantage of using an active switch design increases.

Fig. 7(d) compares saturation throughput of 2-D-RAPID (64 node) for uniform, complement, butterfly, and perfect shuffle traffic traces. We again notice that maximum improvement is seen for complement traffic, with throughput increasing by more than a factor of 2. Fig. 7(e) compares the average latency for complement traffic with and without reconfiguration. As can be clearly seen, with reconfiguration, the system saturates at a much higher load. Fig. 7(f) shows higher throughput and latency as we change the amount of reconfiguration for complement traffic and for a network load of 0.5. In Fig. 7(f), $k = i$ implies i extra transmitters have been turned on. This means that the load can now be evenly distributed between $i + 1$ different transmitters. From the figure, throughput increases as the amount of reconfiguration increases. Average latency on the other hand decreases with reconfiguration due to the fact that a packet sees lower blocking and queuing delays.

VI. Conclusion

In this paper, we propose a multidimensional optoelectronic architecture nD-RAPID for HPC systems that improves fault tolerance and dynamic reconfigurability. While taking maximum advantage of what the high bandwidth optics has to offer, we reduce the network cost by building a scalable architecture where the optical active components are present only on the

board. The architecture is made fault tolerant by employing an n-dimensional $(n = 2, 3)$ structure, which ensures that there is more than one path to every board. Dynamic reconfigurability is achieved by rerouting packets when they come across a faulty link. In order to improve performance in the presence of adversial traffic patterns such as complement traffic, a compact, integratable, and nonblocking optical switch matrix for implementing dynamic bandwidth reallocation was proposed. The switch matrix was designed to reduce cost (in terms of number of lasers) while maintaining the performance benefits and flexibility shown by passive implementation of DBR. When nD-RAPID was compared with other popular networks for HPC systems, it was seen that it consistently outperformed them in most of the traffic patterns tested. Analytical and simulation studies further showed that the proposed active implementation is able to improve performance (throughput and latency) with minimal area, power, and hardware overheads. There is a slight increase in power consumption (0.4% at most for the worst case traffic) using the active switch matrix.

REFERENCES

[1] D. A. B. Miller, "Rationale and challenges for optical interconnects to electronic chips," *Proc. IEEE*, vol. 88, no. 6, pp. 728–749, Jun. 2000.

[2] R. G. Beausoleil, P. J. Kuekes, G. S. Snider, S.-Y. Wang, and R. S. Williams, "Nanoelectronic and nanophotonic interconnect," *Proc. IEEE*, vol. 96, no. 2, pp. 230–247, Feb. 2008.

[3] C. Batten, A. Joshi, J. Orcutt, A. Khilo, B. Moss, C. Holzwarth, M. Popovic, H. Li, H. Smith, J. Hoyt, F. Kartner, R. Ram, V. Stojanovic, and K. Asanovic, "Building manycore processor-to-dram networks with monolithic silicon photonics," in *Proc. 16th Annu. Symp. High-Performance Interconnects*, Aug. 27–28, 2008.

[4] N. Kirman, M. Kirman, R. K. Dokania, J. Martínez, A. B. Apsel, M. A. Watkins, and D. H. Albonesi, "Leveraging optical technology in future bus-based chip multiprocessors," in *Proc. 39th Int. Symp. Microarchitecture*, Dec. 2006.

[5] A. Shacham, K. Bergman, and L. P. Carloni, "On the design of a photonic network-on-chip," in *Proc. 1st Int. Symp. Network-on-Chip*, May 2007, pp. 53–64.

[6] D. Vantrease, R. Schreiber, M. Monchiero, M. McLaren, N. P. Jouppi, M. Fiorentino, A. Davis, N. Binkert, R. G. Beausoleil, and J. H. Ahn, "Corona: System implications of emerging nanophotonic technology," in *Proc. 35th Int. Symp. Comput. Architecture*, Jun. 2008.

[7] 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. Develop.*, vol. 49, no. 4/5, pp. 755–775, Sep. 2005.

[8] S. Banerjee and D. Sarkar, "Hypercube connected rings: A scalable and fault-tolerant logical topology for optical networks," *Comput. Commun.*, vol. 24, pp. 1060–1079, 2001.

[9] Y. Yang and J. Wang, "A fault-tolerant rearrangable permutation network," *IEEE Trans. Comput.*, vol. 53, no. 4, pp. 414–426, Apr. 2004.

[10] B. Helvik and R. Andreassen, "Fault tolerance in optical networks; a study of electronic in- and egress interconnections in torus topologies," in *Proc. 9th Conf. Opt. Network Design Model.*, 2005.

[11] A. K. Kodi and A. Louri, "Rapid: Reconfigurable and scalable all-photonic interconnect for distributed shared memory multiprocessors," *J. Lightw. Technol.*, vol. 22, no. 9, pp. 2101–2110, Sep. 2004.

[12] P. Dowd, J. Perreault, J. Chu, D. Crouse, D. Hoffmeister, R. Minnich, D. Burns, F. Hady, Y. J. Chen, M. Dagenais, and D. Stone, "Lightning network and systems architecture," *J. Lightw. Technol.*, vol. 14, no. 6, pp. 1371–1387, Jun. 1996.

[13] C. M. Qiao, R. Melham, D. Chiarulli, and S. Levitan, "Dynamic reconfiguration of optically interconnected networks with time-division multiplexing," *J. Parallel Distrib. Comput.*, vol. 22, no. 2, pp. 268–278, 1994.

[14] P. Krishnamurthy, R. Chamberlain, and M. Franklin, "Dynamic reconfiguration of an optical interconnect," in *Proc. 36th Annu. Simul. Symp.*, 2003.

[15] A. K. Kodi and A. Louri, "Performance adaptive power-aware reconfigurable optical interconnects for high-performance computing (HPC) systems," in *Int. Conf. High-Performance Comput. Netw. Storage Anal.*, Reno, NV, Nov. 10–16, 2007.

[16] Q. Xu, B. Schmidt, S. Pradhan, and M. Lipson, "Micrometre-scale silicon electro-optic modulator," *Nature Lett.*, vol. 435, pp. 325–327, 2005.

[17] B. A. Small, B. G. Lee, K. Bergman, Q. Xu, and M. Lipson, "Multiple-wavelength integrated photonic networks based on microring resonator devices," *J. Opt. Netw.*, vol. 6, no. 2, pp. 112–120, Feb. 2007.

[18] C. Berger, B. J. Offrein, and M. Schmatz, "Challenges for the introduction of board-level optical interconnect technology into product development roadmaps," in *Proc. SPIE—Int. Soc. Opt. Eng.*, Jan. 2006, pp. 61240.J1–61240.J12.

[19] M. Haurylau, G. Chen, H. Chen, J. Zhang, N. A. Nelson, D. H. Albonesi, E. G. Friedman, and P. M. Fauchet, "On-chip optical interconnect roadmap: Challenges and critical directions," *IEEE J. Sel. Top. Quantum Electron.*, vol. 12, no. 6, pp. 1699–1705, Nov./Dec. 2006.

[20] A. L. Glebov, M. G. Lee, and K. Yokouchi, "Integration technologies for pluggable backplane optical interconnect systems," *Opt. Eng.*, vol. 64, 2007.

[21] C. Gunn, "CMOS photonics for high speed interconnects," *IEEE Photon. Technol. Lett.*, vol. 26, no. 2, pp. 58–66, Mar./Apr. 2006.

[22] P. Dumon, W. Bogaerts, D. Van Thourhout, D. Taillaert, and R. Baets, "Compact wavelength router based on a silicon-on-insulator arrayed waveguide grating pigtalied to a fiber array," *Opt. Express*, vol. 14, no. 2, pp. 664–669, Jan. 2006.

[23] A. Huang, C. Gunn, G. Li, Y. Liang, S. Mirsaidi, A. Narasimha, and T. Pinguet, "10 Gb/s photonic modulator and WDM mux/demux integrated with electronics in 0.13 μm soi CMOS," in *Proc. IEEE Int. Solid-State Circuits Conf.*, Feb. 2006.

[24] 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.

[25] 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.

[26] A. Kodi and A. Louri, "A system simulation methodology of optical interconnects for high-performance computing (HPC) systems," *J. Opt. Netw.*, vol. 6, pp. 1282–1300, Dec. 2007.

[27] M. Galles, "Spider: A high-speed network interconnect," *IEEE Micro*, vol. 17, no. 1, pp. 34–39, Jan./Feb. 1997.

Avinash Karanth Kodi (M'07) received the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Arizona, Tucson in 2003 and 2006, respectively.

Currently, he is an Assistant Professor of Electrical Engineering and Computer Science at Ohio University, Athens. His research interests include computer architecture, optical interconnects, chip multiprocessors (CMPs), and network-on-chips (NoCs).

Ahmed Louri (S'86–M'88–SM'95) received the M.S. and Ph.D. degrees in computer engineering from the University of Southern California, Los Angeles, in 1984 and 1988, respectively.

Currently, he is a Full Professor of Electrical and Computer Engineering at the University of Arizona, Tucson. He is also the Director of the High-Performance Computing Architectures and Technologies (HPCAT) Laboratory. His research interests include computer architecture, parallel processing, optical interconnection networks, and network-on-chips (NoCs).

Prof. Louri is a regular member of OSA.