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

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

Abstract

As communication distances and bit rates increase, opto-electronic interconnects are becoming de-facto standard 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 adversial traffic patterns. Dynamic bandwidth reconfiguration 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 a dynamically reconfigurable architecture called E-RAPID (Extended-Reconfigurable, All-Photonic Interconnect for Distributed and parallel systems) that not only dynamically reallocates bandwidth, but also reduces the power consumption for all traffic patterns. Our proposed LS (Lock-Step) reconfiguration technique combines Dynamic Power Management (DPM) with DBR techniques, achieving a reduction in power consumption of 25% - 50% while degrading the throughput by less than 5%.

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]. This has given rise to opto-electronic networks can that support greater bandwidth through a combination of efficient multiplexing techniques for board-to-board and rack-to-rack interconnects. 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.

Static allocation of wavelengths in optical interconnects offers every node with equal opportunity for interprocessor communication. In our previously proposed RAPID (Reconfigurable All-Photonic Interconnect for Distributed and parallel computing systems)[6], the routing and wavelength assignment (RWA) allocated bandwidth statically between various communicating boards using different wavelengths, fibers and time-slots. While static allocation improved performance for benign traffic patterns, the network congests for adversial traffic patterns due to uneven resource utilization. On the other hand, 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 elements[5], time-slots based bandwidth re-allocation[7] and both time and space based bandwidth switching[8].

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 consume a sizeable fraction of the system power budget (for example, 70% of the switch power budget in IBM Infiniband 8-port 12X switch[9, 10]), researchers have proposed several power-aware techniques to optimize power consumption for HPC systems. Dynamic power reduction techniques such as DVS (Dynamic Voltage Scaling)[11, 12, 13] and DLS (Dynamic Link Shutdown)[14] have been suggested for electrical networks. In DVS, 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 heavily used and turns up the link when

1-4244-0910-1/07/$20.00

needed. In [12], power-aware opto-electronic network design space is explored by regulating power consumption in response to actual network traffic. However, this work enables efficient power regulation 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 tradeoff by balancing power consumption with improving network performance. This enables reducing communication bottlenecks, while optimizing resource utilization leading to *balanced-improved* system architecture design.

In this paper we propose a dynamically reconfigurable optical interconnect called E-RAPID (extended-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) techniques (locally controlled) such as DVS and DLS are applied in conjunction with Dynamic Bandwidth Re-allocation (DBR) techniques (globally controlled) 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/link independently makes decisions, and (2) Re-allocation of bandwidth happens between any system boards without affecting the on-going communication in the overall system.

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.

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 $\lambda_{(d-s)}^{(s)}$ if $s > d$, where B is the total number of boards in the system[6]. 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 computation (RC), virtual-channel allocation (VA) and per-flit steps include switch allocation (SA) and switch traversal (ST)[15]. 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 feasi-

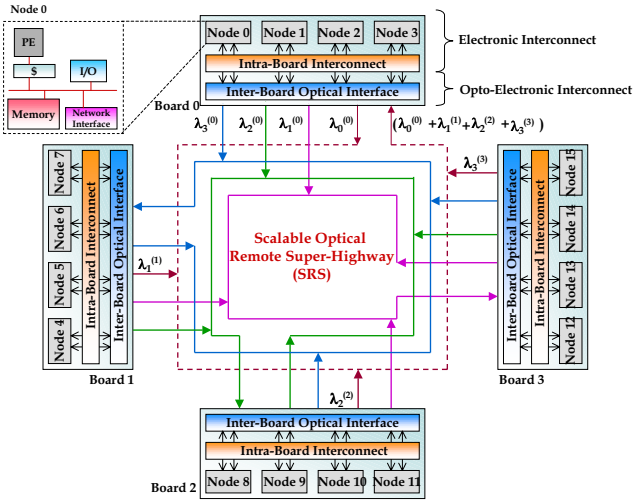

Figure 1. Routing and wavelength assignment in E-RAPID for inter-board communication.

ble, flit management across multiple domains is extremely complicated.

2.2 Technology for Reconfiguration

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 communications 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.

3 Bandwidth-Power Dynamic Reconfiguration

In this section, we describe the implementation of LS technique. To provide more insight into reconfiguration mechanisms, consider Figure 3 which shows various combination of power/non-power aware and bandwidth/non-bandwidth reconfigured network design. The total power consumption of an opto-electronic link scales with the supply voltage (V_{DD}) as well as with the bit rate (BR)[12]. Increasing the bit rate consumes more power as both the voltage and bit rate increases. Suppose, we have 3 power levels, power-low P_L, power-mid P_M and power-high P_H as shown on the left y-axis and the link utilization (measures the amount of time the link is used) corresponding to 3 levels utilization-low U_L, utilization-mid U_M and utilization-high U_H as shown on the right y-axis. Figure 3(a) shows the Non-Power Aware Non-Bandwidth Reconfiguration (NP-NB). 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 Power-Aware Non-Bandwidth Reconfiguration (P-NB) where the link utilizations are regularly measured at the power reconfiguration window, R_w = R_p. P-NB allows link power to scale with utilization, thereby improving the performance at high utilization and saving power at low utilization. If the bandwidth demands further increase, there is no provision for further improving performance. Figure 3(c) shows the Non-Power Aware, Bandwidth Reconfiguration (NP-B), in which depending on the availability of the idle links, performance can be improved by providing additional bandwidth. This can be achieved by monitoring the utilization at bandwidth reconfiguration window, R_w = R_B and thereby re-allocating

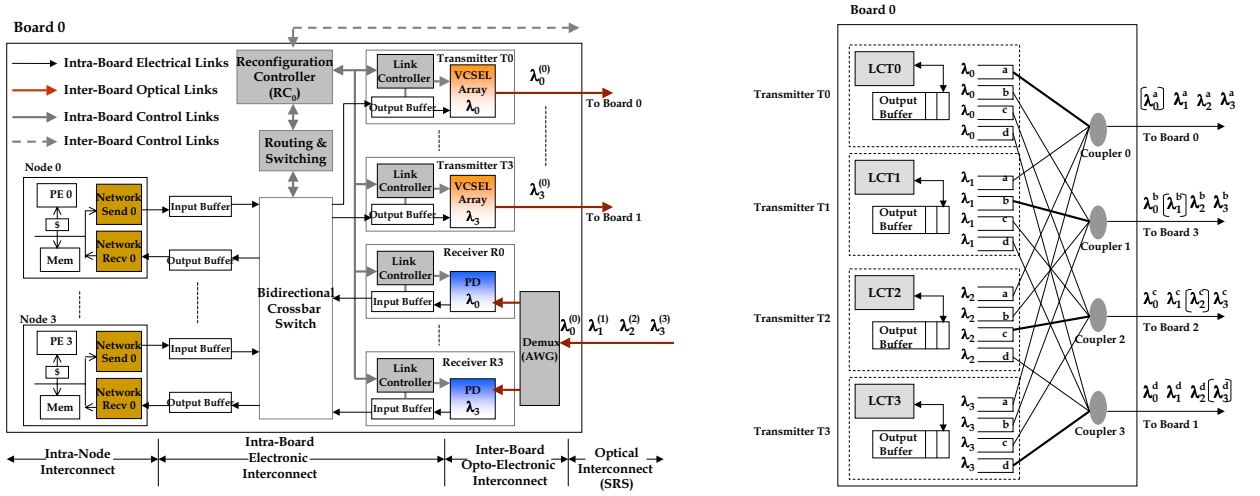

Figure 2. (a) The proposed on-board interconnect for the E-RAPID architecture with reconfiguration controller (RC) and link controllers (LC). (b) The proposed technology for reconfiguration using passive couplers and array of lasers per transmitter port.

idle link bandwidth. While this achieves improved performance, NP-B consumes double the power as shown in the Figure 3(c). Figure 3(d) shows Power Aware Bandwidth Reconfiguration (P-B) where it balances power consumption with bandwidth re-allocation, i.e. it combines power-awareness with bandwidth reconfigurability, thereby improving performance while consuming less power.

In this paper, we propose Lock-Step (LS) technique that 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 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. This reconfiguration phase could be for power-awareness in the network or for bandwidth re-allocation. 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. Each LC is associated with an optical transmitter to measure link statistics, and with an optical receiver to turn on/off the receiver. 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[12]. All these statistics are measured over a sampling time window called *Reconfiguration window* or phase, R_w. Link_{util} provides accurate information regarding whether a link is being used at all, at low-medium network loads, where as Buffer_{util} provides accurate information regarding network congestion at medium-high network load. In what follows, we first explain how power-awareness is implemented, then how bandwidth reconfiguration is implemented and lastly, how they are implemented together.

3.1 Dynamic Power Management (DPM)

An optical link in E-RAPID architecture consists of the transmitter, the receiver and the channel. The total power consumption of an optical link is comprised of the transmitter and receiver power. Transmitter power is consumed at the laser, and laser driver/modulator, where as the receiver power is consumed at the photodetector, transimpedance amplifier (TIA) and clock and data recovery (CDR) circuitry[16]. While both Multiple-Quantum Wells (MQW)[16] with external modulators and VCSELs (vertical-cavity surface emitting lasers)[17, 16] 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 are

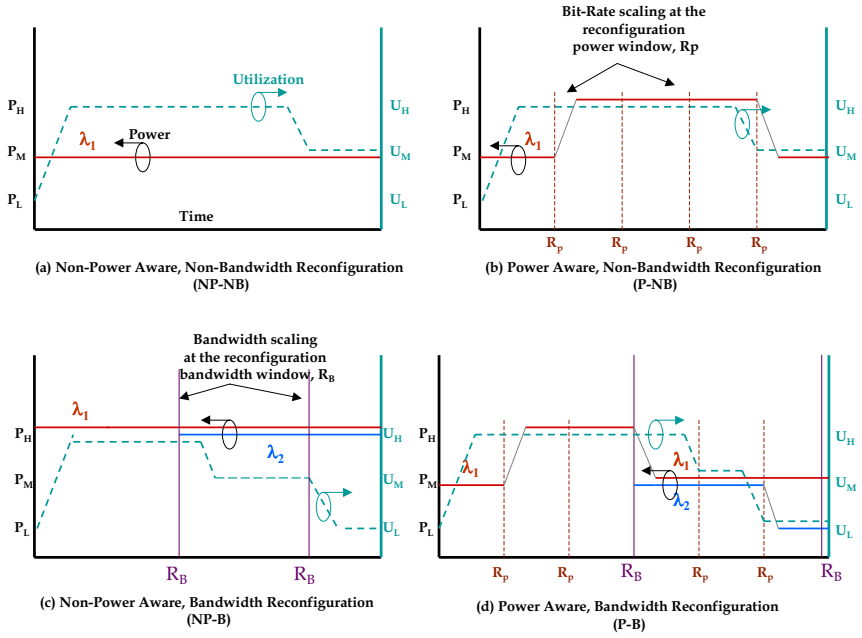

Figure 3. Design space of power-aware, and bandwidth reconfigurability.

used for reconfiguration in E-RAPID. The power scaling trends with supply voltage (V_{DD}) and bit rate (BR) for various optical link components are as follows: VCSEL (V_{DD}), VCSEL driver (V_{DD}^2. BR), TIA ($V_{DD}.BR$) and CDR ($V_{DD}^2.BR$)[12, 16]. When the bit rate scales down, the supply voltage is also reduced of all the above components, resulting in power savings.

In a VCSEL-based transmitter, both the bit-rate and the supply voltage can be controlled by the modulation current from the VCSEL driver, which results in a linear reduction in output optical power. At the receiver, the supply voltages and bit rates can be scaled to save power in the photodetector, TIA and CDR. Scaling the power level focuses on reducing the delay incurred during the slow voltage transitions as compared to frequency transitions[12, 11]. 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. In our power-aware opto-electronic network, different bit rates correspond to different power levels. We consider 3 power levels P_{low}, P_{mid} and P_{high} corresponding to bit rates 2.5 Gbps, 3.3 Gbps and 5 Gbps. While 10 Gbps VCSELs are available, we consider these 3 power levels to match the slower electrical on-board link rates.

Dynamic Power Regulation Algorithm: The power-awareness cycle is triggered by the RC on every system board every R_w. Each RC_j, j = 0,1,...B-1 sends to LC_i, i = 0,1,...D-1, $Power_{Request}$ control packet. When every LC_i receives the packet, it measures the link utilization Link_{util}, and buffer utilization Buffer_{util} for the prior reconfiguration window, R_w and forwards the $Power_{Request}$ to the next LC_{i+1}. The $Power_{Request}$ control packet is finally received by the RC which completes the power-aware cycle. LC_i then decide to scale the bit rates based on link thresholds, L_{min} and L_{max} and buffer threshold B_{max}. If the Link_{util} falls below L_{min}, LC_i scales the bit rate down to the the next power level. If the Link_{util} exceeds L_{max}, LC_i scales the bit rate up to the next power level. If the Link_{util} falls between L_{min} and L_{max}, it retains the same bit rate. 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. The bit rate scaling is locally controlled by the LC. The RC does not receive any information regarding the state of the LCs during the power-aware reconfiguration cycle.

While multiple bit rates can conserve more power by finely tuning the bit rates to the link 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 to be 2000 simulation cycles. By using only 3 power levels in our system architecture, we avoid multiple bit rate transitions. Moreover, we aggressively push the link utilization to the limit. For example, setting the L_{max} to be 0.9 and L_{min} to be 0.7 allows the link to be fully utilized. This ensures that for low loads, we keep decreasing the bit rate until the utilization falls between L_{min} and L_{max}. For medium load, the link is well utilized and we are on the verge of saturating the link. We increase the bit rate if the Link$_{util}$ is greater than L_{max}. Similarly, at high load, we operate at the highest power level. Now, instead of simply scaling the bit rate if the Link$_{util}$ exceeds L_{max}, we incorporate additional power savings by not only saturating the link, but also waiting until the buffer utilization exceeds B_{max}. The bit rate is scaled up only if the link threshold exceeds both L_{max} and B_{max}. As the network link is saturated at high loads, additional power savings can be obtained by reducing the bit rates.

3.2 Dynamic Bandwidth Re-allocation (DBR)

In order to implement DBR, 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 reconfigure, this information is disseminated back to the RCs on other boards as well as the local LCs. Each RC_i is connected 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 λ_0, λ_1, λ_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 Λ = λ_0, λ_1, λ_2 ... λ_{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 re-allocate 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 bandwidth reconfiguration phase. Each LC_i updates the queue statistics Link$_{util}$, and 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 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 Board$_{Request}$ from other RC_i (shown in dotted lines). For example, when board 0 receives Board$_{Request}$ from say board 1, it will update the field for wavelength with which board 0 communicates with board 1, i.e. λ_3 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.

Reconfigure Stage: Now, each RC_i computes if reconfiguration is necessary based on two buffer thresholds, minimum threshold B_{min} and maximum threshold B_{max}. While profiling of traffic traces can provide more accurate information regarding when the network is actually congested, setting the B_{max} to 0.3 is fairly reasonable for most traffic scenarios. This implies that on an average 30% of our buffers are occupied by packets for the given reconfiguration window R_w. We set B_{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 if Buffer$_{util}$ is less than B_{min} (implying that this wavelength can be re-allocated), normal utilized if Buffer$_{util}$ falls between B_{min} and B_{max} (implying the wavelength is well utilized) and over-utilized if Buffer$_{util}$ is greater than B_{max} (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

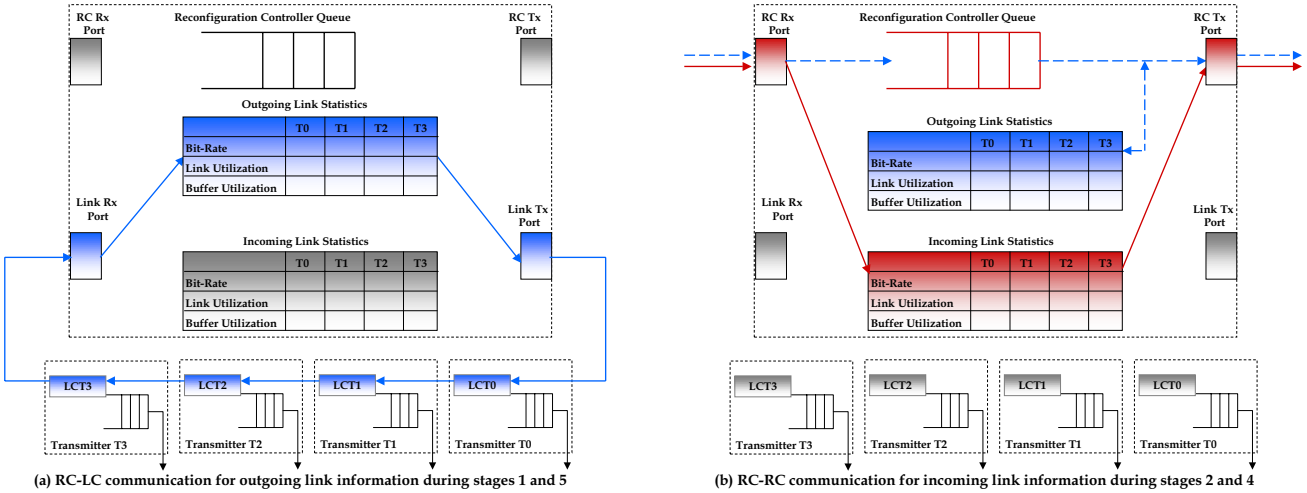

Figure 4. Reconfiguration algorithm implementation.

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. The power-bandwidth reconfiguration algorithm is implemented every R_w by the board reconfiguration controller RC_i. We implement $odd - even$ reconfiguration, where every odd cycle R_w = 1, 3, 5 ..., RC_i triggers power-awareness cycle and every even cycle, R_w = 2, 4, 6, ... the bandwidth reconfiguration cycle is triggered. The reason for doing in this manner is that the power-awareness can be implemented locally, due to one-to-one mapping between the transmitter and the receiver. The bandwidth reconfiguration needs to be implemented globally, knowing all the idle links in the network. As the links that are idle are generally turned off during the power-awareness cycle, power scaling cannot be implemented during the bandwidth reconfiguration cycle. Moreover, as we increase the bit rate only if both Link_{util} and Buffer_{util} exceeds L_{max} and B_{max}, it provides a incremental increase in bandwidth. First, the bit rate is scaled. If that does not stabilize the utilization, DBR allocates spare resources to improve performance.

4 Performance Evaluation

The performance of E-RAPID is evaluated using YACSIM[18] and NETSIM discrete-event simulator and is compared to various non-power/power, 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[15]. 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.

4.1 Simulation Network Parameters

The electrical network router model parameters are shown in Table 1. These parameters reflect the design from SGI Spider routing chip[19]. For the router model designed, the channel width is 16 bits and 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 implemented for a single flit buffer with credits incurring a single cycle channel delay. Routing computation, virtual channel and switch allocation, each takes one router clock cycle. For the optical network, we assume bit rates of 2.5, 3.3 and 5 Gbps. For most of the runs, we maintained a constant packet size of 64 Bytes, resulting in a 8 flit packet size.

At 5 Gbps, the total power consumption of an optical link is 43.03 mW operating at a supply voltage of 0.9 V. The total transmitter power consumed in a link is given by

the sum of VCSEL power and the VCSEL driver power. For an implant VCSEL with a slope efficiency of 0.42 A/W, the modulation current I_m is calculated as 16.6mA and the power consumption is estimated to be 1.5 μW for transmitting a packet of size 64 bytes[16]. The power consumed in a VCSEL driver with a capacitance of C_{driver} 0.62 pF is 1.23 mW[12]. At the receiver side, the photodetector power consumption is calculated as 1.4 μW. The TIA operating at drain-source current of I_{ds} 27.8 mA consumes 25.02 mW[20] and the CDR with a capacitance of C_{CDR} 9.26 pF consumes 17.05 mW of power[12]. Similarly, the minimum operating voltage at 2.5 Gbps is 0.45 V and the total power consumption is 8.6 mW, where as at 3.3 Gbps, the supply voltage is 0.6 V and the total power consumed is 26 mW. These values are shown in Table 1. The CDR delay was estimated from [12], which was normalized to our network clock cycle. In [12], 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. The slower voltage transitions across adjacent levels took 65 clock cycles. In our network simulation, after the control bit rate packet is transmitted, the transmitter conservatively disables the link for 65 cycles.

The performance of E-RAPID was compared to other electrical networks 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. The performance of E-RAPID was compared on the basis of throughput, latency and power consumed.

4.2 Results and Discussion

Throughput, Latency, Power: Figures 5 and 6 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. Due to space constraints, we show the performance for only 64 node network. For uniform traffic, NP-NB (non-power aware non-bandwidth reconfigured) shows similar performance (throughput and latency) as NP-B (non-power aware, bandwidth reconfigured). 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. 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 on-going communication. For P-NB (power aware, non-bandwidth reconfigured) network, there is a marginal degradation in performance (less than 3%) as the network attempts to regulate the power. However for P-B (power aware bandwidth reconfigured) network, there is degradation in throughput of 8%. The power consumptions for both NP-NB and NP-B are identical, as there is no power awareness in the network. However, P-NB and P-B show different power consumption. P-NB shows almost 16% reduction on power consumption where as P-B shows almost 50% reduction in power consumption. In P-NB, the B_{max} is kept at 0.0 and L_{max} is 0.7, as opposed to P-B, where the B_{max} is set to 0.3 and L_{max} is 0.9. In P-NB, the links are not allowed to completely saturate as there are no additional links/bandwidth to provide in case they are saturated. Therefore, we conservatively increase the bit rate when it is about to saturate.

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, network latency and power consumption 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 and latency. We achieve almost 400% improvement in throughput by completely reconfiguring the network. Similarly, the power consumption for a NP-B network is also 300% more than the NP-NB/P-NB networks. However, for P-B networks, while the performance is almost similar to the NP-B, the power consumption is reduced by 25% over NP-B networks. P-B networks consume almost double the power consumption, but provide four fold improvement in performance. Similar performance trends can be seen for perfect shuffle and butterfly as shown in Figure 6. For butterfly traffic pattern, NP-B provides 25% improvement in performance, but consumes almost double the power consumption. P-B, on the other hand, provides similar performance improvement of 25% where as consumes 1.5 times that of NP-NB and P-NB. For perfect shuffle patterns, the throughput improves by 1.7 times by using NP-B and P-B, where as power consumption increases by 70% and 25% for NP-B and P-B. In E-RAPID architecture, power and bandwidth reconfiguration 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 simply the bandwidth to be reconfigured, 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.

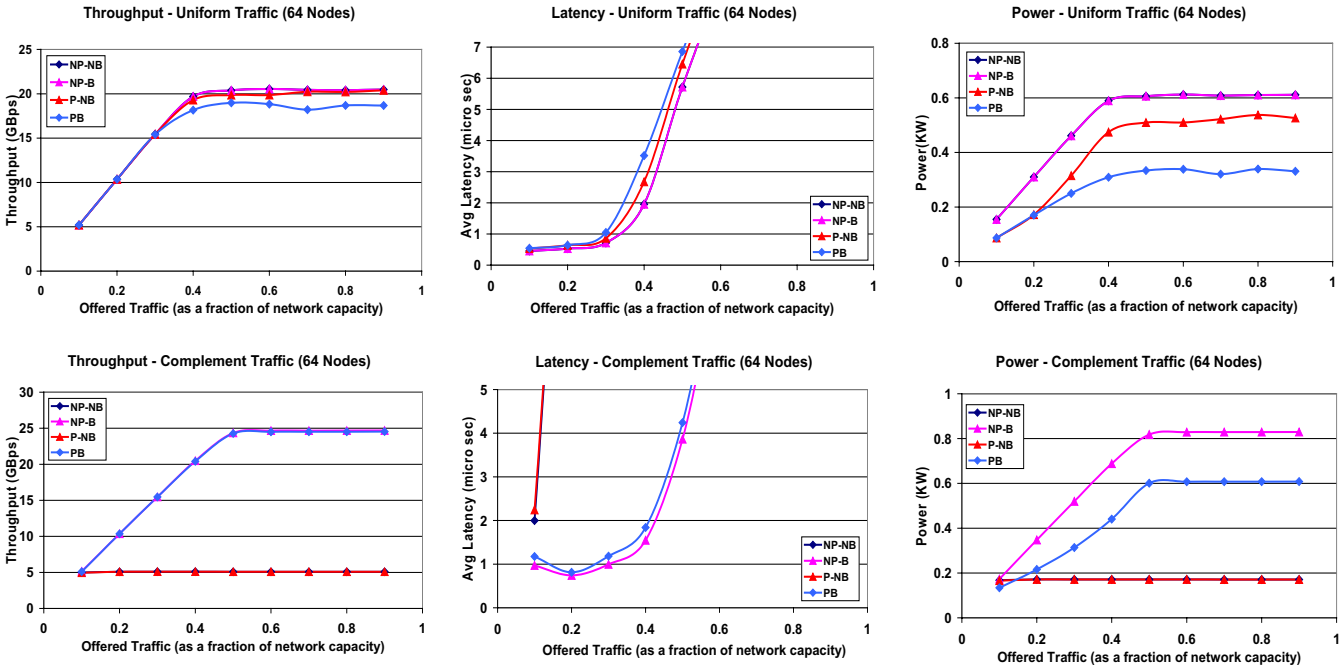

Figure 5. Performance-Power consumption for a 64 node E-RAPID configuration implementing NP-NB, NP-B, P-NB and P-B for Uniform and Complement traffic patterns.

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 in terms of throughput and latency, 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 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.

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

References

[1] A. F. Benner and et.al, "Exploitation of optical interconnects in future server architectures," in *IBM Journal of Research and Development*, 2005, pp. 755–776.

[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.

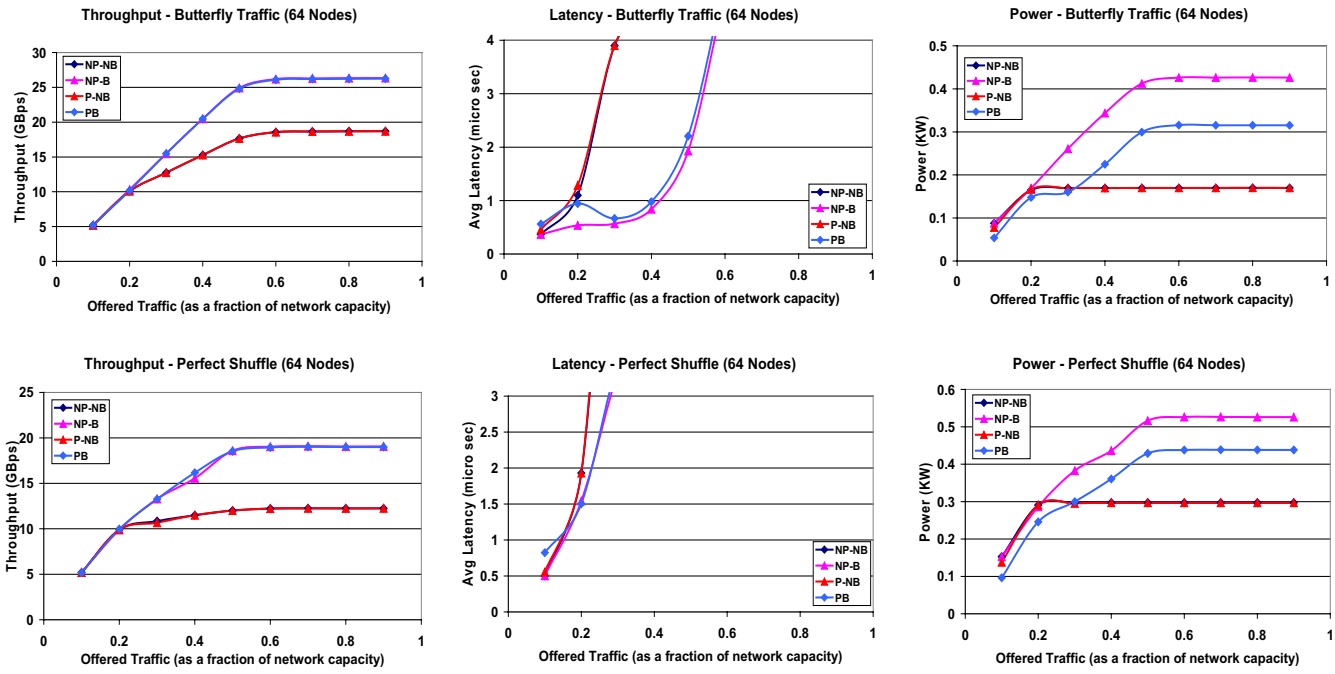

Figure 6. Performance-Power consumption for a 64 node E-RAPID configuration implementing NP-NB, NP-B, P-NB and P-B for butterfly and perfect shuffle traffic patterns.

[6] Avinash Karanth Kodi and Ahmed Louri, "Design of a high-speed optical interconnect for scalable shared memory multiprocessors," *IEEE Micro*, vol. 25, pp. 41–49, Jan/Feb 2005.

[7] 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.

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

[9] The Infiniband Trade Alliance architecture, "http://www.inifiniband.org," .

[10] Shubhendu S. Mukherjee, Peter Bannon, Steven Lang, Aaron Spink, and David Webb, "The alpha 21364 network architecture," *IEEE Micro*, vol. 22, no. 1, January/February 2002.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

[17] A.V.Krishnamoorthy and et.al., "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.

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

[19] Mike Galles, "Spider: A high-speed network interconnect," *IEEE Micro*, vol. 17, pp. 34–39, Jan/Feb 1997.

[20] Daniel A. Van Blerkom, Chi Fan, Matthias Blume, and Sadik C. Esener, "Transimpedance reciever design optimization for smart pixel arrays," *Journal of Lightwave Technology*, vol. 16, January 1998.