

Exploring the Design Space of Power-Aware Opto-Electronic Networked Systems

Xuning Chen[†] Li-Shiuan Peh[†] Gu-Yeon Wei[‡] Yue-Kai Huang[†] Paul Prucnal[†]

[†]Dept. of Electrical Engineering, Princeton University,
Princeton, NJ 08544

[‡]Dept. of Electrical Engineering & Computer Science,
Harvard University, Cambridge, MA 02138

Abstract

As microprocessors become increasingly interconnected, the power consumed by the interconnection network can no longer be ignored. Moreover, with demand for link bandwidth increasing, optical links are replacing electrical links in inter-chassis and inter-board environments. As a result, the power dissipation of optical links is becoming as critical as their speed. In this paper, we first explore options for building high speed opto-electronic links and discuss the power characteristics of different link components. Then, we propose circuit and network mechanisms that can realize power-aware optical links – links whose power consumption can be tuned dynamically in response to changes in network traffic. Finally, we incorporate power-control policies along with the power characterization of link circuitry into a detailed network simulator to evaluate the performance cost and power savings of building power-aware opto-electronic networked systems. Simulation results show that more than 75% savings in power consumption can be achieved with the proposed power-aware opto-electronic network.

1 Introduction

Computer systems are increasingly composed of subsystems connected with an interconnection network fabric – such as clusters of PCs, servers composed of compute and/or storage blades, supercomputers built from boxes and boards of microprocessors. Due to tight cooling budgets, power is becoming the key constraint limiting scalability in these systems. With link circuitry consuming a significant portion of the system power budget (60% of the line card power budget in the Avici TSR router [4], and 70% of the switch power budget in the IBM InfiniBand 8-port 12X switch [6]), there is a clear incentive to focus on improving the power efficiency for this part of the system.

As bandwidth demands increase, opto-electronic links are becoming the de-facto interconnect between boxes, and moving into the board-to-board domain as well. While an optical link enables high bit rates, it does not ease the power consumption problem, prompting companies and researchers to find ways to reduce power in optical links [22]. In this paper, we investigate a power-aware architecture that uses power-control policies to dynamically control the bit rate and power consumption of opto-electronic links.

Power-aware networks that regulate their power consumption in response to actual traffic utilization were first proposed in [24], which explored the use of dynamic voltage scalable (DVS) electrical links in networks, with routers controlling and setting the link bit rates. Since then, there have been various studies on power-aware networks – exploring the impact of DVS on on-chip interconnects' transmission energy

and bit error rate [30], exploring power-aware networks where electrical links are turned completely on and off [26], investigating the control of DVS links in clusters of workstations through routing table reconfiguration [11]. These prior works focused on network design, exploring different policies for the controlling of the power-aware network, glossing over the circuit design issues of a power-aware link, and its impact on the control policy and the overall network architecture.

The motivation to explore the possibility of designing power-aware opto-electronic network systems is two fold. First, as with electrical links, opto-electronic links consume significant power, and have a power profile that does not vary significantly with actual utilization, so the interconnection fabric in these systems can consume high power even when the system is lightly-loaded. Second, since real-life network traffic exhibits substantial temporal and spatial variance [14, 26], opto-electronic networks that can regulate their own power consumption at run-time by tuning their bit-rate and supply voltage with respect to network traffic stand to gain significant power savings.

In this paper, we explore the design space of power-aware opto-electronic networks from the bottom up, first investigating ways to incorporate run-time power-control into each component of a link, understanding the limits, then designing and architecting a complete power-aware networked system with routers that dynamically control the power consumed by the opto-electronic links. Section 2 dives into the various components of an opto-electronic link to show how each works, their power characteristics, and ways to incorporate power-control into each component. Section 3 presents a complete power-aware opto-electronic network design, which differs from traditional network design in both topology as well as router microarchitecture. To evaluate this design, network simulations are run using both synthetic and actual traffic traces. The simulation setup and results are shown in Section 4. Section 5 concludes the paper and discusses our next steps in prototyping the proposed system. This paper is the result of a collaborative effort between optics, electronics, and networks researchers to make power-aware opto-electronic networks a reality.

2 Power-aware opto-electronic link design

Fig. 1 presents the overall architecture of a typical opto-electronic link. First, at the link transmitter, serialized electronic data forms the input to the electrical *modulator driver* which generates the respective voltage signals corresponding to 1s and 0s, controlling the *modulator* to switch the light from the *laser source* “on” and “off”. The modulated optical signal is then transmitted to the receiver through an optical fiber. At the receiver, this optical signal feeds a *photodetector* which converts the optical bit stream back into electrical current signals. These current signals are then transformed into an

amplified voltage signal via a *transimpedance amplifier (TIA)*. Finally, the *clock and data recovery (CDR)* circuit tracks the amplified voltage signal and extracts the digital 1s and 0s.

In the subsequent subsections, we outline design considerations for applying dynamic power control mechanisms to an opto-electronic link. For each link component, we first explain how it works, characterize what affects its power dissipation, then propose ways to incorporate dynamic power-control for a power-aware network system. Our proposals for incorporating power-control into opto-electronic links build upon prior research of variable-frequency electronic links [28, 12]. Essentially, there are two ways of realizing power-awareness – controlling just the link bit rate, or controlling both bit rate and supply voltage, with respect to the network traffic. We'll discuss the implications of these alternatives for every link component.

2.1 Transmitter: Laser source and modulator

There are two basic alternatives for implementing the light source opto-electronic links used in board-to-board and box-to-box networks: (1) Vertical cavity surface emitting lasers (VCSELs), where light is generated on chip and directly modulated by electrical drive currents; (2) An external laser source housed in a separate chassis, feeding the hundreds to thousands of transmitters within the system, with multiple-quantum-well (MQW) modulators [16] that switch light “on” and “off” based on electrically driven signals.

2.1.1 Transmitter: Directly modulated VCSELs & driver

Operation. VCSELs are controlled electronically – when the input driving current is above a certain threshold (I_{th}) the VCSEL is stimulated and it emits light. At high bit rates, this threshold not only affects light generation, but also the time required for the stimulated emission to stabilize [16]. Therefore, a VCSEL is usually constantly biased at a current above this threshold (I_{bias}). The VCSEL driver modulates the driving current to the VCSEL based on the input bit patterns, so that the driving current, I , is $I_m + I_{bias}$ (I_m is the modulation current) for bit 1, and just I_{bias} for bit 0. The VCSEL then converts 1s and 0s into high and low light intensities, respectively.

The design of the VCSEL driver can be quite simple. As shown in Fig. 2, it simply consists of a string of cascaded inverters, where the size of each inverter is β (a constant factor typically between 3 and 4) times the size of the previous one to control the propagation delay, when driving a large load.

Power characteristics. The power consumption of a VCSEL hinges on the threshold current, as that is the minimal fixed power consumption regardless of activity or the transmitted bit patterns. Once biased above the threshold, the emitted optical power of a VCSEL, P_e , grows linearly with the driving current, I :

$$P_e = S \cdot (I - I_{th}) \quad (1)$$

where I_{th} is the threshold current, and S is the slope efficiency that defines the conversion ratio in *Watts/Amps*.

When assuming equal probabilities of 1s and 0s, the power consumption by a VCSEL is:

$$P_{VCSEL} = [I_{bias} + I_m/2] \cdot V_{bias} \quad (2)$$

where $I_{bias} + I_m/2$ is the average driving current.

For the VCSEL driver, dynamic power is consumed to charge/discharge the capacitance of the inverter chain for

every data transition. We can model the driver power as:

$$P_{VCSEL\ driver} = \alpha_1 \cdot C_{LD} \cdot V_{dd}^2 \cdot BR \quad (3)$$

where α_1 is the switching efficient (the probability of bit transitions) of the input data stream, BR is the bit rate of the link, and C_{LD} is the total switched capacitance (the sum of the capacitance within the laser driver and the gate capacitance of NMOS $N1$).

Dynamic power control. The electrical laser driver's power can be dynamically controlled through bit-rate and voltage scaling, its power consumption has a scaling trend close to $V_{dd}^2 \cdot BR$. However, power control of the laser driver has an effect on the VCSEL's output light intensity. Scaling of its V_{dd} along with bit rate affects I_m to VCSEL. With fixed V_{bias} , the scaling of I_m with voltage controls VCSEL's power consumption as shown in Eq. 2 and output light intensity.

2.1.2 Transmitter: External laser source with MQW modulator

Operation. In this scheme, the laser source is housed separately from the system, in an external chassis, with its own power supply and cooling [21]. Light is directed from this central laser source to the transmitters at each link of the system through optical fibers. Since the maximum optical power (the light intensity) required for each link is only tens or hundreds of μW (25 μW at receiver for 10 Gb/s link), a typical mode-locked laser can support up to hundreds or even thousands of links [20]. So, optical power (light) can be split and distributed to each link using static optical power splitters, such as fused-fiber optical couplers [17], which introduce very low insertion loss (a maximum of 13.6dB for 1 to 16 splitting)¹.

At each link transmitter of the system, a multiple-quantum-well (MQW) modulator receives light from the external mode-locked laser. Based on the electrical voltage applied to the modulator, which is controlled by the modulator driver, the modulator absorbs the light (for data value 0 – “off” state), or allows light to pass through (for data value 1 – “on” state). A MQW modulator is characterized by its capacitance, insertion loss (IL) and contrast ratio (CR), where insertion loss is the amount of optical power that is lost upon passing through the modulator for “on” state, and contrast ratio is the ratio of the optical power that passes through the modulator for the “on” and “off” states.

The modulator driver amplifies the serialized signal to drive the modulator. The simplest driver can be a string of cascaded inverters as shown in Fig. 2 which is similar to the inverter chain for VCSEL driver. For input data 0, voltage applied to modulator is V_{bias} , for bit 1 voltage applied to the modulator is $V_{bias} - V_{dd}$. A large V_{dd} is desirable for a high contrast ratio.

Power characteristics. Given that the laser source is external to the network system, we assume that its power is not part of the system's power budget, nor contributes to its cooling costs. In this transmitter scheme, we ignore the laser's power consumption and only characterize the power dissipated by the modulator driver and the modulator.

Power dissipation in the modulator is due to the *absorbed* optical power. The modulator dissipates more power in the “off” state, because much more light is absorbed. Eq. 4

¹Insertion loss refers to the amount of light lost as a result of splitting. For instance, a 10% insertion loss indicates that 10% of the optical power is lost through splitting, so if the original optical power is 0dB, the resulting optical power after 1 to 16 splitting is -12dB.

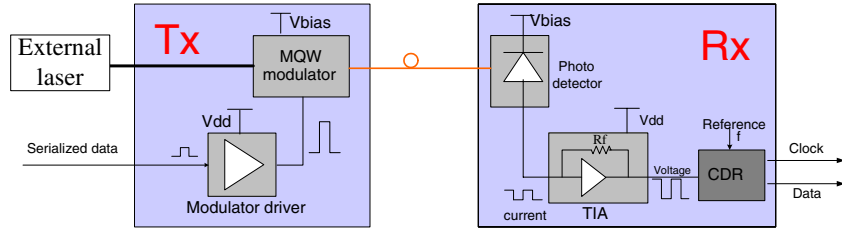


Figure 1. Opto-electronic link architecture.

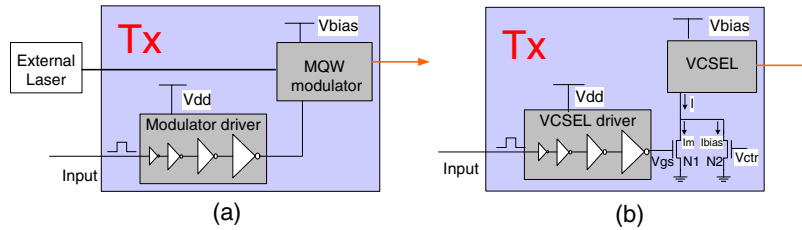


Figure 2. Alternatives for laser source and modulators. (a) An external laser source with an on-chip MQW modulator. (b) A directly modulated VCSEL laser source.

shows the average power assuming the same probability of 1s and 0s.

$$P_{modulator} = 0.5 \cdot R_s \cdot P_I \cdot [IL \cdot (V_{bias} - V_{dd}) + (1 - \frac{1 - IL}{CR}) \cdot V_{bias}] \quad (4)$$

where R_s is the conversion efficiency from optical power to electrical current, P_I the input optical power, V_{bias} the bias voltage, and V_{dd} the supply voltage.

Again, modulator driver's power consumption is due to the charging/discharging of capacitances in the inverter chain. The maximum output voltage swing corresponds to the supply voltage, so total power dissipation can be approximated as:

$$P_{modulator\ driver} = \alpha_2 \cdot C_{md} \cdot V_{dd}^2 \cdot BR \quad (5)$$

where α_2 is the switching efficiency (probability of bit transitions) of the input data stream, BR is the bit rate of the link, and C_{md} is the total capacitance it drives (the sum of the capacitance of the driver and the modulator).

Dynamic power control. The modulator can be made power-aware by varying the light intensity delivered to each link by inserting a tunable attenuator after the splitter output for each link.

The electrical modulator driver can also be made power-aware through bit-rate scaling. While scaling the supply voltage with bit rate offers additional power savings, that would reduce the voltage swing to the modulator. This reduction degrades the contrast ratio [7] making it harder to detect the data at the receiver. Due to the already low supply voltage levels, we opt to keep the supply voltage for modulator driver fixed. So, the power consumed by the modulator driver only scales with bit rate.

2.2 Receiver design

At the receiver, as shown in Fig. 1, the *photodetector* accepts the light and converts the optical bit stream back into electrical current signals, which are then transformed into amplified voltage signals by the *transimpedance amplifier*

(*TIA*). Finally, the *clock and data recovery* (*CDR*) circuit tracks the amplified voltage signal and extracts the digital 1s and 0s. We explain the operation and real-time power-control of these components in detail in the following subsections.

2.2.1 Receiver: Photodetector

Operation. The photodetector converts the optical signals into photon-current. To guarantee an acceptable bit error rate (BER) (typical BER for inter-chassis and inter-board links is 10^{-12}), a minimum amount of optical power is required by the detector, defined as the receiver sensitivity (P_{rec}). Higher bit rates require higher receiver sensitivity to achieve the same BER.

Power characteristics. The detector dissipates power as it absorbs photons to generate current. The average power dissipation is shown in Eq. 6 [10]:

$$P_{detector} = P_{rec} \cdot \frac{q}{h\nu} \cdot V_{bias} \cdot \frac{CR + 1}{CR - 1} \quad (6)$$

where q is the charge of an electron, h is Planck constant, ν is the optical frequency, and CR the optical intensity contrast ratio for bit 1 and 0. V_{bias} is the bias voltage to the photodetector.

Dynamic power control. Given that the photodetector's power dissipation is much lower than other components ($<1\text{mW}$ [10]), no additional power control mechanisms are considered.

2.2.2 Receiver: Transimpedance amplifier (TIA)

Operation. The TIA typically consists of an internal common-source amplifier with a feedback impedance R_f as shown in Fig. 1. It transforms the photon current (I_p) from the detector to a voltage swing $I_p \cdot R_f$. It works well up to a maximum bit rate (BR_{max}), which is regulated by the bias current of the internal amplifier [1]:

$$I_{bias} = c \cdot BR_{max} \quad (7)$$

where c is a constant for a given TIA implementation.

Power characteristics. The power consumption for TIA depends on the bias current (I_{bias}), photon-current (I_p) and dark current (I_d) of the photo-detector [1]. However, as the power incurred by the photon-current (less than $100\mu A$) and dark current (several nA) are negligible [10], compared to the total TIA power consumption (around hundreds of mW) [1], the power consumed by TIA can be simplified as in Eq. 8:

$$P_{TIA} = I_{bias} \cdot V_{dd} = c \cdot BR_{max} \cdot V_{dd} \quad (8)$$

Dynamic power control. When the bit rate scales down, the maximum affordable bit rate BR_{max} can be reduced by the same degree. Thus, the bias current can scale with the bit-rate by tuning the supply voltage as the bias current has an almost linear relation to the supply current [19]. Then, the TIA's power scales with $V_{dd} \cdot BR$. Another benefit is that the TIA output voltage swing $I_p \cdot R_f$ can be smaller when supply voltage decreases. So with R_f fixed, less I_p is required for a lower supply voltage.

2.2.3 Receiver: Clock and data recovery circuitry (CDR)

Operation. The CDR is a key component of both optical and electrical receivers. It consists of a clock recovery circuit that re-times an internal clock with respect to the incoming data signals and decision circuitry that extracts digital data from the received signals. Once in lock, the CDR can recover a constant stream of data received at a fixed rate. However, in the event of a sudden change in the bit rate, the CDR requires time (set by the bandwidth of the timing recovery loop) to recapture lock before it can again operate reliably.

Power characteristics. A straight forward implementation of a CDR is to use a PLL (phase-locked loop) structure [12]. In a CDR, the PLL and clock buffers are the dominant power consumers, so power consumption does not change much with actual bit patterns. Assuming CDRs are mostly comprised of digital circuitry, the main power consumption comes from charging and discharging capacitors at high frequency. Its power consumption can be approximated as:

$$P_{CDR} = \alpha_3 \cdot C_{CDR} \cdot V_{dd}^2 \cdot BR \quad (9)$$

where α_3 is the switch efficient for CDR representing the probability of charging or discharging the capacitance, C_{CDR} is the capacitance of the CDR.

Dynamic power control. Like the VCSEL driver and the TIA, the CDR can similarly be frequency and voltage-scaled, as bit rate varies. Therefore, its power consumption has a scaling trend close to $V^2 \cdot f$ [12]. Whenever bit rate changes, we make the conservative assumption that the CDR needs to relock to the bit rate and re-synchronize the clock with the incoming data, so it is disabled for a time period called bit-rate transition delay T_{br} .

2.3 Design issues

Comparison of VCSELs vs. MQW modulators with an external laser. Due to its ease of integration with CMOS technologies, small footprint, and simpler connection design over MQW modulators, VCSELs are commonly used as the light source in box-to-box and board-to-board opto-electronic links. However, an external laser source coupled with MQW modulators has also been proposed as an alternative for such links [9] and offers some advantages. Here, we

compare the relative merits of using VCSELs versus MQW modulators for a power-aware network, first considering various attributes of each scheme with respect to performance and power. While this comparison is qualitative, Section 4 will evaluate both schemes using detailed network simulations to provide a more quantitative comparison.

One advantage of using MQW modulators with an external laser source is that the technology has been extensively used for telecommunications. Hence, we can leverage existing technology advancements, while VCSEL technology is relatively immature. For example, current modulator-based links have been demonstrated for 40 Gbps operation and beyond [20]. In comparison, the highest modulation speed for VCSELs is still at 10GHz [18]. Another significant benefit of using MQW modulators can be derived from the stable optical power available from the external laser source. This stability offers relatively lower noise operation. Comparatively, the VCSEL output is sensitive to various factors such as temperature and the operating voltage environment, thus, requiring additional circuit complexity to stabilize the system.

In addition to the performance benefits, modulator-based optical links has potential power advantages as well. Using an external laser source allows us to move the primary heat source in opto-electronic networked systems away (physically) from the actual system, into a separate chassis with its own power supply and cooling. This separation allows us to focus on the network system's power in order to ease thermal constraints and only contend with the power dissipation associated with the MQW modulators in the overall system power budget.

While one of the disadvantages of VCSELs stems from the threshold current that consumes constant power, recent progress in VCSEL technology, such as oxide-aperture-confined structure [18, 10], has significantly reduced the threshold current to hundreds of micro-amps. Thus, for the on-board power dissipation, VCSEL based transmitters dissipate power comparable to MQW modulator based transmitters [10]. Moreover, there is lower complexity to build and control a VCSEL based power-aware opto-electronic link. So, both schemes are explored in this paper and their relative efficiency in adjusting network power consumption with respect to traffic is simulated and explored in Section 4.

Interaction between power-control mechanisms and operation of link components. Implementing power control into the various link components of a power-aware network requires careful consideration of its interaction with the components. When bit rate scales down, also reducing the supply voltage to each of the components (e.g., VCSEL/modulator driver, TIA, and CDR) offers significant power savings, as shown by Eqs. 3, 5, 8 and 9. However, if the output swing of the modulator driver (set by the power supply voltage) drops, there will be a dramatic increase in insertion loss and a big decrease in contrast ratio. Those effects will adversely affect the photodetector's operation. Therefore, only bit rate (frequency) control can be used to reduce power consumption in the modulator driver.

In contrast, for a VCSEL-based transmitter, both bit rate and supply voltage can be controlled, as a decrease in modulation current (due to a lower supply voltage) only leads to a linear reduction in the optical output power, preserving a high contrast ratio. Moreover, both the photodetector and TIA are able to operate at lower light levels and supply voltages as bit rate decreases. Hence, it is possible to maintain acceptable BER performance by carefully balancing the impact of lower light intensity.

These simple examples show that the power-control mechanisms (i.e., frequency and voltage control) in each com-

ponent of an opto-electronic link have to work in concert, in order to allow link power consumption to vary with bit rates while ensuring correct operation. These nuances are factored into our overall networked system design and simulated to facilitate a quantitative exploration of the design space of power-aware opto-electronic networks.

3 Power-aware opto-electronic networked system

Advances in VLSI and network link bandwidths are making it cost-effective to connect multiple processing elements to each communication router, forming clustered systems. While flat interconnection network architectures are traditionally assumed, hierarchical, clustered architectures are gaining interest in large scalable parallel systems—examples include the IBM Blue Gene [5], Intel Paragon [8], Stanford DASH [15], and Cray T3D [3]. We thus target such systems with power-awareness using power-controlled opto-electronic links, as proposed in the previous section, for both inter-rack as well as inter-board interconnections.

3.1 System architecture

In this section, we present in detail an example system architecture using MQW-modulator-based links, owing to their higher complexity. A network system with VCSEL-based links is similar to the example system except that no external laser source is required. Moreover, light intensity is controlled directly by the VCSEL driver instead of requiring external optical power control.

Figs. 3 and 4 sketch the design of our proposed power-aware opto-electronic clustered network system comprised of 64 clusters, each with 8 processing nodes. A general two-dimensional mesh topology is chosen for the inter-cluster network, with eight processing nodes inside each cluster. This topology combines both the scalability of meshes and the cost-effectiveness of clustered designs (Fig. 3(a)). Each eight-node cluster, along with its communication router, is placed within the same rack. The eight boards within the rack each houses a processing node, and are attached to the communication router on another board through opto-electronic fiber links (Fig. 4(a)). Inter-rack communications occur through the router nodes, which are connected to neighboring routers in adjacent racks based on the mesh topology, also using opto-electronic fiber links. While a VCSEL-based scheme incorporates the lasers into the transmitter of each link, a MQW modulator based system requires an additional component illustrated in Fig. 3(b). An external laser source provides light to the transmitters for all of the links through fibers originating from the central laser source, where the optical power for each rack is controlled by an attenuator. The attenuator is managed through the control lines driven back from the routers, indicating the power level each should be set to.

The router microarchitecture is presented in Fig. 4(b). It consists of eight ports (ports 0-7), which are injection/ejection ports connected to the eight processing elements within that rack (cluster). The other four ports (ports 8-11) are for inter-router connections from north, south, east and west. A power-aware policy controller sits in the router node for every link, setting the power levels of each opto-electronic link in response to actual utilization. An additional power controller sends control messages to the external laser to set optical power levels. While it is possible to make the router itself power-aware, we chose to run it at a fixed frequency in order

maintain a consistent reference clock across links running at different bit rates. Thus, a router in this power-aware network continues to operate on fixed-size flits² in the midst of variable bit rates.

3.2 Adjusting bit rate and power levels

In power-aware systems, one must carefully choose the number of different power levels and the granularity of these levels. In our power-aware opto-electronic link, different power levels correspond to different bit rates. Power control is achieved via the variable bit rate and corresponding voltage variations for the link components and different optical power levels. Optimal power control depends on two key factors: (1) The requirements of a particular network traffic pattern, to ensure good overall network power-performance tradeoffs, and (2) The number and granularity that can be supported by the underlying optical and electrical circuits. Here, we discuss how we arrive at a suitable range for the number and granularity of power levels given circuits constraints. Section 4 evaluates the effectiveness of this range given a variety of network traces.

3.2.1 Electrical bit rate and voltage levels

Given the long transition time for a large step in both bit-rate and voltage variation, small steps are preferred in frequency variations. Moreover, in order to avoid the much longer delay overhead incurred by voltage transitions compared with frequency transitions, voltage will be pulled up before the frequency is increased, which will be orchestrated by the power-aware system's control policy. Conversely, voltage is reduced only after the frequency decreases. Since the voltage transitions always meets performance requirements, the link can function properly during slow voltage transitions, so we only assume that the links are disabled during the frequency transitions.

3.2.2 Optical power levels for transmitters

For a VCSEL-based system, since VCSEL's output optical power is directly modulated by its driving current I (see Eq. 1), which is almost proportional to the supply voltage of VCSEL driver, the optical power level is automatically tuned by supply voltage scaling.

For MQW-modulator based system, as we fixed the supply voltage to modulator driver to guarantee proper link operation, the optical power level is controlled by the attenuators of the external laser source. The long delay (around $100\mu s$) required to switch between levels motivates the use of fewer, coarser-grain optical power levels. Hence, we must balance the power savings that more levels avail with the corresponding performance degradation and increased system complexity. In this light, we chose to investigate two scenarios. First, use a fixed power level, which obviates the external laser source controller and reduces design complexity. Second, implement three optical power levels – P_{low} , P_{mid} and P_{high} (where $P_{low} = 0.5 \cdot P_{mid}$, $P_{mid} = 0.5 \cdot P_{high}$) correspond to three bit rate intervals – BR_{low} ($<4\text{Gb/s}$), BR_{medium} (4 to 6Gb/s), and BR_{high} (6 - 10Gb/s), where we assume 10Gb/s to be the maximum BR. In order to minimize the performance impact of changing optical power levels, we follow a control policy similar to that used for adjusting voltage levels, e.g. for low-to-high bit-rate transitions and corresponding optical power levels, the optical power is changed in advance and then the bit rate.

²Flits stand for flow control units, and are fixed-size segments of a packet

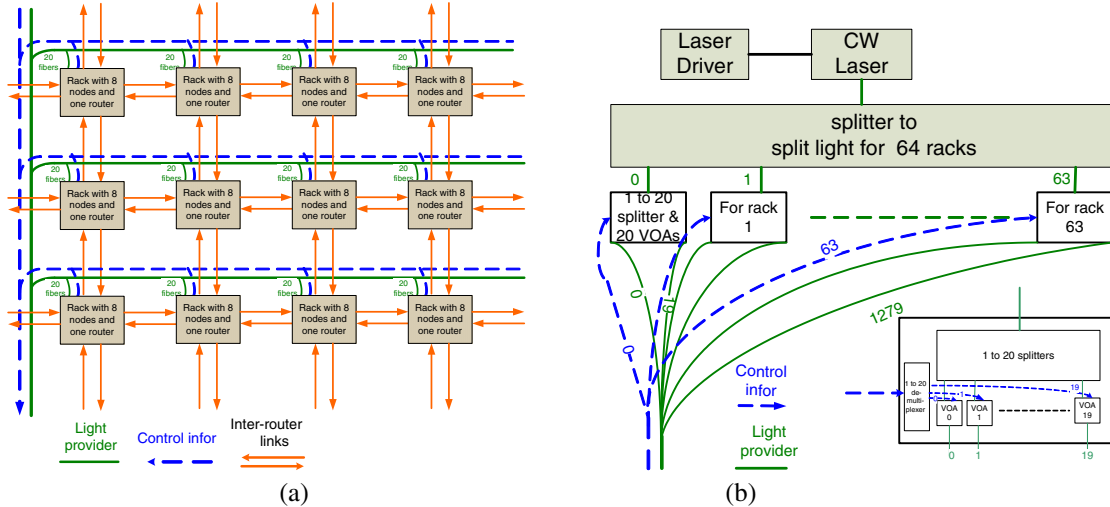


Figure 3. (a) Inter-cluster (rack) interconnection network, where the bold lines correspond to the fibers providing light from the external laser source, the gray ones correspond to the power-aware opto-electronic links used for inter-cluster communication, and the dashed lines are control lines back to the external rack housing the external laser that feed the power control units, (b) External laser source that is statically split to all 64 racks and the 20 fibers within each rack, through two levels of 1:64 followed by 1:20 splitting. A power control unit (a VOA or tunable optical switch) varies the optical power for each outgoing fiber, and is controlled through a control fiber from the racks.

3.3 Control policies

The policy controllers in the router decide *when* and *how* to transition between power levels. The link policy controller regulates the transitions between electrical bit rate and voltage levels, while the external laser source controller regulates transitions between optical levels. Here, we adopt policies previously proposed for electrical DVS links [24], extending it to target the multiple policy controllers required to manage the multiple time scales of our power-aware network.

Link policy controller. A policy controller sits at every link, shown in Fig. 4(b), predicting future workload based on historical traffic statistics. Basically, if the link utilization is greater than a threshold, T_H , it will change to the next higher bit rate level. Conversely, when link utilization is lower than T_L , it will move to the next lower bit rate level. If link utilization falls between the two thresholds, the level remains unchanged. Buffer utilization is used as an indication of network congestion, selecting different sets of T_H and T_L for when the network is congested and when it is relatively idle. The details of the policy are as follows.

Historical statistics are collected with hardware counters sitting at each router port—statistics on link utilization L_u (the percentage of router clock cycles where a flit traverses the output link) and buffer utilization B_u (the average percentage of buffers used in the next router’s port) over a time window T_w as shown in Eq. 10.

$$L_u = \frac{\sum_{t=0}^{T_w-1} A(t)}{T_w}, 0 \leq L_u \leq 1$$

$$B_u = \frac{\sum_{t=0}^{T_w-1} F(t)/B}{T_w}, 0 \leq B_u \leq 1$$
(10)

where $A(t)$ equals 1 if traffic passes in cycle t , otherwise 0 if no traffic passes in cycle t , F_t is the number of occupied buffers in time t , and B is the input buffer size. T_w , the time

window, is an important parameter, since decisions are made at the beginning of every T_w based on the statistics collected during the previous T_w . If T_w is too small, the policy controller will tune the bit rate frequently for short-term traffic fluctuations. This results in performance degradations since the link is disabled for a significant portion of the time in T_w , in the presence of bit-rate transitions. On the other hand, if T_w is too long, the policy controller cannot adequately adapt the bit-rate to accommodate large fluctuations in workload. We use network simulations (Section 4) to vary T_w (from 5 to 500 T_{br}) to find the optimal time window.

B_u (an indicator of network congestion) helps to select the thresholds for link utilization (T_H, T_L). When B_u is greater than the congestion threshold $B_{u,con} = 0.5$, the network is congested, so packets must wait a long time in the router buffer before it can be processed. This delay can mask the additional latency due to lower bit rate operation, so the control policy can afford to be more aggressive as shown in Table 1. These values have been set arbitrarily due to time constraints, and will be optimized in the future through profiling of the traffic traces during simulations.

Table 1. Thresholds for link utilization

Thresholds	$B_u < B_{u,con}$	$B_u \geq B_{u,con}$
Low, T_L	0.4	0.6
High, T_H	0.6	0.7

In order to make our system robust to short-term traffic fluctuations, our policy incorporates a mechanism to average the statistics across N time windows. The statistics of link utilization is stored in a sliding window, as shown here:

$$L_{u,a} = 1/N \cdot \sum_{i=0}^{i=N-1} L_u(i)$$
(11)

With these statistics, the link policy controller decides whether to increase/decrease or keep the bit-rate. At the beginning of

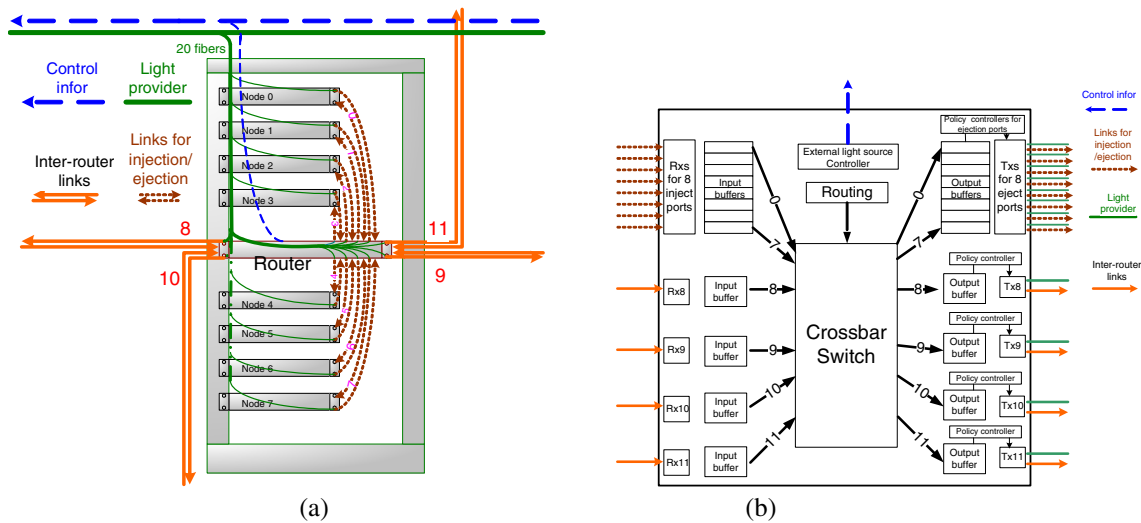


Figure 4. (a) Sketch of the internals of a rack, consisting of eight boards, each with a processing node, and the communication router board, (b) Router micro-architecture, with 8 injection and ejection ports to the intra-rack cluster, and 4 ports with two uni-directional links for inter-rack communications.

every T_w , $L_{u,a}$ is compared with two link utilization thresholds (T_H, T_L). When $L_{u,a}$ is greater than T_H , the policy controller increases bit rate by one level. Similarly, when $L_{u,a}$ is lower than T_L , the bit rate will decrease by a step. If $L_{u,a}$ falls between the two thresholds, bit rate does not change. For a modulator-based system with multiple optical power levels, the link policy controller cooperates with the external laser source controller, which is described next.

External laser source controller for modulator-based systems. To accommodate the long transition/response times of the attenuators at the external laser source, the external laser source controller seeks to track much longer trends in network traffic. Since the response time of the attenuators is around $100\mu s$, the link policy controller will decide whether to change the optical power levels every $200\mu s$. If in this $200\mu s$, the bit rate always remains at a power level that can function with a lower optical power, the policy controller will send a P_{dec} request, which triggers the external laser source controller to tell the external laser source to halve its optical power. On the other hand, if the link policy controller needs to increase link bit rate above that which can be supported by the current optical power level, it will instantly send a P_{inc} request to the external laser source controller (with the electrical bit-rate and voltage remaining constant until the optical power increases), which will prompt a doubling of the optical power. Otherwise, the input optical power remains fixed.

4 Evaluation results

To evaluate the performance impact and potential power savings of power-aware opto-electronic networks, we simulate the details of the entire power-aware network system as described in the previous two sections. Here, we present the simulation setup and experimental findings.

4.1 Network simulator & experimental setup

An event-driven flit-level interconnection network simulator [23] with 5-stage pipelined routers was modified

to include the detailed power-performance characteristics of a power-aware opto-electronic link, the external laser source with attenuators, the routers with the policy controllers, modeling a complete 64-rack power-aware networked system. In our experiments, the routers run at 625MHz, have 16 flits buffers per input port, where each flit is 16 bits wide. The maximum bandwidth out of each input port is set to be 10Gb/s. The simulator can be configured to model either VCSEL-based or MQW-modulator-based opto-electronic links.

Separate clock domains are used for the router core and its links. Functional modules inside the router core operate off of a fixed system clock, while each link has its own clock dynamically tuned by the link policy controller to follow the traffic changes. In the MQW-modulator-based links with multiple optical power levels, the external laser source policy controller in each router dynamically controls the attenuators in the external chassis.

The power consumption for the various components of an opto-electronic link is estimated using the power models described in Section 2, based on parameters from [1, 2, 13], assuming $0.18\mu m$ CMOS technology is used to implement all of the link circuitry. A rough breakdown of the power consumed for various link components operating at the maximum bit rate of 10Gb/s is listed in Table 2, along with approximate power-scaling trends for each component derived from the power models presented in Section 2. We assume that the required supply voltage to the VCSEL driver, TIA, and CDR will linearly scale with bit rate [12, 28], while the modulator driver's voltage is fixed to ensure correct operation.

As shown in Table 2, the transmitter of our uni-directional 10Gb/s opto-electronic link takes approximately 40mW, while the receiver dissipates approximately 250mW, a total of 290mW per link. With bit rate scaling from 10Gb/s to 5Gb/s with 6 bit-rate levels, the supply voltage scales from 1.8V to 0.9V accordingly, excluding the supply voltage to the modulator driver. This lowers link power consumption to 61.25mW at 5Gb/s for a VCSEL-based links, allowing a potential power savings of about 80%.

To be conservative, the link will be disabled for

Table 2. Power consumptions and scaling trends of the link components.

	VCSEL	VCSEL driver	Modulator driver	TIA	CDR
Power (mW)	30	10	40	100	150
Scaling trend \sim	V_{dd}	$V_{dd}^2 \cdot BR$	BR	$V_{dd} \cdot BR$	$V_{dd}^2 \cdot BR$

20 network cycles³ after the bit-rate transitions to give the CDR time to relock the clock to the input data. The transitions of supply voltage is slower, taking 100 network cycles. These delays are estimated and extrapolated based on characterizations of prior circuit designs of variable-frequency links [28, 12]. The impact of these transition delays will be investigated through network simulations.

Latency, throughput, power dissipation, and power-latency product are the metrics used to evaluate our power-aware policies. Latency refers to the time from the creation of the first flit of the packet till the ejection of its last flit from the network at the destination, throughput is defined as the injection rate at which average network latency exceeds twice the latency at zero network load, and the power dissipated by our power aware network is expressed as a percentage of that consumed by a non-power-aware network with all links at the maximum bit rate of 10Gb/s. Averages are computed across all nodes in the network. Power-latency product multiplies average latency with average power dissipation, encapsulating in a single metric the power-performance of a network.

4.2 Workloads

Network workloads that accurately reflect the high temporal and spatial traffic variance of many applications, with dynamic fluctuations, bursts and hot-spots, are most useful for evaluating the performance of our power-aware network design. In this paper, we present results for three sets of traces: (1) Uniform random traffic, where each node has equal probability of sending to any other node, at a constant injection rate; (2) A time-varying hot-spot traffic trace, where packets are injected at different injection rates at different phases of the simulation (temporal variance), and node 4 in rack(3,5) accepts four times the traffic injected into others (spatial variance); (3) Actual traffic traces extracted from SPLASH2 parallel applications [29] running on the RSIM multiprocessor simulator with default parameters [27].

Uniform random traffic is one of the most widely-used network loads due to its simplicity that lends readily to analysis. For power-aware networks, its constant injection rate poses a worst-case scenario to the policy controller, since the lack of variance provides little opportunity for frequency/voltage scaling. We thus use it to stress our power-aware policies. The time-varying hotspot trace is used to evaluate the responsiveness of our power-aware network design to fairly marked fluctuations in network traffic, essentially to stress our power-aware circuit mechanisms, highlighting the impact of link circuitry overheads. The SPLASH traffic traces are used to evaluate the realistic power-performance impact of a power-aware network.

4.3 Simulation results and performance evaluation

4.3.1 Uniform random traffic

Effect of policy parameters. We first explore the impact of our policy's sampling window size (T_w), i.e. how frequently it samples statistics, on a modulator-based network (trends are

similar for VCSEL-based networks), varying it from 100 to 10000 cycles.

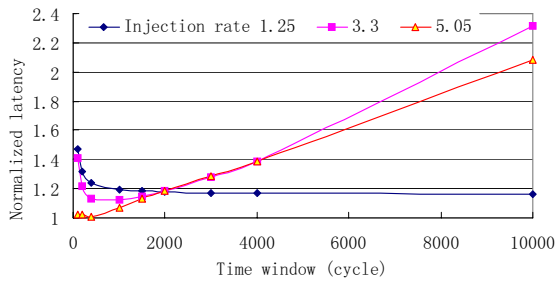
Fig. 5(a) presents the average network latency (averaged across all links in the network) relative to that of a non power aware network under light (1.25 packets/cycle), medium (3.3 packets/cycle) and heavy (5 packets/cycle) injection rates. It shows that a larger latency penalty is incurred for the shortest windows ($T_w = 100$). This is because a control policy that uses shorter windows yields more power-level transitions, which in turn incurs more transition penalties. Larger latency penalty is also incurred for a bigger window size under medium and heavy traffic since it cannot respond quickly to traffic variations. The effect on power dissipation (see Fig. 5(b)) shows shorter time windows leading to higher power consumption for all traffic injection rates except under 1.25 packets/cycle. At such light traffic, the power-aware network essentially transitions to the lowest bit-rate supported (5Gb/s) and remains at that level. Fig. 5(c) presents the effect of window size on power-latency product. It shows that window size around 1000 cycles yields acceptable power latency product for all injection rates. While one might expect a shorter time window to enable the policy to be more responsive, better track traffic fluctuations, and thus reduce power, more frequent bit-rate transitions also lead to higher power consumptions. Since the network must disable the links more often, it must compensate for this loss in link activity with higher bit rates when links are active, resulting in higher power consumption. Based on these results, we choose $T_w = 1000$ cycles for all subsequent simulations.

Figs. 5(d),(e),(f) present the effect of link utilization thresholds (T_H and T_L) on average latency, power consumption and power-latency product, respectively. The difference between high and low utilization thresholds ($T_H - T_L$) is fixed at 0.1 as simulations show better power-performance. Intuitively, higher thresholds lead to more aggressive scaling of link bit rates, which result in larger delays as well as lower power consumption, as is evident for the injection rate of 3.3 packets/cycle. At light traffic (1.25 packets/cycle), the power-aware network relegates to few transitions – mimicking a network whose link bit rates are statically set at startup. At high traffic (5.05 packets/cycle), network latency does not increase either with more aggressive thresholds, as the network is highly congested, so flits are queued up in the routers for long periods of time anyway, masking the additional link delay due to power-aware networks. We choose an average threshold of 0.5 for subsequent simulations in order to balance the impact on power-performance. If a larger average threshold of 0.6 is chosen, higher power savings can be attained, demonstrating the tradeoff designers have to make.

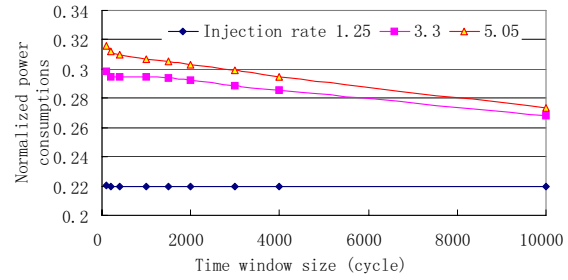
Effect of link technology. Fig 5(g) compares the average latency of two power-aware network configurations – one whose link bit rates vary from 5 to 10Gb/s, vs. one whose link bit rates vary from 3.3 to 10Gb/s. Our simulations show the network with 5-10Gb/s links not hurting network throughput, saturating at the same point as the non-power-aware network. When 3.3-10Gb/s links are used, however, throughput suffers, degrading to about 3 packets/cycle. If link bit rates are statically set at 3.3Gb/s, throughput will be severely affected, dropping to lower than 2 packets/cycle.

Fig. 5(h) presents the average power consumption of power-aware network systems. Power savings can be at-

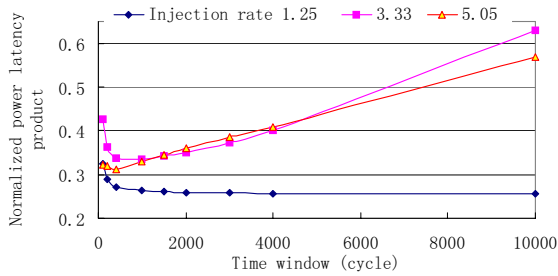
³A network cycle is basically one flit time.



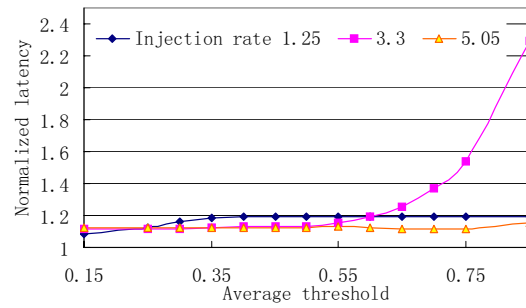
(a) Latency over window size



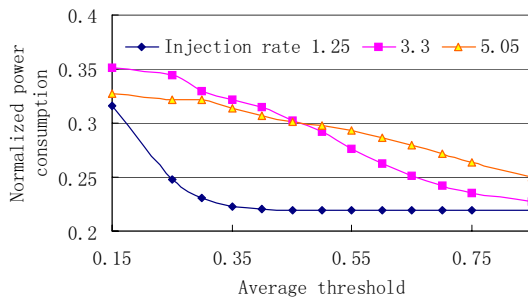
(b) Power over window size



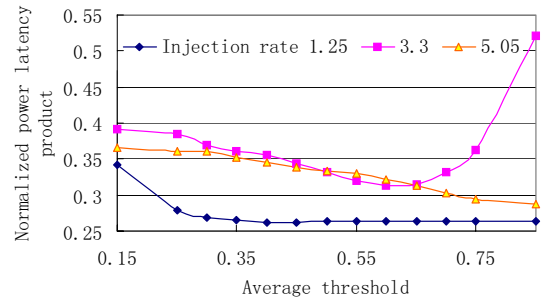
(c) Power latency product over window size



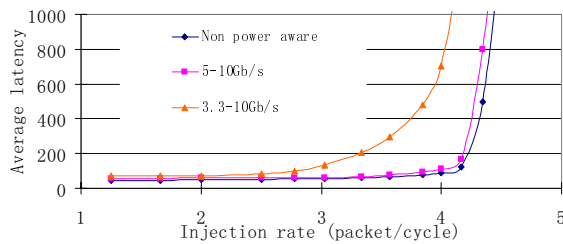
(d) Latency over threshold



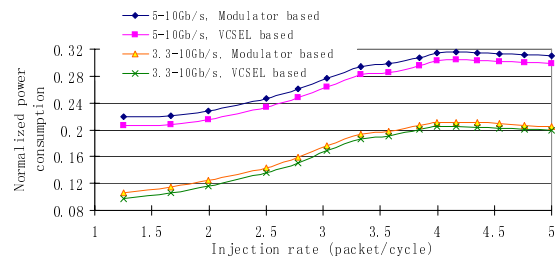
(e) Power over threshold



(f) Power latency product over threshold



(g) Latency over injection rate



(h) Power over injection rate

Figure 5. Under random uniformly distributed traffic: (a) Average latency for different injection rates normalized to non power-aware network over time window size. (b) Power consumption of a power-aware network relative to a non-power-aware network over time window size, (c) Power latency product of a power-aware network normalized to a non-power network over time window size, (d) Normalized average latency over average link utilization threshold, (e) Normalized power over average link utilization threshold, (f) Normalized power latency product over average link utilization threshold, (g) Average latency over injection rate for power aware and non power-aware network systems, (h) Power consumption of a power-aware network relative to a non-power-aware network.

tained even when the inter-router links are completely saturated, as the injection/ejection links into the cluster router are lowly utilized for uniform random traffic. Before network saturation, power dissipation increases as more traffic is injected. Beyond that, the power-aware policy controller can be more aggressive in adjusting bit rates to enable more power savings even as traffic increases. It shows that power aware opto-electronic networks are more effective at light and heavy ends of traffic conditions where sensitivity of overall network latency to link delay is lower.

It is again important to point out that more power savings is also achievable at the expense of increased latency penalty by using a lower minimum bit-rate as shown in Fig 5(g),(h). These plots clearly demonstrate the tradeoff between latency and power savings for power-aware networks. If power consumption is our main concern, greater than 90% savings in power consumption can be achieved, with VCSEL-based networks having a slight advantage. We select 5-10Gb/s range henceforth since that gives better power-performance.

4.3.2 Time-varying hot-spot traffic

Fig. 6 presents the network simulation results using the time-varying hotspot traffic trace shown in Fig. 6(a).

Effect of link circuitry constraints. The number and range of bit-rate (and corresponding voltage) levels and transition delays between levels are key link circuitry constraints that impact overall network performance. Having an understanding of their effect on overall network performance allows a link designer to better tune its link specifications.

To explore the effect of transition delay overheads, we zero out T_v and/or T_{br} . As expected, Fig. 6(b) and (c) shows that a system without power awareness has the lowest latency, since all of the links operate at the highest bit rate, and thus consume the most power. Given the control policy of changing the supply voltage in advance of increasing the bit rate, links are able to always operate in the presence of voltage transitions. Hence, the voltage transition penalty has negligible impact on performance. These results also show that for a timing window size (T_w) of 1000 cycles, the relatively small penalty of disabling the link for 20 cycles during bit-transitions (\bar{T}_{br}) has little impact on performance.

VCSEL vs. MQW modulator-based power-aware links. Fig. 6(c) provides insights into the impact of multiple optical power levels for a network with MQW modulators. For small increases in the network traffic (from 1.1e6 to 1.3e6 cycles), which corresponds to a small increase in bit rate, the optical power level does not change and there is no additional latency penalty. However, for a larger jump (between 1e6 and 1.1e6 cycles), a change in the optical power level is triggered, incurring higher latency as the network has to wait $100\mu s$ for an optical power transition. These simulation results verify that choosing a smaller number of coarse-grain optical power levels yields lower latency penalties. While this additional delay penalty is infrequent, subsequent simulations assume a single optical level to simplify comparisons between systems using VCSELs and MQW modulators.

Given these promising results, Fig. 6(d) then compares the normalized power consumption for systems using VCSELs versus MQW modulators, relative to non-power-aware networks. The slight advantage of a VCSEL-based system stems from the fact that its laser driver's power can scale with both bit rate and voltage, while the modulator driver operates off of a fixed supply voltage and only scales with bit rate.

4.3.3 Traffic traces of SPLASH2 parallel benchmarks on RSIM

Three SPLASH2 benchmarks are used to evaluate our power-aware network – fast fourier transform (FFT), matrix decomposition (LU), integer sort kernel (Radix). The benchmarks are parallelized onto 64 nodes housed in 8 racks of our proposed networked system. Since traces are long (several hundreds of millions of cycles), we snapshot just a portion where large fluctuations in injection rate are seen (Fig. 7). Figs. 7(a)(c)(e) show the injection rate over time (the average packet size is 48-flit) while Figs. 7(b)(d)(f) present the power consumption of the modulator-based power aware system under those 3 traces normalized to the power of non power aware system. We see that our power-aware network tracks workload fluctuations effectively. In addition, since link bit-rate is only changed for big variations in link utilization, the power curves filter out small fluctuations in the injection rate curves and are thus smoother. On average, more than 75% savings in power consumption is achieved with less than double network latency, resulting in more than 60% savings in network power latency product as shown in Table 3. The latency impact for FFT is much lower than the others as its traffic peaks and troughs occur over a longer period of time, making it easier for the policy to accurately predict trends. The increase in this raw network transmission latency needs to be taken in the context of overall network delay, which includes hundreds and thousands of cycles of network interface and software latency.

Table 3. Power-performance numbers for power aware networks normalized against non-power-aware networks.

Traces	FFT	LU	Radix
Average latency	1.08	1.50	1.60
Average power consumption	0.22	0.25	0.23
Average power latency product	0.24	0.38	0.37

5 Conclusions

In this paper, we explored the design space of power-aware opto-electronic networks. To the best of our knowledge, this is the first complete power-aware network design that explores link circuits, router microarchitecture, and overall network system architecture in tandem, as well as the first proposal of a detailed power-aware link architecture. Joint investigation across these areas allow link constraints to be accurately factored into network design, and system requirements to be incorporated into the underlying circuit design process.

Through detailed network simulations with both synthetic and actual traces, we show that power-aware opto-electronic networks can lower power consumption by approximately $4\times$, with VCSEL-based power-aware opto-electronic links consistently turning in slightly better power-performance across all traces. Moreover, this simulation environment has enabled us to understand the impact of various control parameters and design choices on power and performance, and find optimal settings.

We are currently working towards the design of opto-electronic link components implemented in a $0.18\mu m$ CMOS process with power-control capabilities that are compatible to the proposed power-aware network. Experimental results extracted from the test-chip prototype can then be fed into our network system simulator, in place of current models, to more accurately evaluate system performance and power

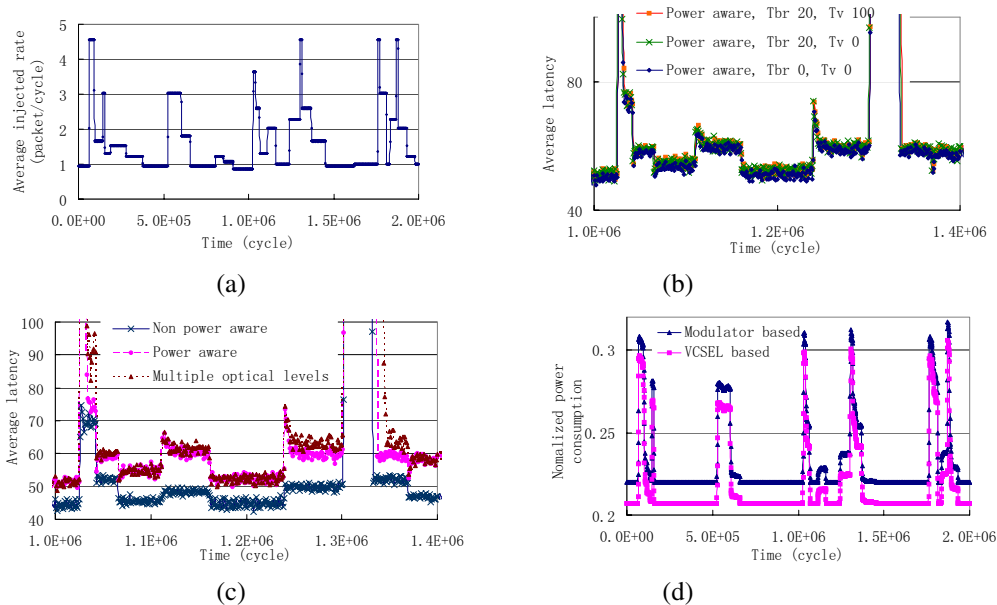


Figure 6. Under time-variance hot-spot traffic: (a) The average number of packets injected into the whole network over time (injection rate), (b) The average network latency for power aware systems w/o transition delays, (c) The average network latency for power aware systems with single or multiple optical power levels compared with non power aware systems, (d) Power dissipation of both VCSEL based and modulator based power-aware network system relative to that of non-power-aware networks.

savings. We hope this work will pave the way for the deployment of power-aware network systems in the future.

Acknowledgments

The authors would like to thank Jose Duato of the Technical University of Valencia for passing us a traffic trace of LU that significantly enhanced our insights in the submitted version [25]. We will also like to thank Amit Kumar of our research group at Princeton for characterizing and collecting all the RSIM multiprocessor traffic traces used in this camera-ready version. In addition, we appreciate the useful discussions with Li Shang, Lei Xu, Camille Bres and Darren Rand of Princeton University. This work is partially supported by NSF ITR grant CCR-0324891, NSF CAREER grant CCR-0237540, as well as the MARCO Gigascale Systems Research Center.

References

- [1] A. Apsel and A. G. Andreou. Analysis of short distance optoelectronic link architectures. In *International symposium on circuits and systems*, volume 4, pages IV-840-IV843, May. 2003.
- [2] D. Blerkom, C. Fan, M. Blume, and S. Esener. Transimpedance receiver design optimization for smart pixel arrays. In *IEEE Journal of Lightwave Technology*, volume 16, Jan. 1998.
- [3] Cray research Inc. Cray T3D system architecture overview, 1993.
- [4] W. J. Dally, P. Carvey, and L. Dennison. The Avici terabit switch/router. In *Proc. Hot Interconnects 6*, Aug. 1998.
- [5] IBM. IBM Blue Gene project. <http://www.research.ibm.com/bluegene/>.
- [6] IBM. IBM InfiniBand 8-port 12x switch. <http://www-3.ibm.com/chips/products/infiniband/>.
- [7] T. Ido, H. Sano, S. Tanaka, D. J. Moss, and H. Inoue. Performance of strained InGaAs/InAlAs multiple-quantum-well electroabsorption modulators. *Journal of Lightwave Technology*, 14(10):2324-2331, Oct. 1996.
- [8] Intel Corporation. Paragon XP/S product overview, 1991.
- [9] I. Keslassy, S.-T. Chuang, K. Yu, D. Miller, M. Horowitz, O. Solgaard, and N. McKeown. Scaling internet routers using optics. In *Proc. SIGCOMM*, Aug. 2003.
- [10] O. Kibar, D. A. V. Blerkom, C. Fan, and S. C. Esener. Power minimization and technology comparisons for digital free-space optoelectronic interconnections. *Journal of Lightwave Technology*, 17(4):546-555, April 1999.
- [11] 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 *Proc. Int. Symp. Low Power Electronics and Design*, pages 459-464, Aug. 2003.
- [12] J. Kim and M. Horowitz. Adaptive supply serial links with sub-1V operation and per-pin clock recovery. In *Proc. Int. Solid-State Circuits Conf.*, pages 216-221, Feb. 2002.
- [13] B. Lee, M. Hwang, S. Lee, and D. Jeong. A 2.5-10gb/s CMOS transceiver with alternating edge sampling phase detection for loop characteristic stabilization. In *IEEE Journal of solid-state circuits*, volume 38, Nov. 2003.
- [14] W. Leland, M. Taqqu, W. Willinger, and D. Wilson. On the self-similar nature of ethernet traffic (extended version). *IEEE/ACM Transactions on Networking*, 2(1):1-15, Feb. 1994.
- [15] D. Lenoski, J. Laudon, K. Gharachorloo, and et al. The Stanford DASH multiprocessor. In *IEEE Computer*, pages 63-79, 1990.
- [16] D. A. B. Miller. Rationale and challenges for optical interconnects to electronic chips. *Proceedings of the IEEE*, 88(6):728-748, June 2000.
- [17] Newport Corporation. Fiber Optic Couplers and Wavelength Division Multiplexers. <http://www.newport.com>.

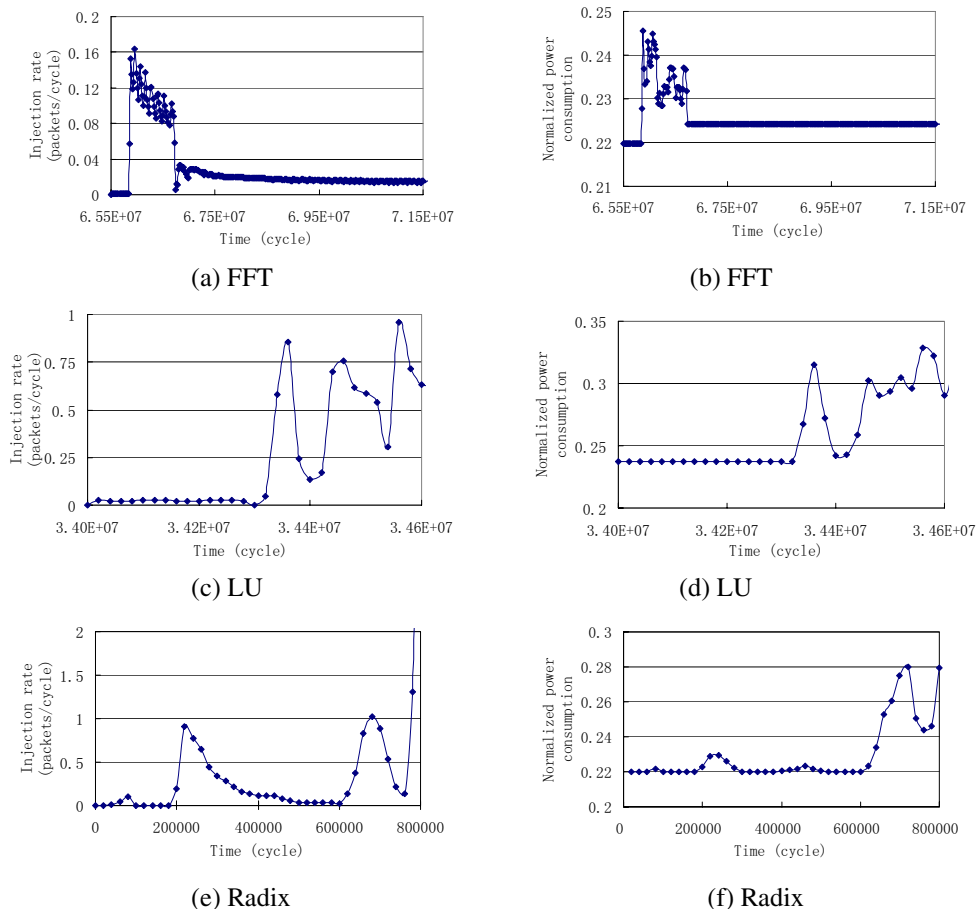


Figure 7. Under SPLASH2 benchmark traces: (a)(c)(e) injection rate for FFT, LU, and Radix respectively. (b)(d)(f) Normalized power consumption of power-aware networks relative to a non-power-aware network.

[18] M. Ortsiefer, R. Shau, F. Mederer, R. Michalzik, J. Roskopf, G. Bohm, F. Kohler, C. Lauer, M. Maute, and M.-C. Amann. High-speed modulation up to 10 Gbit/s with 1.55 μ m wavelength ingaalaal vesels. *Electronics Letters*, 38(20), Sep. 2002.

[19] Philips Semiconductors. TZA3013A;TZA3013B SDH/SONET STM16/OC48 transimpedance amplifier. <http://www.semiconductors.philips.com>.

[20] PriTel Inc. Actively Mode-locked Fiber Lasers. <http://www.pritel.biz>.

[21] P. R. Prucnal, S. D. Elby, and K. B. Nichols. Optical transmitter for fiber optic interconnects. In *Optical Engineering*, volume 30(5), pages 511–516, 1991.

[22] R. C. Johnson, EE Times. Darpa to Fund Optical Interconnect Research. <http://www.eetimes.com/story/OEG20030910S0040>.

[23] L. Shang. Popnet:a network simulator. <http://www.princeton.edu/~lshang/popnet.html>.

[24] L. Shang, L.-S. Peh, and N. K. Jha. Dynamic voltage scaling with links for power optimization of interconnection networks. In *Proc. Int. Symp. High Performance Computer Architecture*, pages 79–90, Feb. 2003.

[25] F. Silla, M. P. Malumbres, J. Duato, D. Dai, and D. K. Panda. Impact of adaptivity on the behavior of networks of workstations under bursty traffic. In *Proceedings of the International Symposium on circuits and systems*, volume 10-14, pages 80–87, Aug. 1998.

[26] V. Soteriou and L.-S. Peh. Design-space exploration of power-aware on/off interconnection networks. In *Proc. International Conference on Computer Design*, Oct. 2004.

[27] V. S. Pai, P. Ranganathan and S. V. Adve. RSIM Reference Manual. www-ece.rice.edu/~parthas/publications/rsim_manual.ps.

[28] G. Wei, J. Kim, D. Liu, S. Sidiropoulos, and M. Horowitz. A variable-frequency parallel I/O interface with adaptive power-supply regulation. *Journal of Solid-State Circuits*, 35(11):1600–1610, Nov. 2000.

[29] S. Woo, M. Ohara, E. Torrie, J. P. Singh, and A. Gupta. The SPLASH-2 programs: Characterization and methodological considerations. 1995.

[30] F. Worm, P. Jenne, P. Thiran, and G. D. Micheli. An adaptive low power transmission scheme for on-chip networks. In *Proc. Int. System Synthesis Symposium*, Oct. 2002.