# A New Dynamic Bandwidth Re-Allocation Technique in Optically Interconnected High-Performance Computing Systems

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

## Abstract

*As bit rates increase, optical interconnects based high-performance computing (HPC) systems improve performance by increasing the available bandwidth (using wavelength-division multiplexing (WDM) and space-division multiplexing (SDM)) and decreasing power dissipation as compared to traditional electrical interconnects. While static allocation of wavelengths (channels) in optical interconnects provide every node with equal opportunity for communication, it can lead to network congestion for non-uniform traffic patterns. In this paper, we propose an opto-electronic interconnect for designing a flexible, high-bandwidth, low-latency, dynamically reconfigurable architecture for scalable HPC systems. Reconfigurability is realized by monitoring traffic intensities, and implementing dynamic bandwidth re-allocation (DBR) technique that adapts to changes in communication patterns. We propose a DBR technique - Lock-Step (LS) that balances the load on each communication channel based on past utilization. Simulation results indicate that the reconfigured architecture shows 40% increased throughput and and 20% reduced network latency as compared to HPC electrical networks.*

## 1. Introduction

In order to meet high bandwidth and low power requirements in scalable high-performance computing (HPC) systems, opto-electronic interconnects are becoming common at board-to-board and box-to-box communication distances, thereby overcoming some of the fundamental signalling limitations of current electrical interconnects[1, 2, 3]. Opto-electronic interconnects provide maximum flexibility for HPC systems by partitioning electronic processing functionalities with high bandwidth optical communication capabilities, thereby optimizing cost to performance ratio.

In our previously proposed optical interconnect called RAPID (Reconfigurable, All-Photonic Interconnect for Distributed and parallel systems)[4], the routing and wavelength 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 non-uniform communication patterns. On the other hand, dynamic re-allocation 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 electro-optic switching elements[5], time-slots based bandwidth re-allocation[6] and both time and space based bandwidth switching[7].

In this paper we propose a dynamically reconfigurable optical interconnect called E-RAPID (extended-RAPID) that achieves dynamic reconfiguration without using any active components. Instead, E-RAPID relies solely on passive optical components and arrays of multiple wavelength transmitters for designing the reconfigurable network. In addition, we propose a dynamic bandwidth re-allocation (DBR) technique called Lock-Step (LS) that adapts to changes in communication traffic on E-RAPID. LS is a history-based distributed reconfiguration algorithm that triggers reconfiguration phases, disseminates state information, re-allocates system bandwidth, and re-synchronizes the system periodically with low control overhead to achieve optimized resource utilization. LS has several advantages including: (1) LS is completely decentralized such that every board independently makes re-allocation decisions. (2) Re-allocation of bandwidth could happen between any system boards without affecting the on-going communication in the overall system, and (3) Maximum bandwidth can be provided for system boards

Proceedings of the 14th IEEE Symposium on High-Performance Interconnects (HOTI'06)
0-7695-2654-3/06 $20.00 © 2006 IEEE

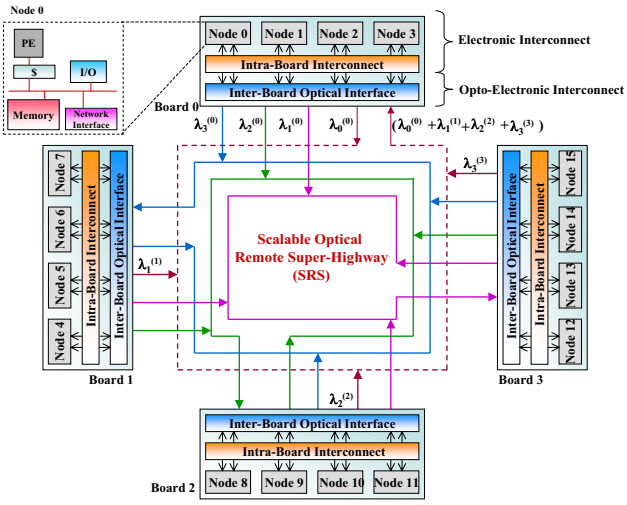


**Figure 1. Routing and wavelength assignment in E-RAPID for inter-board communication.**

for gather/scatter or hot-spot/bursty traffic pattern, where extremely high load is placed for a short duration of time. The proposed DBR technique results in reduced communication bottlenecks and optimized resource utilization leading to *balanced-improved* system architecture design. The proposed E-RAPID and LS are discussed in the next sections.

## 2. Optical Reconfigurable Architecture: E-RAPID

A E-RAPID network is defined by a 3-tuple:(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. Figure 1 shows an E-RAPID system with C = 1, B = 4 and D = 4. All nodes are connected to the scalable electrical Intra-Board Interconnect (IBI). The IBI connects the nodes for local (intra-board communication) as well as to the Scalable Remote Optical Super-Highway (SRS) for remote (inter-board communication). All interconnects on the board are implemented using electrical interconnects, where as the interconnections from the board to SRS are implemented using optical fibers using multiplexers and demultiplexers. The WDM and SDM features are exploited by the SRS for maximizing the inter-board connectivity as explained next.

**Inter-board and Intra-board Communication:** The static routing and wavelength allocation (RWA) for inter-board communication for a R(1,4,4) system is shown in Figure 1. For inter-board communication, different wavelengths from various boards are selectively merged to separate channels to provide high connectivity. Inter-board wavelengths are indicated by $\lambda_i^{(s)}$, where $i$ is the wavelength and $s$ is the source board number from which the wavelength originates. 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[4]. For example, if any node on board 1 needs to communicate with any node in board 0, the wavelength used is $\lambda_1^{(1)}$ and for reverse communication, the wavelength used is $\lambda_3^{(0)}$. The multiplexed signal received at the board is demultiplexed such that every optical receiver detects a wavelength.

Figure 2 shows the intra-board interconnections for board 0 of Figure 1. The network interface at every node is composed of send and receive ports. These send and receive ports at each node are connected to the optical transmitter and receiver ports through the bidirectional switch. Each packet, consisting of several fixed-size units called flits, that arrives on the physical input buffers progress through various stages in the router before it is delivered to the appropriate output port. The progression of the packet can be split into *per-packet* and *per-flit* steps. The per-packet steps include route computation (RC), virtual-channel allocation (VA) and per-flit steps include switch allocation (SA) and switch traversal (ST)[8]. A link controller (LC) is associated with each optical transmitter and receiver and a Reconfiguration Controller (RC) is associated with each system board. The co-ordination between RCs and LCs are essential for implementing the reconfiguration algorithm.

One significant distinction should be made in E-RAPID: Flits from different nodes are interleaved in the electrical domain using virtual channels whereas packets from different boards are interleaved in the optical domain. Although flit transmission in the optical domain is feasible, flit management across multiple domains is extremely complicated.

**Technology for Reconfiguration:** From Figure 2, each optical transmitter is composed of an array of vertical-cavity surface emitting lasers (VCSELs). The enabling technology for reconfigurability in E-RAPID is shown in Figure 3. Each optical transmitter is associated with 4 output ports (a, b, c and d) as there are 4 boards in the system. The notation $\lambda_x^{(y)}$ is used here to indicate wavelength $x$ originating from port $y$ for a given transmitter. The statically assigned wavelengths for inter-board communication are enclosed in a bracket.

The ability to dynamically switch multiple wavelengths through different ports of a given transmitter simultaneously to different system boards using passive couplers forms the basis for system reconfigurability in E-RAPID. This provides the flexibility in E-RAPID where more than one wavelength can be used for board-to-board communications in

Proceedings of the 14th IEEE Symposium on High-Performance Interconnects (HOTI'06)
0-7695-2654-3/06 $20.00 © 2006 IEEE

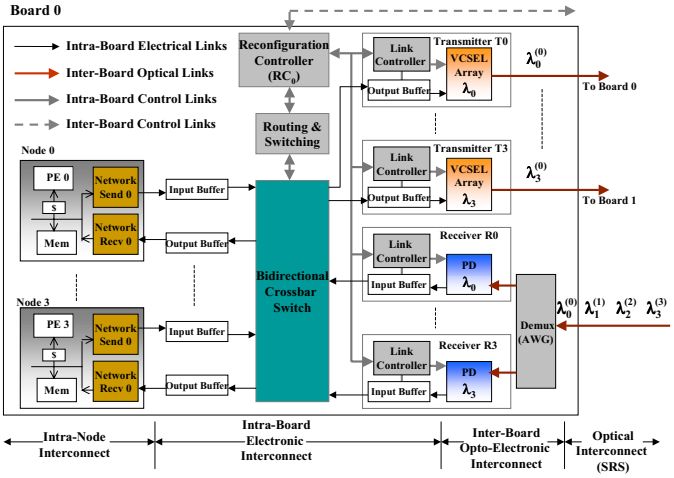


**Figure 2. E-RAPID architecture.**

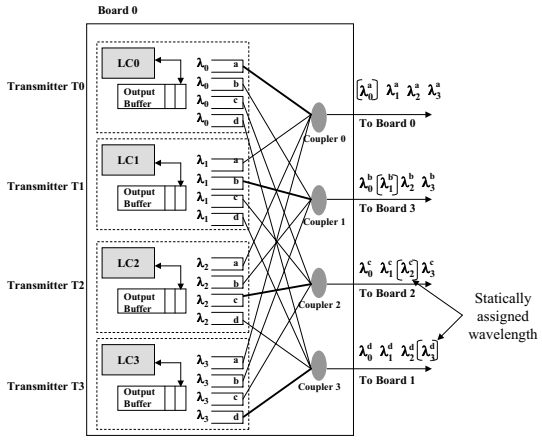


**Figure 3. Technology for reconfiguration.**

case of increased traffic loads. The basis of reconfiguration is to combine, at a given coupler, different wavelengths from similar numbered ports, but from different transmitters. Referring to Figure 3, the multiplexed signal appearing at coupler 1 is composed of all the signals inserted by same numbered $b$ ports ($\lambda_0^{(b)}$, $\lambda_1^{(b)}$, $\lambda_2^{(b)}$ and $\lambda_3^{(b)}$), but from different transmitters. Now, when needed, different destination boards can be reached by more than one static wavelength, thereby enabling the dynamic reconfigurability of the proposed architecture. For example, assume that the traffic intensity from board 0 to 2 is high. The static wavelength assigned for communication to board 0 to 2 is $\lambda_2^{(c)}$ at coupler 2. The other wavelengths $\lambda_0^{(c)}$, $\lambda_1^{(c)}$ and $\lambda_3^{(c)}$ appearing at the same coupler 2, could be used if other boards (board 1, 2 or 3) release their statically allocated wavelengths (with which they can communicate with board 2) to board 0. If board 1 releases wavelength $\lambda_1$ to board 0, then board 0 can start using port $c$ at transmitter 1 ($\lambda_1^{(c)}$) in addition to port $c$ at transmitter 2 ($\lambda_2^{(c)}$), thereby doubling the bandwidth and reducing communication latency. The physical link over which both the wavelengths $\lambda_1^{(c)}$, and $\lambda_2^{(c)}$ propagate are the same, where as the different channel is formed between transmitters 1 and 2 at board 0 with different receivers on board 2. This allows contending traffic, not only to use multiple wavelengths, but also to spread the traffic on the transmitter board, thereby increasing the throughput of the network.

## 2.1. Dynamic Reconfiguration Algorithm: Lock-step (LS)

In order to achieve (DBR), a new technique called Lock-step (LS) is proposed. LS re-allocates wavelengths associated with idle channels to busy channels based on historical information. In LS, each reconfiguration phase works in several circular stages, each stage is implemented either as a request or a response stage between RC and LC. Each RC triggers the reconfiguration phase, communicates with the local LCs and other RCs to determine the network load based on state information (link and buffer utilizations) collected during the previous phase. RCs evaluate the state information and re-allocate the bandwidth for the current phase based on previous phase reconfiguration statistics. After RCs have decided which links to reconfigure, this information is disseminated back to the RCs on other boards as well as the local LCs. The key requirement of LS is to minimize the impact of reconfiguration latency on the on-going communication in the network. In addition, the time to reconfigure should also be minimized so that the reconfiguration algorithm is responsive to transient traffic changes.

**Reconfiguration Statistics:** Historical statistics are collected with the hardware counters located at each LC. While each LC is associated with an optical transmitter and receiver, in this paper we consider the LC implementation with respect to the transmitter. The link utilization $\text{Link}_{util}$ tracks the percentage of router clock cycles where a packet is being transmitted in the optical domain from the transmitter queue. The buffer utilization $\text{Buffer}_{util}$ determines the percentage of buffers being utilized before the packet is transmitted[9]. All these statistics are measured over a sampling time window called *Reconfiguration window* or phase, $R_w$. $\text{Link}_{util}$ provides accurate information regarding whether a link is being used at all, at low-medium network loads, where as $\text{Buffer}_{util}$ provides accurate information regarding network congestion at medium-high network load.

**Reconfiguration Implementation:** Each $RC_i$ is connected

Proceedings of the 14th IEEE Symposium on High-Performance Interconnects (HOTI'06)
0-7695-2654-3/06 $20.00 © 2006 IEEE

to $RC_{i+1}$ in a simple electrical ring topology separated from the optical SRS. A ring topology with unidirectional flow of control ensures that what information is sent in one direction is always received in another. Figure 4 shows the 2 communication stages, RC-LC and RC-RC of the reconfiguration implementation. Figure 4 shows the RC, with RC transmit/receiver ports, LC transmit/receive ports, an RC queue, an outgoing link statistic and an incoming link statistic table. Each transmitter associated with every wavelength $\lambda_0$, $\lambda_1$, $\lambda_2$ ... on a given system board has a on/off value. This binary value indicates which lasers within a transmitter are either on (1) or off (0).

The symmetry of E-RAPID with respect to the number of wavelengths provides the insight into reconfiguration algorithm. For example, if $\Lambda = \lambda_0$, $\lambda_1$, $\lambda_2$ ... $\lambda_{W-1}$ is the total number of wavelengths associated with the system, we can see that this is exactly the number of wavelengths transmitted/received from each system boards. In other words, the number of *outgoing* or *incoming* links per system board is the same. Therefore, in order to balance the load and reallocate wavelengths on any given link, the system board needs all link statistics on its $incoming$ links. This is achieved by the co-ordination between the LCs and RCs as explained in the 5 stage reconfiguration mechanism for a R(1, 4, 4) system. Figure 4(a) shows the RC-LC communication used for Link Request and Link Response stages and Figure 4(b) shows the RC-RC communication used for Board Request and Board Response stages.

**Link Request Stage**: From Figure 4(a), at each board, $RC_i$, ($i = 0, 1, ...3$) sends out $Link_{Request}$ packets to the each of the LCs, $LC_0$, $LC_1$, ... $LC_3$ sequentially at the beginning of the reconfiguration phase. Each $LC_i$ updates the queue statistics $\text{Link}_{util}$, and $\text{Buffer}_{util}$, and forwards the packet to the next $LC_{i+1}$. When this packet is received by the $RC_i$, it updates all the *outgoing* link statistics.

**Board Request Stage**: From Figure 4(b), each $RC_i$ now sends out $\text{Board}_{Request}$ for all its *incoming* link information (shown in straight line). As it sends out, due to the symmetry of the ring architecture, it receives $\text{Board}_{Request}$ from other $RC_i$ (shown in dotted lines). For example, when board 0 receives $\text{Board}_{Request}$ from say board 1, it will update the field for wavelength with which board 0 communicates with board 1, i.e. $\lambda_3$ using the data stored in its $outgoing$ link statistic. When the board $RC_i$ receives its own $\text{Board}_{Request}$ packet, it updates all the incoming link statistics.

**Reconfigure Stage**: Now, each $RC_i$ computes if reconfiguration is necessary based on two thresholds, maximum threshold $T_{max}$ and minimum threshold $T_{min}$. It has been documented that the network is congested if the buffer threshold exceeds 0.5[9]. While profiling of traffic traces can provide more accurate information regarding when the network is actually congested, setting the $T_{max}$ to 0.5 is fairly reasonable for most traffic scenarios. This implies that on an average 50% of our buffers are occupied by packets for the given reconfiguration window $R_w$. We set $T_{min}$ to 0.0 which indicates no packets are queued. Each incoming link statistic is classified into three categories using $Buffer_{util}$ as under-utilized (implying that this wavelength can be re-allocated), normal utilized (implying the wavelength is well utilized) and over-utilized (implying that additional wavelengths are needed). RC would allocate the under-utilized links to the over-utilized links. In this way load can be balanced on all the links $incoming$ on a given system board.

**Board Response Stage**: From Figure 4(b), each $RC_i$ now sends out $Board_{Response}$ to all the remaining board $RCs$ to update their outgoing link statistics. As in board request stage, $RC_i$ updates the information received from other $RCs$ for the transmitters with which $RC_i$ communicates with those boards into its $outgoing$ link statistics.

**Link Response Stage**: From Figure 4(a), each board $RC_i$ sends out $Link_{Response}$ packets using the data received from its outgoing link statistics to each of the $LC_i$. Each $LC_i$ updates the state information received, thereby either turning on/off the lasers.

The entire protocol works in *lock-step* fashion, i.e. as a new control packet is transmitted by the $RC_{i+1}$, it receives a control packet from the previous $RC_i$. This provides synchronization as the $RC_{i+1}$ will not service the newly received control packet from $RC_i$ until it transmits its own control packet.

## 3. Performance Evaluation

The performance of E-RAPID is evaluated using YACSIM and NETSIM[10] discrete-event simulator and is compared to various electrical interconnects for both uniform and non-uniform traffic traces[11]. The most common electrical networks for clustering such as the 2D-torus, hypercube and fatree topologies were chosen for comparison.

**Simulation Methodology:** We use cycle accurate simulations to evaluate the performance of E-RAPID and other electrical interconnects. Packets were injected according to Bernoulli process based on the network load for a given simulation run. The network load is varied from $0.1 - 0.9$ 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[8]. The electrical network router model parameters are similar to the SGI Spider routing chip. For the router model designed, the channel width and the phit size is 16 bits, flit size is 64 bits and channel speed is 400 Mhz, resulting in a unidirectional bandwidth of 6.4 Gbps and per-port bidirectional bandwidth of 12.8 Gbps. Credit-based flow control is im-

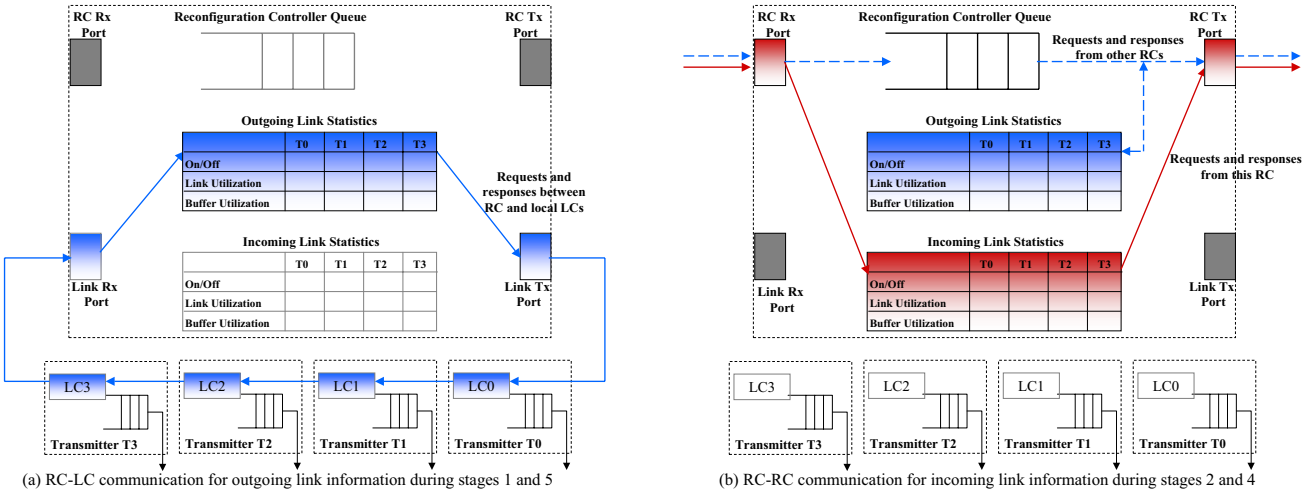


**Figure 4. Reconfiguration algorithm implementation.**

plemented for a single flit buffer with credits incurring a single cycle channel delay. For the optical network, we assume a channel speed of 5 Ghz, based on current optical technology. At 5 Gbps data rates, the transmission of an 8 byte flit takes around $12.8nsec$.

**Simulation Results:** The performance of E-RAPID was compared to other electrical networks for several communication patterns including uniform, bit-reversal, butterfly, complement, matrix transpose, neighbor and perfect shuffle for varying network size from 16 to 1024. Due to space constraints, only a few results are shown. Figures 5(a), 5(b), 5(c) and 5(d) show the throughput and latency plots for uniform and complement traffic for 64 nodes with 8 nodes per board. Under uniform traffic condition, the load is well balanced among all the system boards. In such a scenario, the non-reconfigured E-RAPID network shows identical performance the reconfigured E-RAPID. The plot for throughput shows that E-RAPID outperforms the closest electrical networks by 20% and is saturated beyond electrical networks. More significantly, with reconfiguration, there is no excess latency penalty. This implies that LS independently evaluates if reconfiguration is necessary. If it cannot reconfigure the network, it does not hinder the ongoing communication. The best case performance is observed for complement traffic. In complement traffic, nodes 0, 1, 2 ... 7 on board 0 communicates with node 63, 62, 61, ... 56 on board 7. Therefore, in the non-reconfigured E-RAPID, the network is saturated even for low load. LS algorithm re-allocates the entire board bandwidth for communicating between boards 0 and 7. This results in almost 400% improvement as compared to the non-reconfigured E-RAPID. This also results in 30% improved performance as compared to the hypercube network. For the remaining traffic patterns, the performance of reconfigured E-RAPID falls between uniform and complement traffic results.

Figure 5(e) shows the system sensitivity to time varying traffic for 16 node network. Each reconfiguration phase ($R_w$, $w$ = 1, 2, ... 10) is triggered after 2000 simulation cycles. Initially, complement traffic behavior is simulated from board 0 to board 3 with a single wavelength. The average latency is shown only for packets from board 0. As the buffer utilization increases, the reconfiguration algorithm re-allocates the entire incoming bandwidth of board 3 to board 0 at $R_2$. This causes the average packet latency to drop during the $R_2$ and $R_3$. After the $R_3$, another node 4 from board 1 begins communication to board 3. As the wavelength was previously re-allocated to board 0, it waits for the next reconfiguration phase to begin. At $R_4$, the statically allocated channel is re-acquired by board 1. This reduces the number of wavelength allocated to board 0 to 3. At the same time, this causes the average latency to increase as there are 4 nodes communicating using 3 wavelengths. After $R_5$, another node 8 from board 2 starts communication to board 3. This further reduces the number of wavelengths allocated to board 0 to 2 and results in a slight increase in latency. After this increase, the average latency stabilizes as shown. Therefore, LS re-allocates bandwidth from idle boards to highly utilized boards. At the same time, should any of these idle boards should become active, it releases the wavelength from the re-allocated board. Figure 5(f) shows the effect of the reconfigured traffic on non-reconfigured traffic. After reconfiguration, reconfigured traffic interacts with non-reconfigured traffic, with traffic from both nodes arriving at the same transmitter queue. While this interaction exists, its impact on the overall latency is minimal as seen.

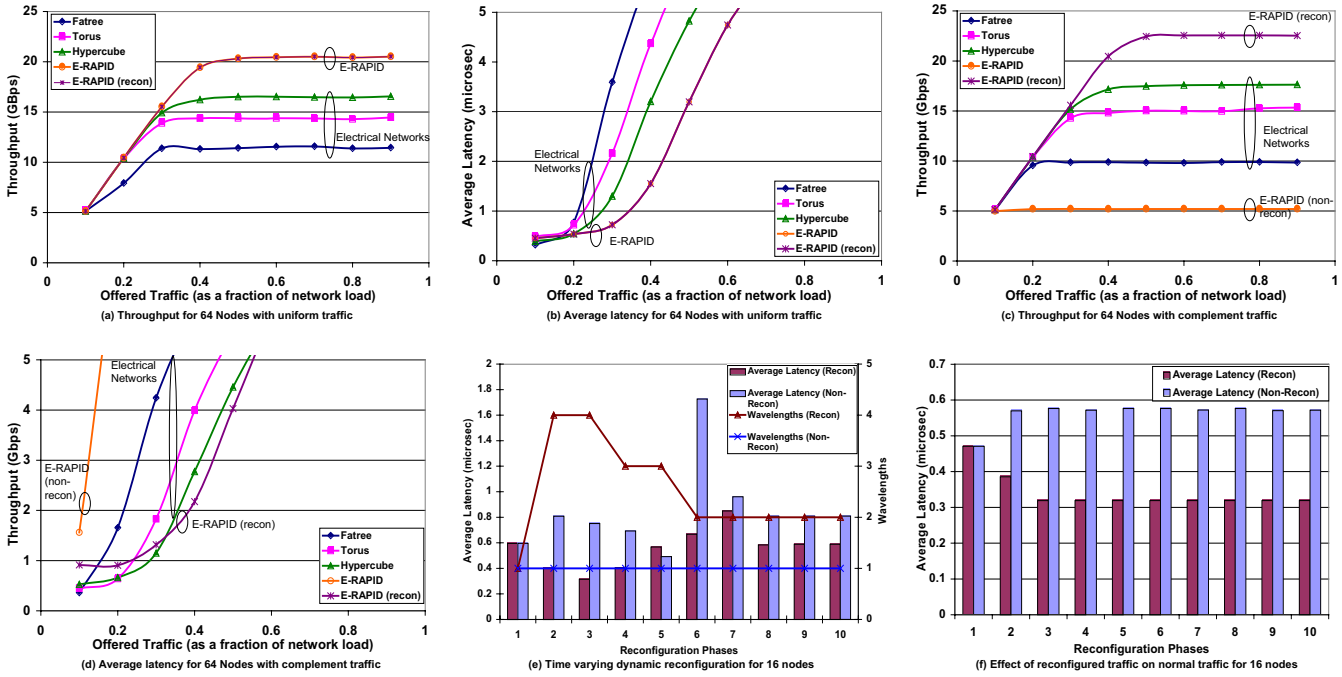


**Figure 5. Performance evaluation for 64 and 16 node E-RAPID network.**

## 4. Conclusion

In this paper, we proposed a dynamically reconfigurable optically interconnected architecture called E-RAPID that 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. The proposed dynamic bandwidth re-allocation (DBR) technique called Lock-Step (LS) further improves the performance of the proposed architecture by adapting to non-uniform communication patterns. In case of complement traffic saturating the network at low loads, the reconfigured E-RAPID can reduce network congestion by reallocating more bandwidth that result in increased throughput and reduced network latency.

**Acknowledgement** This research is supported by NSF grants CCR-0309537, CCF-0538945, Connection One and a grant from Intel Corporation.

## References

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

[2] Edris Mohammed and et.al., "Optical interconnect system integration for ultra-short-reach applications," *Intel Technology Journal*, vol. 8, pp. 114–127, 2004.

[3] Jeff Kash and et.al, "Bringing optics inside the box: Recent progress and future trends," in *16th Annual Meeting of the IEEE/LEOS*, October 2003, p. 23.

[4] Avinash Karanth Kodi and Ahmed Louri, "Rapid: Reconfigurable and scalable all-photonic interconnect for distributed shared memory multiprocessors," *Journal of Lightwave Technology*, vol. 22, pp. 2101–2110, September 2004.

[5] Patrick Dowd and et.al., "Lighnting network and systems architecture," *Journal of Lightwave Technology*, vol. 14, pp. 1371–1387, 1996.

[6] Chunming M. Qiao and et.al., "Dynamic reconfiguration of optically interconnected networks with time-division multiplexing," *Journal of Parallel and Distributed Computing*, vol. 22, no. 2, pp. 268–278, 1994.

[7] Praveen Krishnamurthy, Roger Chamberlain, and Mark Franklin, "Dynamic reconfiguration of an optical interconnect," in *36th Annual Simulation Symposium*, 2003.

[8] Principles and Practices of Interconnection Networks, *William James Dally and Brian Towles*, Morgan Kaufmann, San Fransisco, 2004.

[9] X. Chen, Li-Shiuan Peh, Gu-Yeon Wei, Yue-Kai Huang, and Paul Pruncal, "Exploring the design space of power-aware opto-electronic networked systems," in *11th International Symposium on High-Performance Computer Architecture (HPCA-11)*, February 2005, pp. 120–131.

[10] J. Robert Jump, "Yacsim reference manual," *Rice University Available at http://www-ece.rice.edu/ rppt.html*, March 1993.

[11] Fabrizio Petrini, Eitan Frachtenberg, and Adolfy Hoisie, "Performance evaluation of the quadrics interconnection network," *Cluster Computing*, vol. 6, pp. 125–142, 2003.