Performance Adaptive Power-Aware Reconfigurable Optical Interconnects for High-Performance Computing (HPC) Systems

Avinash Kodi[*]
Department of Electrical Engineering and
Computer Science
Ohio University
Athens, OH 45701
avinashk@ece.arizona.edu

Ahmed Louri
Department of Electrical and Computer
Engineering
University of Arizona
Tucson, Arizona 85721
louri@ece.arizona.edu

ABSTRACT

As communication distances and bit rates increase, opto-electronic interconnects are being deployed for designing high-bandwidth low-latency interconnection networks for high performance computing (HPC) systems. While bandwidth scaling with efficient multiplexing techniques (wavelengths, time and space) are available, static assignment of wavelengths can be detrimental to network performance for non-uniform (adversial) workloads. Dynamic bandwidth re-allocation based on actual traffic pattern can lead to improved network performance by utilizing idle resources. While dynamic bandwidth re-allocation (DBR) techniques can alleviate interconnection bottlenecks, power consumption also increases considerably. In this paper, we propose to improve the *performance* of optical interconnects using DBR techniques and simultaneously optimize the *power* consumption using Dynamic Power Management (DPM) techniques. DBR re-allocates idle channels to busy channels (wavelengths) for improving throughput and DPM regulates the bit rates and supply voltages for the individual channels. A reconfigurable opto-electronic architecture and a performance adaptive algorithm for implementing DBR and DPM are proposed in this paper. Our proposed reconfiguration algorithm achieves a significant reduction in power consumption and considerable improvement in throughput with a marginal increase in latency for various traffic patterns.

Categories and Subject Descriptors

B.4.3 [**Hardware**]: Input/Output and Data Communications—*Interconnections(Subsystems)*; C.0 [**Computer Systems Organization**]: General

[*]This work was performed at the High-Performance Computing Architecture and Technologies (HPCAT) Laboratory, University of Arizona, Tucson.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
SC07 November 10-16, 2007, Reno, Nevada, USA
Copyright 2007 ACM 978-1-59593-764-3/07/0011 ...$5.00.

General Terms

Design, Performance

Keywords

Performance Modeling, Power-Aware, Reconfigurable Optical Interconnects, High-Performance Computing (HPC)

1. INTRODUCTION

The increasing bandwidth demands at higher bit rates and longer communication distances in high-performance computing (HPC) systems are constraining the performance of electrical interconnects [1, 2, 3, 4, 5, 6]. This has given rise to opto-electronic networks that can support greater bandwidth through a combination of efficient multiplexing techniques (wavelength-division, time-division, and space-division) for board-to-board and rack-to-rack interconnects. Opto-electronic interconnects provide maximum flexibility for HPC systems by augmenting electronic processing functionalities with high bandwidth optical communication capabilities, thereby optimizing cost to performance ratio.

In an optically interconnected network, it is often that the wavelengths or the channels are statically allocated to nodes or boards using different wavelengths, fibers and time-slots [6, 7, 8, 9]. Static allocation of wavelengths in optical interconnects offers every node with equal opportunity for inter-processor communication. While static allocation improves performance for uniform or benign traffic patterns, the network congests for non-uniform or adversial traffic patterns due to uneven resource utilization. Based on the enormous bandwidth demands (in excess of Terabytes per second) of future HPC systems, optical interconnects will need to be much more flexible to adapt to various application communication patterns. Therefore, dynamic re-allocation of bandwidth based on actual traffic utilization can improve performance by utilizing idle resources in the network. Prior work on dynamic reconfiguration have used active electro-optic switching element [5], time-slots based bandwidth re-allocation [10] and both time and space based bandwidth switching [11].

While opto-electronic networks can improve performance with higher bit rates and dynamic re-allocation of bandwidth, power consumption is still a critical problem for HPC systems. As interconnection network consumes a sizeable fraction of the system power budget, researchers have proposed several power-aware techniques to optimize power consumption for HPC systems. Dynamic power reduction techniques such as DVFS (Dynamic Voltage and Frequency Scaling) [12, 13, 14] and DLS (Dynamic Link Shutdown) [15] have been suggested for electrical networks. In DVFS, voltage and frequency of the electrical link are dynamically adjusted to different power levels according to traffic intensities to minimize power consumption. DLS, on the other hand turns down the link if it is not used and turns up the link when needed. In [13], power-aware opto-electronic network design space is explored by regulating power consumption in response to actual network traffic. However, they have designed efficient power regulation control policies without bandwidth re-allocation.

The motivation for designing dynamically reconfigurable, power-aware opto-electronic network for HPC systems is two fold. First, as bandwidth demands increase, networks that can dynamically re-allocate bandwidth by adapting to shifts in network traffic can gain significant improvement in performance. Second, as spatial and temporal locality exists due to inter-process communication patterns, opto-electronic power-aware networks can optimize their power consumption and thereby improve performance by scaling bit rates and supply voltage. While scaling the bit rates allows opto-electronic networks to reduce their power consumption, this can adversely affect performance by increasing latency. Similarly, dynamically re-allocating bandwidth can improve the network performance, but at the same time consume more power. Taken together, this work evaluates the power-performance trade-off by balancing power consumption with improving network performance.

In this paper we propose a dynamically reconfigurable optical interconnect called E-RAPID that not only dynamically re-allocates bandwidth, but also reduces the power consumption while delivering high-bandwidth, and high connectivity. Dynamic Power Management (DPM) technique such as DVFS is applied in conjunction with Dynamic Bandwidth Re-allocation (DBR) technique based on prior network utilization for various communication patterns. We propose a dynamic reconfiguration algorithm called Lock-Step (LS) technique that adapts to changes in communication patterns. LS is a history-based distributed reconfiguration algorithm that triggers reconfiguration phases, disseminates state information, re-allocates system bandwidth, regulates power consumption and re-synchronizes the system periodically with minimal control overhead. LS has several advantages including: (1) Decentralized power scaling such that every board independently makes power control decisions. (2) Re-allocation of bandwidth happens between any system boards without affecting the on-going communication in the overall system, and (3) Maximum bandwidth can be provided for system boards for hot-spot/bursty traffic pattern, where extremely high load is placed for a short duration of time.

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.

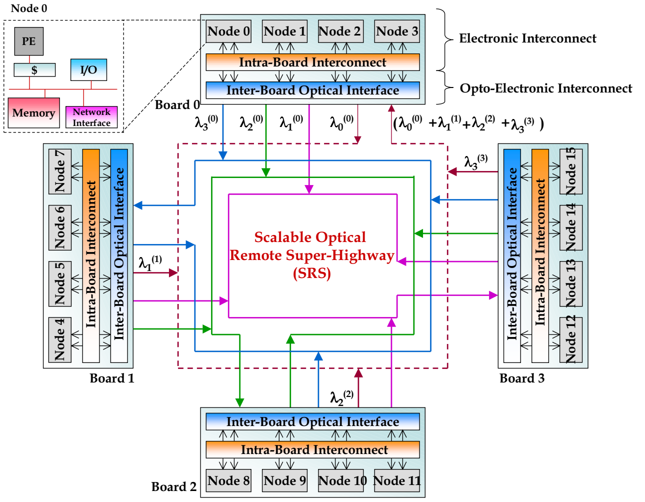


Figure 1: Routing and wavelength assignment in E-RAPID for inter-board communication.

2.1 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 $\lambda_{(s-d)}^{(s)}$ if $s > d$, where B is the total number of boards in the system. 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(a) shows the intra-board interconnections for board 0. 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

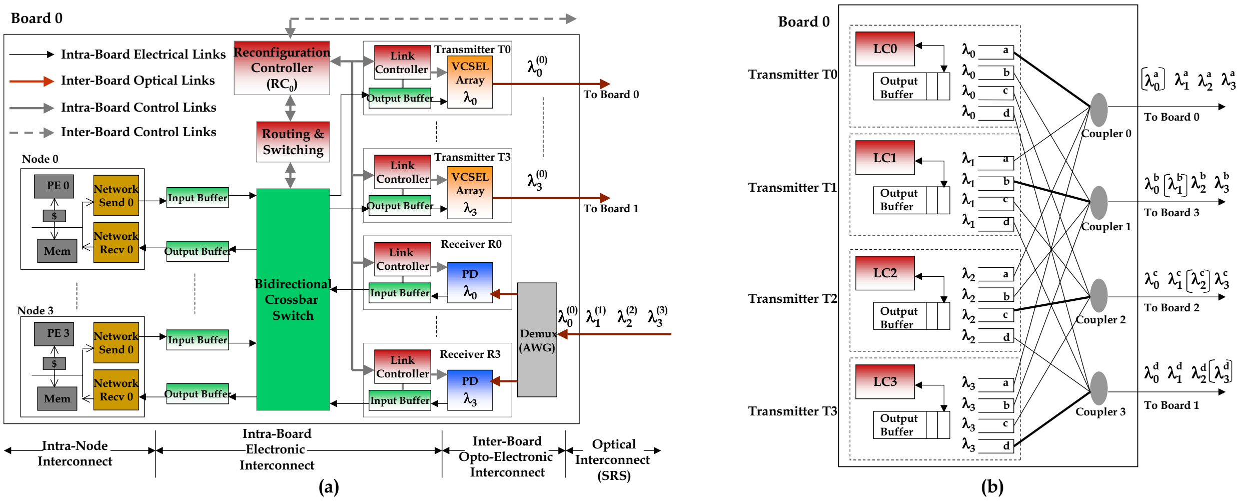


Figure 2: (a) The proposed on-board interconnect for the E-RAPID architecture with reconfiguration controller (RC) and link controllers (LC). (b) The proposed technology for reconfiguration using passive couplers and array of lasers per transmitter port.

computation (RC), virtual-channel allocation (VA) and per-flit steps include switch allocation (SA) and switch traversal (ST)[16]. 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.

2.2 Technology for Dynamic Bandwidth Re-Allocation (DBR)

From Figure 2(a), each optical transmitter is composed of an array of similar wavelength lasers. The enabling technology for reconfigurability in E-RAPID is shown in Figure 2(b). 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 wavelength as per the communication requirements from section 2.1 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 communication in 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 2(b), 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 λ_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.

2.3 Dynamic Power Management (DPM) of Optical Interconnects

An optical link in E-RAPID architecture consists of the transmitter, the receiver and the channel. Considering a passive channel, the total power consumption of an optical link depends on the transmitter and the receiver power. Transmitter power is consumed at the laser and laser driver, where as the receiver power is consumed at the photodetector, transimpedance amplifier (TIA) and clock and data recovery (CDR) circuitry [12, 17]. While both Multiple-Quantum Wells (MQW) [17] with external modulators and VCSELs (vertical-cavity surface emitting lasers)[17, 18] can be considered as light sources, we assume a VCSEL (vertical-cavity surface emitting laser) as the laser source, which eliminates the need for the external modulator. Moreover, there are commercial vendors who provide one-dimensional multiple-wavelength VCSEL arrays which can be used for reconfiguration in E-RAPID [19]. In the next subsection, we evaluate the power dissipated in an opto-electronic link and device parameters that can be controlled to regulate the power consumption.

2.3.1 Power Calculations

The total power consumed by an entire opto-electronic link is given by:

$$P_T = (P_{Driver}+P_{VCSEL})_{TX}+(P_{Photodiode}+P_{TIA}+P_{CDR})_{RX} \tag{1}$$

The superbuffer in the VCSEL driver is a set of cascaded inverters, and the size of each inverter is larger than the previous one by a constant factor δ. The total power dissipated in the driver stages is calculated as

$$P_{Driver} = \gamma C_L V_{dd}^2 B_R \tag{2}$$

where γ is the switching factor, C_L is the total load capacitance of the superbuffers (of n inverters), V_{dd} is the supply voltage and B_R is the bit rate. The total capacitance is the sum of input and output capacitance of all the inverters, and is given as [17]

$$C_L = C_{Load} - C_{in} + \Sigma_{k=0}^{n-1}(C_{in} + C_{out})\delta^k \tag{3}$$

where C_{Load} is the load capacitance of the inverter chain, C_{in} and C_{out} are the input and output capacitances of the minimum sized inverters. We adopt the VCSEL with a CMOS driver from [17], where the driver circuitry consists of two NMOS transistors providing the threshold and modulation currents and a superbuffer driving the gate that delivers the modulation current. The VCSEL power consumed is given as

$$P_{VCSEL} = I_{Total}.V_{source} = (I_{th}+I_m\gamma)(V_{th}+I_mR_s+V_{dd}-V_{tn}) \tag{4}$$

The total current is the sum of threshold (I_{th}) and modulation currents times the switching factor. The total voltage is the sum of the VCSEL threshold voltage (V_{th}), the voltage drop across the series resistance (R_s) and the minimum source-drain voltage (V_{dd} - V_{tn}) to ensure the gate that delivers the modulation current is in saturation.

At the receiver, we determine the power consumed by the photodetector and the TIA. This is modeled similar to [20], which consists of the photodetector as a current source ($I_d + \alpha\beta I_m$) and a common source amplifier connected by a feedback resistance, R_f. I_d is the dark current, α is the VCSEL efficiency in A/W and β is the detector efficiency in W/A. The input capacitance to the amplifier $C_{in} = C_D+C_g$, where C_D is the diode capacitance and $C_g = C_{ox}WL$ is the gate capacitance. The VCSEL needs to generate enough light which depends on I_m such that the receiver will produce an output signal of amplitude $\triangle V_0$, which can then be amplified by further receiver stages. This can be approximated as [20]

$$\triangle V_0 = \frac{\gamma I_m}{\beta\alpha R_f} \tag{5}$$

Therefore, the power consumption of the VCSEL is defined by the needs of the receiver for a given B_R and V_{dd}. The total power dissipated in the TIA based receiver circuit is then given as

$$P_{TIA} = I_b V_{dd} + I_d^2 V_{dd} + \gamma(\alpha\beta I_m)^2 R_f \tag{6}$$

where I_b is the bias current of the internal amplifier and is given by $I_b = \omega_{3dbint} V_e C_0$ where ω_{3dbint} is the 3db bandwidth of the internal amplifier, V_e is the early voltage, and C_0 is the output capacitance. The gain-bandwidth product of the internal amplifier is $GBW = A(\omega)\omega_{3dbint} = g_m/C_0$, where $w = 2\pi B_R$ and g_m is the transconductance. The relationship between the internal amplifier bandwidth and the maximum bit rate is given as $\omega = 0.35\omega_{3dbint}$. The bandwidth of TIA is assumed to be half the bandwidth of the internal amplifier, therefore, the 3dB bandwidth of TIA is approximated as

$$\omega_{3dbtia} = \frac{A(\omega)}{R_f C_i n} = \frac{w}{0.7} \tag{7}$$

Then the total power dissipated at the receiver can be obtained as

$$P_{TIA} = \frac{0.7A(\omega)I_d^2}{2\pi C_{in} B_R} + (\frac{2\pi V_e C_0 V_{dd}}{0.35} + \frac{2\pi\gamma\triangle V_0^2 C_{in}}{0.7A(\omega)})B_R \tag{8}$$

Then the desired I_m at the transmitter can be obtained by solving (5), (7) and (8). The power dissipated at the clock and data recovery is given as [12]

$$P_{CDR} = \gamma C_{CDR} V_{dd}^2 B_R \tag{9}$$

where C_{CDR} is the capacitance of the clock and data recovery unit.

2.3.2 Dynamic Power

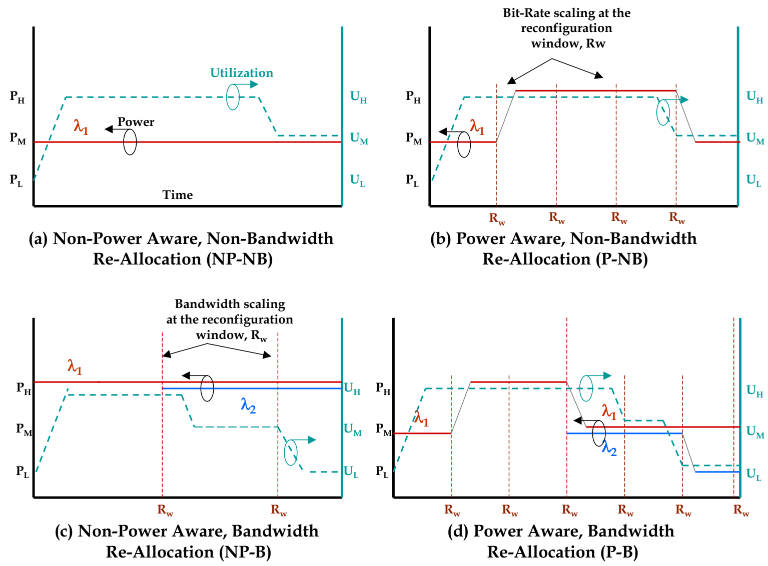
At the transmitter, VCSEL is generally biased at threshold current I_{th}, and the dynamic power consumed by VCSEL grows with the modulation current I_m. I_m is controlled by the receiver's minimum voltage swing required as given by equation (5). For the VCSEL driver, the dynamic power is consumed by charging/discharging the capacitor chain and scales with B_R and V_{dd}^2. At the receiver, TIA consumes maximum power and it depends on V_{dd} and B_R as given by equation (8). The CDR can be frequency and voltage scaled as bit rate varies as (V_{dd}^2 and B_R) from equation (9).

When the bit rate scales down, the supply voltage is also reduced of all the above components, resulting in power savings. Scaling the power level focuses on reducing the delay incurred during the slow voltage transitions as compared to frequency transitions [12, 13]. As the link can be operational during the slow voltage transitions, increasing the link speed involves increasing the voltage before scaling the frequency. Similarly, the frequency is decreased before scaling the voltage. The delay penalty is limited to frequency transitions as this requires the CDR (implemented as phase-locked loop) to relock the bit-rate and re-synchronize the clock with the incoming data.

3. DYNAMIC RECONFIGURATION

3.1 Power-Performance Trade-Offs

To provide more insight into power and performance tradeoffs, consider Figure 3 which shows various combination of power regulation and bandwidth re-allocation techniques. These techniques include four cases namely, Non-Power Aware Non-Bandwidth Re-allocation (NP-NB), Power-Aware Non-Bandwidth Re-allocation (P-NB), Non-Power Aware, Bandwidth Re-allocation (NP-B) and Power Aware Bandwidth Re-allocation (P-B).



(a) Non-Power Aware, Non-Bandwidth Re-Allocation (NP-NB)

(b) Power Aware, Non-Bandwidth Re-Allocation (P-NB)

(c) Non-Power Aware, Bandwidth Re-Allocation (NP-B)

(d) Power Aware, Bandwidth Re-Allocation (P-B)

Figure 3: Design space of power-aware, and bandwidth reconfigurability.

Let us consider three power levels, namely, low-power P_L, mid-power P_M and high-power P_H as shown on the left y-axis of Figure 3, and three link utilization levels, low-utilization U_L, mid-utilization U_M, and high-utilization U_H as shown on the right y-axis of Figure 3. Link utilization measures the amount of time the link is in use. Moreover, assume that these link utilization levels are measured at every reconfiguration window, R_w. Reconfiguration statistics will be gathered from the past reconfiguration window to predict the future link utilization, and the corresponding power and bandwidth levels.

Figure 3(a) shows the NP-NB technique. In this case, irrespective of the link utilization, the power consumption remains constant and the network cannot react to fluctuations in traffic patterns. Figure 3(b) shows P-NB technique, where the link utilization is measured at every R_w. P-NB technique allows link power to scale with link utilization. This technique is shown to reduce power consumption, but is not able to respond to increases in bandwidth demands. Figure 3(c) shows the NP-B, where the link utilization is measured at every R_w. NP-B technique allows bandwidth re-allocation to adapt to link utilization. This technique is also shown to improve performance, but is unable to regulate power. Figure 3(d) shows P-B where both power is regulated and bandwidth is re-allocated upon changes in link utilization. This technique is shown to not only reduce power, but also improve performance.

Power consumption and latency $(1/B_R)$ are inversely related, i.e. there is a minimum power at which the latency increases asymptotically to infinite and a minimum latency at which the power consumption increases asymptotically to infinite. However, between these extremes exist several design points at which either power or latency can be optimized. Dynamic power management (DPM) allows power scaling by controlling the bit rates and supply voltages. Dynamic bandwidth re-allocation (DBR) technique allows multiple links to be operational for a given communication. Taken together, this provides a two-dimensional design space optimization problem. In this work, we re-allocate channels for overloaded links by DBR and regulate power consumption by DPM for all the links in the network.

3.2 Dynamic Reconfiguration Technique

LS technique re-allocates link bandwidth, scales the bit rates and supply voltages 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 reconfiguration controller (RC) and link controller (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. LS protocol works in the background and does not affect the on-going communication, thereby minimizing the impact of reconfiguration latency on the overall network latency.

Reconfiguration Statistics: Historical statistics are collected with the hardware counters located at each LC. Each LC is associated with an optical transmitter to measure link statistics, and to turn on/off the laser. The link utilization Link$_{util}$ tracks the percentage of router clock cycles when a packet is being transmitted in the optical domain from the transmitter queue. The buffer utilization Buffer$_{util}$ determines the percentage of buffers being utilized before the packet is transmitted. At low-medium network loads, link$_{util}$ provides accurate information regarding whether a link is being used at all, where as Buffer$_{util}$ provides accurate information regarding network congestion at medium-high network load. All these statistics are measured over a sampling time window called *Reconfiguration window* or phase, R_w. This sampling window impacts performance, as reconfiguring finely incurs latency penalty and reconfiguring coarsely may not adapt in time for traffic fluctuations. We utilize network simulations to determine the optimum R_w.

Each RC_i, $i = 0, 1, \ldots\ B-1$ is connected to all the LC_j, $j = 0, 1, \ldots\ D-1$ on the board. In addition, each RC_i is also connected to $(RC_{i+1}) modulo B$ 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. Each LC associated with a transmitter has a link utilization counter, a buffer utilization counter and a on/off binary value for every wavelength λ_0, λ_1, λ_2 ... on a given system board.

The symmetry of E-RAPID with respect to the number of wavelengths provides the insight into reconfiguration algorithm. For example, if $\Lambda = \lambda_0$, λ_1, λ_2 ... λ_{B-1} is the total number of wavelengths associated with the system, we can see that this is exactly the same number of wavelengths transmitted/received from every system board. In other words, the number of *outgoing* or *incoming* wavelengths per system board is the same. Therefore, in order to balance the load and re-allocate wavelengths on a 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 LS algorithm.

LS Algorithm: In order to implement LS, RCs evaluate the state information and re-allocate the bandwidth for the current R_w based on previous R_w. After RCs have decided which links to re-allocate, this information is disseminated back to the RCs on other boards. RCs then determine the power level for each link and convey re-allocation and power level information to the LCs. The pseudo code of the LS algorithm is shown in Table 1. After R_w, in Step 2, RC_i sends out $Link_{Request}$ packets to LC_i as shown in Figure 4(a). When this packet is received by RC_i, it updates all the *outgoing* link statistics. In Step 3, each RC_i sends $Board_{Request}$ packet to obtain all the link statistics for its *incoming* links as shown in Figure 4(b). As it sends out, due to the symmetry of the ring architecture, it receives Board$_{Request}$ from other RC_i. For example, when board 1 receives BR_0 from say board 0, it will update the field for wavelength with which board 1 communicates with board 0, i.e. λ_1 using the data stored in its *outgoing* link statistic. When the board RC_i receives its own Board$_{Request}$ packet, it updates all the incoming link statistics.

In step 4, DBR is implemented. Now, each RC_i computes if reconfiguration is necessary based on buffer congestion, B_{con} and minimum link utilization L_{min}. While profiling of traffic traces can provide more accurate information regarding when the network is actually congested, setting the B_{con} 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 L_{min} to 0.0 which indicates no packets are being transmitted on the link. Each incoming link statistic is classified into three categories as under-utilized if Link$_{util}$ is less than L_{min} (implying that this wavelength can be re-allocated), normal utilized if Buffer$_{util}$ less than B_{con} and Link$_{util}$ is greater than L_{min} (implying the wavelength is well utilized) and over-utilized if Buffer$_{util}$ is greater than B_{con} (implying that additional wavelengths are needed). RC would allocate the under-utilized links to the over-utilized links.

In Step 5 and 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.

In Step 6 DPM is implemented. The power level for the next R_w is computed based on two buffer thresholds, B_{min} and B_{max}. While other researchers have used link utilizations to regulate power levels, link utilization does not allow for aggressive power regulation. In our power regulation technique, we aggressively push the link to be fully utilized and then evaluate based on buffer thresholds. If the Buffer$_{util}$ falls below B_{min}, the power level of the link is scaled down to the next lower power level, P_{n-1}. If the Buffer$_{util}$ exceeds B_{max}, the link power is scaled up to the next power level, P_{n+1}. If the Buffer$_{util}$ falls between B_{min} and B_{max}, the link retains the same power level, P_n. While multiple bit rates can conserve more power by finely tuning the bit rates to the buffer utilization, it increases the delay penalty by re-clocking the CDR circuitry every time the bit rate is scaled. Similarly, if R_w is too small, the bit rates will be tuned too often, again incurring excess delay penalty. If R_w is too large, the bit rates cannot scale to accommodate large fluctuations. We use network simulation to determine an optimum value of R_w. By using 6 power levels in our system architecture, we avoid multiple bit rate transitions.

In Step 7 and 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 and re-clocking to the new power level. As there is one-to-one mapping between the transmitter and the receiver, the transmitter LC_i injects a bit rate control packet on the link and stops transmission for the duration while the frequency and voltage transitions occur. When this bit rate control packet is received, the optical receiver then re-clocks to the new bit rate.

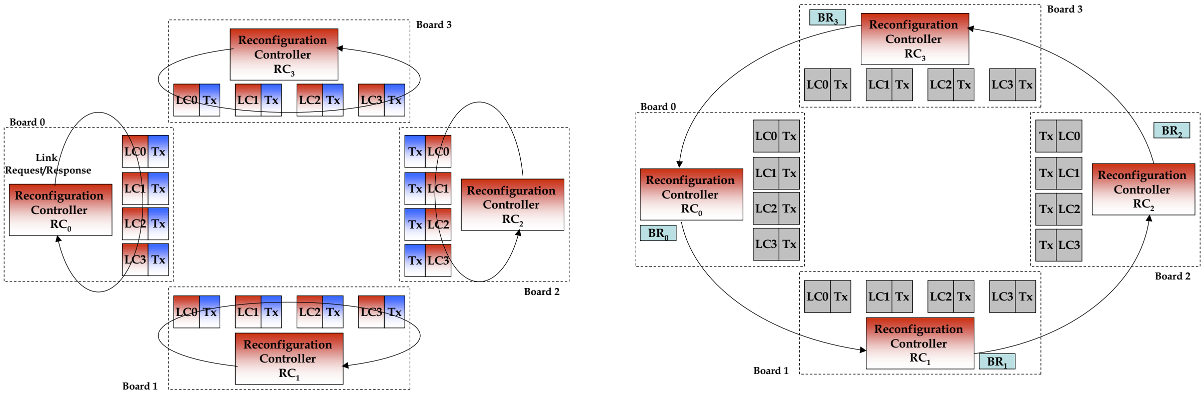
4. PERFORMANCE EVALUATION

4.1 Simulation Network Parameters

The performance of E-RAPID is evaluated using YACSIM and NETSIM discrete-event simulator and is compared to various non-power/power aware, non-bandwidth/bandwidth reconfigured network configurations. We use cycle accurate simulations to evaluate the performance of E-RAPID. 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[16]. The simulator was warmed up under load without taking measurements until steady state was reached. Then a sample of injected packets were labelled during a measurement interval. The simulation was allowed to run until all the labelled packets reached their destinations.

For the on-board router model designed for E-RAPID architecture, we considered the channel width to be 32 bits and the router speed to be 400 Mhz, resulting in a unidirectional bandwidth of 12.8 Gbps and per-port bidirectional bandwidth of 25.6 Gbps. It takes a single router cycle for routing, virtual channel allocation and switch allocation. For most of the runs, we maintained a constant packet size of 128 Bytes, resulting in a 8 flit packet size.

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 HPC sys-


Board 3
Reconfiguration Controller RC$_3$
LC0 Tx
LC1 Tx
LC2 Tx
LC3 Tx
Board 0
Link Request/Response
Reconfiguration Controller RC$_0$
Board 2
Reconfiguration Controller RC$_2$
Board 1
Reconfiguration Controller RC$_1$
BR$_3$
BR$_0$
BR$_2$
BR$_1$

(a) RC-LC communication

(b) RC-RC communication

Figure 4: Reconfiguration algorithm implementation.

tems. The power-performance of E-RAPID utilizing various techniques such as NP-NB, NP-B, P-NB and P-B were evaluated for several communication patterns including uniform, butterfly (a_{n-1},a_{n-2}, ...,a_1, a_0 communicates with a_0,a_{n-2},...,a_1,a_{n-1}), complement (a_{n-1},a_{n-2}, ...,a_1,a_0 communicates with node $\overline{a_{n-1}, a_{n-2}, ..., a_1, a_0}$), and perfect shuffle ($a_{n-1}$,$a_{n-2}$, ...,$a_1$,$a_0$ communicates with with node a_{n-2}, a_{n-3},...,a_0,a_{n-1}) for network size of 64 nodes. While networks of varying sizes were modelled, due to space constraints, we describe the performance (throughput, latency and power) for a 64 node network.

Optical Network Modelling: In the calculations for a oxide based VCSEL [17], we considered a 50% duty cycle, γ = 0.5, threshold current $I_{th} = 0.1\ mA$, series resistance R_s = 250 ohm, threshold voltage, V_{th} = 2V, efficiency $\beta = 0.3$, and V_{tn} = 0.38 V. For the driver we considered, C_{load} = 50pF, input and output capacitance of minimum sized inverters, $C_{in} = C_{out}$ = 2 pF. For the receiver [20], we considered a minimum voltage swing $\triangle V_0$ = 100mV, detector efficiency $\alpha = 0.4$ A/W, amplifier gain A = 10, L = 0.25 μ, μ_n = 1300 $cm^2/V - sec$, V_e = 20 V, C_D = 0.05 pF and C_0 = 0.05 pF, I_d = 100 nA, and CDR capacitance C_{CDR} = 9.26 pF.

From the above parameters and solving equations from section 2.3.1, we estimated the various power dissipated in the link at the transmitter and the receiver. The link power is dominated by the receiver power consisting of the TIA and CDR where as the VCSEL and driver dissipate minimal power. The receiver power can be further reduced by considering other low impedance resistive circuits instead of the TIA [20]. The total power dissipated at 10 Gbps is approximately 535 mW. With the bit rate scaling from 10 Gbps to 5 Gbps and the supply voltage scaling from 1.8 V to 0.9 V, the power dissipation for a 5 Gbps link reduces to almost 108 mW, an 80% reduction in power savings. For the optical network, we considered 6 bit rates corresponding of 5, 6, 7, 8, 9 and 10 Gbps and V_{dd} scaling from 0.9 to 1.8 V giving us 6 different power levels {108.8mW, 163.7mW, 232.5mW, 316.0mW, 417.0mW, 535.0mW}. The CDR delay was estimated from [13], which was normalized to our network clock cycle. In [13], the link was disabled for 12 network clock cycles (for frequency scaling) after the bit rate transitions to give CDR to re-lock to the input data. In our network simulation, after the control bit rate packet is transmitted, the transmitter conservatively disables the link for 65 cycles.

4.2 Results and Discussion

Reconfiguration Window R_w: In order to determine the optimum reconfiguration window size R_w, we performed simulation by varying the window size from 500 simulation cycles to 4000 cycles. We evaluated the latency and normalized power dissipation for complement traffic pattern. Normalized power dissipation is calculated by averaging the various links operating at different bit rates and normalizing it to the maximum bit rate. The latency and power dissipation are evaluated for low (0.2) and medium (0.9) network loads in Figures 5(a) and 5(b) respectively. At low load of 0.2, the latency increases marginally with R_w, where as at high load of 0.5, the latency increases almost 8x for 4000 cycles as compared to 500 cycles. At high loads, most packets that need reconfiguring are already saturating the links, therefore increasing R_w worsens the situation. However at low loads, the number of packets saturating the network is less and therefore the impact on latency is much lesser. For low load of 0.2, the power dissipation increases with increasing R_w. This is because as the network warms up, the demand for reconfiguring at low loads may not exist, however at higher R_w, the need to reconfigure grows faster. At high loads, the network is already experiencing saturating, therefore for all values of R_w, reconfiguration is necessary indicating an almost equal power dissipation. Therefore, to balance the two constraints, we choose R_w of 1000 simulation cycles.

Buffer Thresholds, B_{min}, B_{max}: In order to determine the optimum values of buffer thresholds, we consider two conditions as shown in Figure 6(a) and Figure 6(b). In Figure 6(a), we set $B_{min} = 0.1$ and $B_{max} = 0.3$. As the network warms up, the bit rate is reduced and power savings is obtained. However, by $R_w = 4$, the bit rate starts increasing as the buffer utilization is greater than B_{max}. When the link operates at the peak rate of 10 Gbps, the buffer utilization starts falling, as maximum bandwidth is provided to transmit packets. Once it falls below B_{min}, the bit rates again scales down. This continues until the buffer utilization falls

Table 1: Lock-Step Algorithm for DBR and DPM Implementation

```
Step 1: Wait for Reconfiguration window, $R_w$

Step 2: Each $RC_i$ sends the $Link_{Request}$ control packet to all its outgoing $LC_i$
Step 2a: Each $LC_i$ computes the Link$_{util}$ and Buffer$_{util}$ for the previous $R_w$ and updates the field in the
$Link_{Request}$ packet and forwards to the next $LC_{i+1}$ and finally to $RC_i$

Step 3: Each $RC_i$ sends the $Board_{Request}$ control packet to all $RC_j$, $i \neq j$
Step 3a: $RC_i$ updates the Link$_{util}$ and Buffer$_{util}$ for the link (wavelength) with which it communicates with
$RC_j$ when it receives the $Board_{Request}$ packet from $RC_j$

Step 4: $RC_i$ receives its $Board_{Request}$ packet containing utilization information for all its incoming links
Step 4a: $RC_i$ classifies every $B - 1$ incoming links for DBR as
        If Link$_{util} \leq L_{min}$ => Under-Utilized
        If Link$_{util} \geq L_{min}$ & Buffer$_{util} < B_{con}$ => Normal-Utilized
        If Buffer$_{util} > B_{con}$ => Over-Utilized
        Re-allocates Under-Utilized links to Over-Utilized links

Step 5: Each $RC_i$ sends the $Board_{Response}$ control packet with updated link information to $RC_j$, $i \neq j$
Step 5a: $RC_i$ updates the wavelength re-allocation for the link with which it communicates with
$RC_j$ when it receives the $Board_{Response}$ packet from $RC_j$

Step 6: Each $RC_i$ performs DPM and classifies each link as
        If $B_{min} \geq$ Buffer$_{util}$ => Decrease Power Level ($P_{n-1}$)
        If $B_{min} \leq$ Buffer$_{util} \leq B_{max}$ => Maintain Power Level ($P_n$)
        If Buffer$_{util} > B_{max}$ => Increase Power Level ($P_{n+1}$)

Step 7: Each $RC_i$ sends the $Link_{Response}$ control packet to all its outgoing $LC_i$ with updates link
re-allocation information and new power level information
Step 7a: In response to DBR, each $LC_i$, turns off/on the lasers for wavelength re-allocation
Step 7b: In response to DPM, each $LC_i$, sends new $Power_{Level}$ packets if the new power level is different from
previous power level

Step 8: Go to step 1
```

between B_{min} and B_{max}. If the difference between B_{min} = 0.1 and B_{max} = 0.2, is lower as shown in Figure 7(b), the bit rate will fluctuate more often and at many instances operate at peak bit rates. Therefore to have a larger range of stable operating points, we choose B_{min} = 0.1 and B_{max} = 0.3. Increasing the difference between B_{min} and B_{max}, will prevent fluctuations, but the latency penalty will also increase.

Throughput, Latency, Power: Figures 7, 8 and 9 show the throughput, latency and overall power consumption for 64 nodes for uniform, complement, perfect shuffle and butterfly traffic patterns. All traffic patterns selected are adversial traffic patterns except uniform. From Figure 7, for uniform traffic, NP-NB (non-power aware non-bandwidth reconfigured) and NP-B (non-power aware bandwidth reconfigured) shows identical performance. Both P-NB (power aware non-bandwidth reconfigured) and P-B (power aware bandwidth reconfigured) show a 4% decrease in throughput. This is mainly due to the power awareness algorithm that attempts to regulate the power consumption which affects the throughput. For uniform traffic pattern, all nodes are equally probable to communicate with every other node. This balances the load on all links, thereby having no underutilized links to reconfigure. The worst case traffic pattern for E-RAPID is complement traffic, where all nodes on a given source board communicate with a destination board. For a 64 node network, nodes 0, 1, 2 ... 7 on board 0 communicates with node 63, 62, 61, ... 56 on board 7. Therefore, the network is saturated even for low load for E-RAPID architecture. As seen, NP-NB and P-NB, the network is saturated at very low loads. The throughput remains the same for both NP-NB and P-NB. With reconfiguration, all the remaining links can be provided to the system board, i.e. NP-B and P-B provide improved performance in terms of throughput. We achieve almost 400% improvement in throughput by completely reconfiguring the network. For perfect shuffle and butterfly patterns, the improvement in throughput is 37% and 33%. In these communication patterns, all nodes do not communicate with other boards. As these patterns have some component of local communication within the board, the percentage of improvement is reduced.

From Figure 8, for uniform traffic, the network saturates at 0.4 for NP-NB and NP-B where as the saturation point is slightly shifted for P-NB and P-B to 0.37 due to power awareness being implemented. There is no excess reconfiguration penalty for P-B and NP-B. This implies that LS independently evaluates if reconfiguration is necessary. If it cannot reconfigure the network, it does not hinder the on-going communication. For complement traffic, P-B and NP-B techniques show superior saturation values of 0.5 as opposed to P-NB and NP-NB, where the network is saturated for extremely low load of 0.1. The same is true for

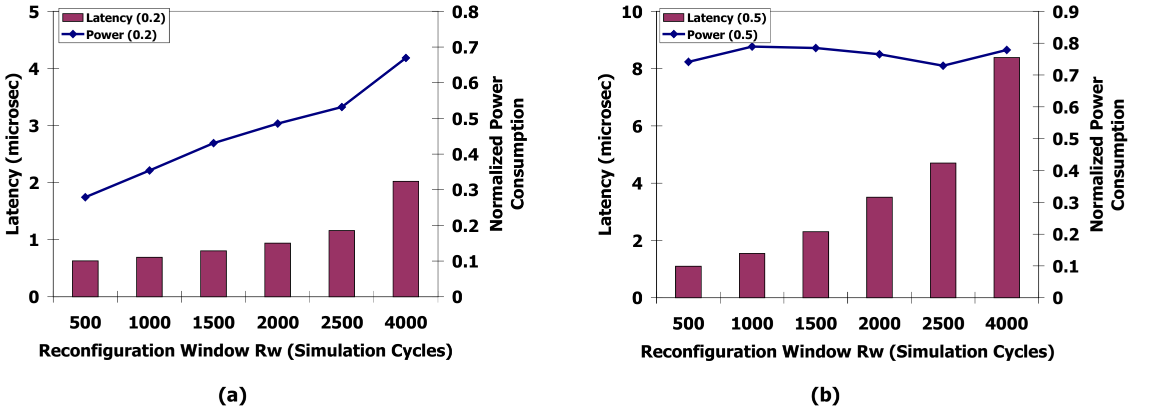


Figure 5: Reconfiguration window sizing for a network load of (a) 0.2 and (b) 0.5.

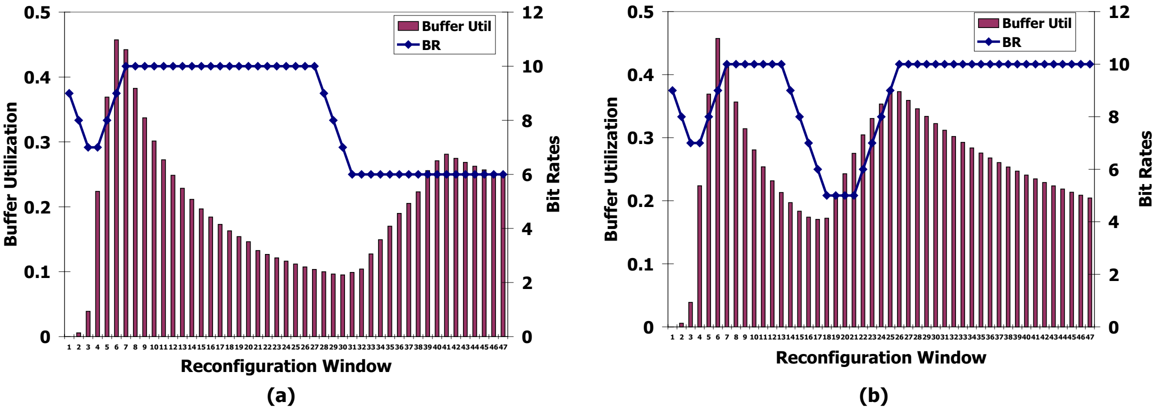


Figure 6: (a) Buffer utilization and bit rate comparisons for Complement traffic pattern with (a) B_{min} = 0.1 and B_{max} = 0.3 and (b) B_{min} = 0.1 and B_{max} = 0.2.

butterfly which saturates at 0.5 and perfect shuffle which saturates at 0.2. The latency is marginally more for P-B technique because of power regulation being implemented.

From Figure 9, the normalized power dissipation for all the traffic patterns are shown. For NP-NB and NP-B techniques, the power dissipated is at the maximum as all links are operational at the peak data transmission rate of 10 Gbps. For uniform traffic, by applying power awareness, we can reduce power consumption by almost 40% using either P-NB or P-B techniques. For complement traffic pattern, the power dissipated improves from 50% for low network loads to 20% at high loads. At high loads, bandwidth re-allocation causes more links to be active, thereby consumes more power. The sam is true for both perfect shuffle and butterfly patterns where the power dissipated in P-B technique is more than P-NB technique due to bandwidth re-allocation. In E-RAPID architecture, power regulation and bandwidth re-allocation allows the network, not only to improve performance by re-allocating idle links, but also to save power by bit rate and voltage scaling. NP-B allows only the bandwidth to be re-allocated, and P-NB allows only power to be scaled. This new P-B allows both, power as well as bandwidth to be reconfigured leading to improved network performance.

Degree of Reconfiguration: Figure 10 shows the degree of reconfiguration for complement and butterfly traffic in terms of throughput for varying network loads of 0.1, 0.5 and 0.9. Degree of reconfiguration indicates the number of links provided for re-allocating. From Figure 10(a), at low loads of 0.1, the throughput is insensitive to the amount of available bandwidth. At medium (0.5) and high loads (0.9), the network is sensitive to the amount of bandwidth availability. For a network load of 0.9, at N = 4, the improvement in throughput as compared to N = 2, is 27% where as at N = 8 as compared to N = 4, the improvement in throughput is almost 47%. Therefore, allocating more links is advantageous for complement traffic as all nodes use the same link for communication. From Figure 10(b), for butterfly traffic, the improvement in throughput at high loads (0.9) is lower. At N = 4, the improvement over N = 2, is 5% and the improvement at N = 8 as compared to N = 4, is 16%. For butterfly traffic pattern, allocating the entire bandwidth does not improve throughput significantly. These results show that based on the traffic patterns, link re-allocation can be optimized such that the performance is improved at much lower power.

5. CONCLUSION

In this paper, we combined dynamic bandwidth re-allocation (DBR) techniques with dynamic power management (DPM) techniques and proposed a combined technique called Lock-Step (LS) for improving the performance of the opto-electronic interconnect, while consuming substantial less power. We implemented LS on our proposed opto-electronic E-RAPID architecture and compared the performance of non-power/power aware and non-bandwidth/bandwidth reconfigured networks. Our proposed LS technique implemented the power-bandwidth (P-B) reconfiguration technique and achieved similar throughput and latency performance as a fully bandwidth reconfigured network while consuming almost 50% to 25% lesser power. More power levels and corresponding bit rates can further improve the performance as power scaling can follow the traffic pattern more accurately. The dynamic bandwidth re-allocation techniques proposed in this paper provides complete flexibility to re-allocate all system bandwidth for a given board. Cost-effective design alternatives that provide limited flexibility for reconfigurability may reduce performance, but lower the cost of the network. In the future, we will evaluate multiple power scaling techniques along with limited bandwidth reconfigurability for improving the system performance, reducing the power consumption and reducing the overall cost of the architecture. In addition, we will also evaluate the performance of DBR and DPM on HPC benchmarks.

6. ACKNOWLEDGMENTS

This research is supported by NSF grants CCR-0309537 and CCF-0538945.

7. REFERENCES

[1] 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.

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

[4] J.H. Collet, D. Litaize, J. V. Campenhut, C. Jesshope, M. Desmulliez, H. Thienpont, J. Goodman, and A. Louri, "Architectural approaches to the role of optics in mono and multiprocessor machines," *Applied Optics, Special issue on Optics in Computing*, vol. 39, pp. 671–682, 2000.

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

[6] Joon-Ho Ha and T.M.Pinkston, "The speed cache coherence for an optical multi-access interconnect architecture," in *Proceedings of the 2nd International Conference on Massively Parallel Processing Using Optical Interconnections*, 1995, pp. 98–107.

[7] 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.

[8] 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 *Proceedings of the 39th International Symposium on Microarchitecture*, December 2006.

[9] A. Shacham, B.A. Small, O. Liboiron-Ladouceur, and K. Bergman, "A fully implemented 12x12 data vortex optical packet switching interconnection network," *Journal of Lightwave Technology*, vol. 23, pp. 3066–3075, Oct 2005.

[10] 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.

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

[12] Li Shang, Li-Shiuan Peh, and Niraj K. Jha, "Dynamic voltage scaling with links for power optimization of interconnection networks," in *Proceedings of the 9th International Symposium on High Performance Computer Architecture*, November 2003.

[13] 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.

[14] Qiang Wu, Philo Juang, Margaret Martonosi, Li-Shiuan Peh, and Douglas W. Clark, "Formal control techniques for power-performance management," *IEEE Micro*, vol. 25, no. 5, September/October 2005.

[15] E.J.Kim, K.H.Yum, G.M.Link, N.Vijaykrishnan, M.Kandemir, M.J.Irwin, M.Yousif, and C.R.Das, "Energy optimization techniques in cluster interconnects," in *Proceedings of the 2003 International Symposium on Low Power Electronics and Design (ISLPED 03)*, August 2003.

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

[17] Osman Kibar, A. Van Blerkom, Chi Fan, and Sadik C. Esener, "Power minimization and technology comparisons for digital free-space optoelectronic interconnections," *IEEE Journal of Lightwave Technology*, vol. 17, pp. 546–555, April 1999.

[18] 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 x 16 vcsel array flip-chip bonded to cmos vlsi circuit," *IEEE Photonics Technology Letters*, vol. 12, no. 8, pp. 1073–1075, August 2000.

[19] A. Lindstrom, "Parallel links transform networking equipment," *FiberSystems International*, pp. 29–32, February 2002.

[20] A. Apsel and A. G. Andreou, "Analysis of short distance optoelectronic link architectures," in *Proceedings of the 2003 International Symposium on Circuits and Systems*, May 2003.

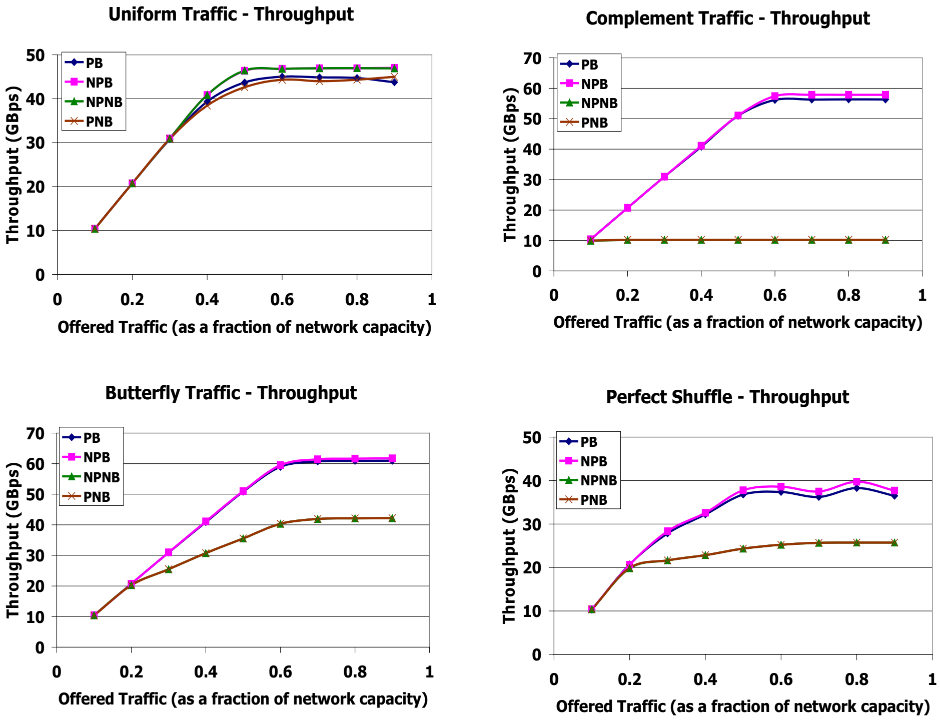


Figure 7: Throughput for a 64 node E-RAPID configuration implementing NP-NB, NP-B, P-NB and P-B for Uniform, Complement, Butterfly and Perfect shuffle traffic patterns.

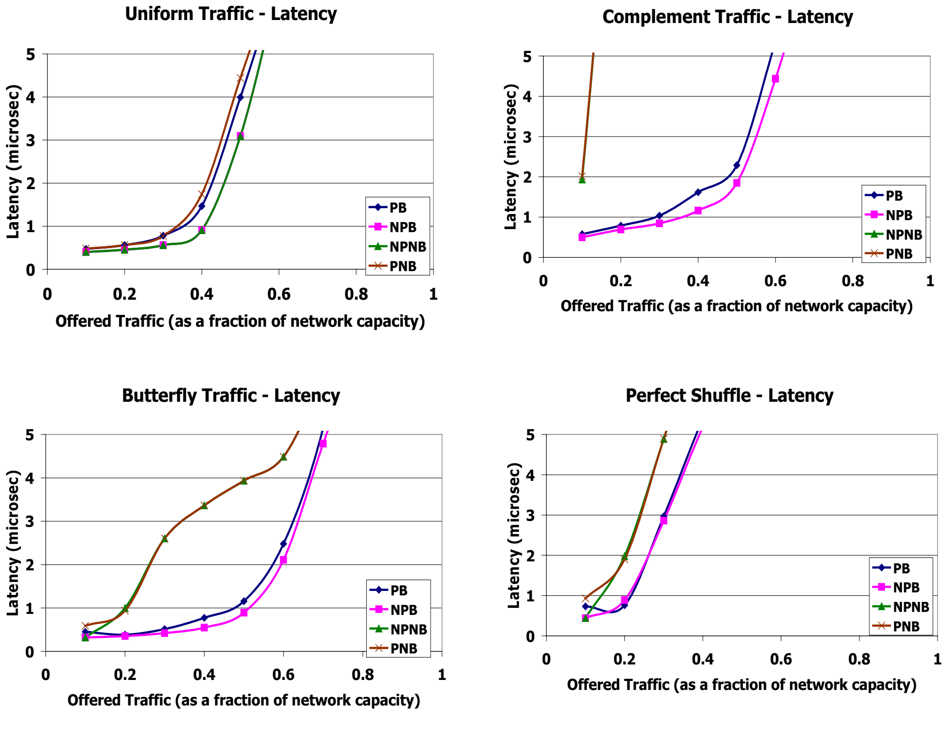


Figure 8: Average latency for a 64 node E-RAPID configuration implementing NP-NB, NP-B, P-NB and P-B for Uniform, Complement, Butterfly and Perfect shuffle traffic patterns.

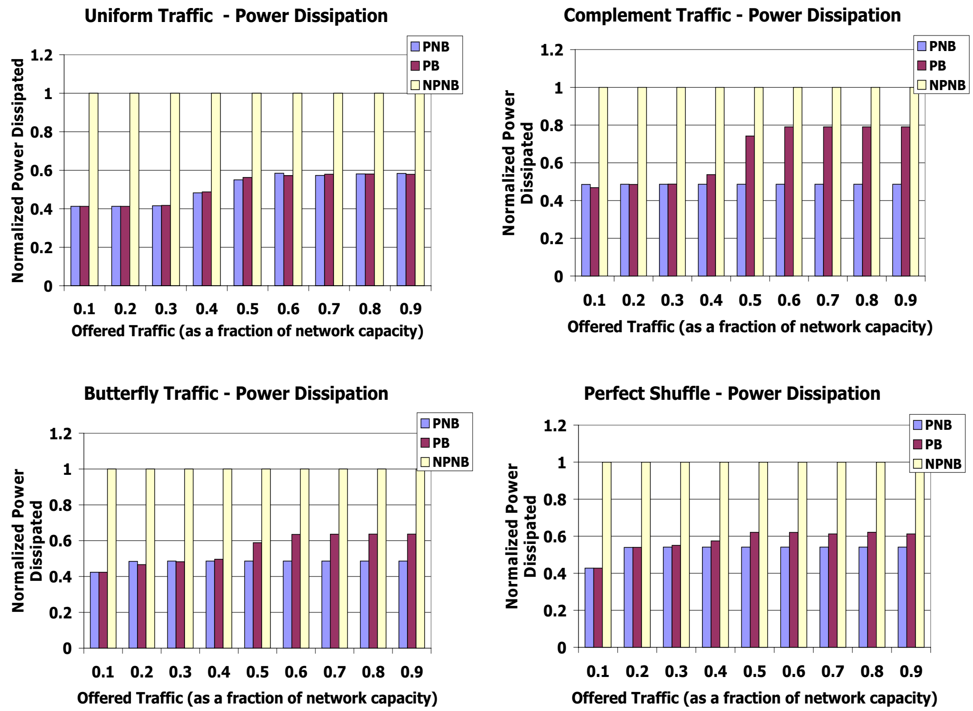
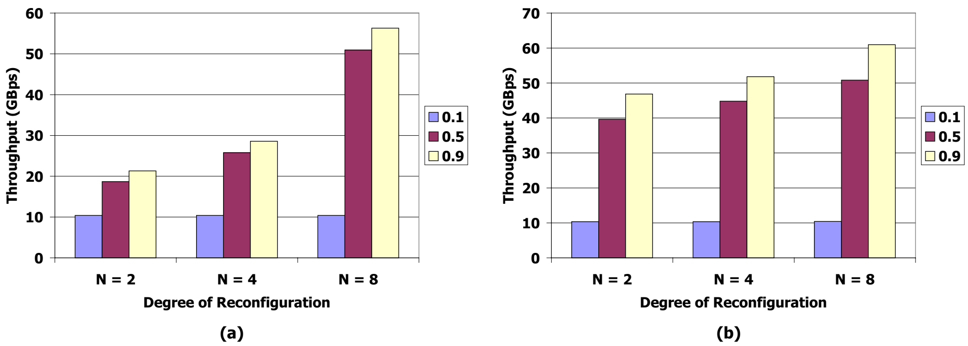


Figure 9: Power consumption for a 64 node E-RAPID configuration implementing NP-NB, NP-B, P-NB and P-B for Uniform, Complement, Butterfly and Perfect shuffle traffic patterns.



(a)

(b)

Figure 10: Degree of reconfiguration at network loads of 0.1, 0.5 and 0.9 for (a) Complement traffic and (b) Butterfly traffic patterns.