Energy-Efficient Error Control for Tightly Coupled Systems Using Silicon Photonic Interconnects

Xuezhe Zheng, Pranay Koka, Michael O. McCracken, Herb Schwetman, James G. Mitchell, Jin Yao, Ron Ho, Kannan Raj, and Ashok V. Krishnamoorthy

***Abstract*—Future computer systems will require new levels of computing power and hence new levels of core and chip densities. Because of constraints on power and area, optical interconnection networks will play a critical role in these new systems. In this paper, we describe the macrochip, a multi-chip node with an embedded silicon photonic interconnection network that consists of thousands of optical links. For such a large-scale wavelength division multiplexing optical network, we show how to use an energy-efficient error control scheme employing variable-length cyclic redundancy check codes to achieve a desirable residual bit error rate (BER) of 10^{-23} for reliable system operation with the individual link BER at 10^{-12} or higher. We use a discrete-event network simulation of the macrochip using uniform random traffic to show that our scheme incurs minimal impact on performance compared to a perfect system with no error control. Using link level energy efficiency and network throughput analysis, we estimate and report network level energy efficiency using the metric of energy per useful bit.**

***Index Terms*—Error control; Error control protocol; Macrochip; Silicon photonic interconnects; WDM point-to-point network.**

I. Introduction

Ever-increasing demand for computer system performance is driving a trend toward dense, powerful compute blocks, integrating tens to hundreds of processing cores on a single die [1–3]. Applications driving this demand require fast access to large data sets stored in main memory. Looking forward, we envisage that chip power will be dominated by the network transporting data between cores and memory [4].

Current electronic I/O technologies offer substantial bandwidth between chips and memories but will be hard-pressed to scale up to the levels needed for such hundred-core chips. Because the areal density of off-chip I/O and package routes dramatically lags that of on-chip wires [5], such links are by necessity over-clocked and serialized and therefore relatively energy inefficient. Newer approaches using coupled data communication [6,7] bypass soldered I/O and package routing and instead transmit signals between chips through direct capacitive or inductive coupling, thus carrying all data over efficient and parallel on-chip wire buses [8]. While such systems enable modest arrays of chips, their ultimate scalability is limited by the low speed of on-chip wires, especially over distances longer than 10 mm. Therefore, building highly parallel systems with many hundred-core chips connected through electronic I/O technologies, either traditional or coupled links, limits off-chip bandwidth and will result in poor overall performance.

Optical interconnects, by comparison, promise higher inter-chip bandwidth with channels with lower energy per bit and can potentially better support such large-scale systems. Critical to the promise of such an optical system is the use of wavelength multiplexing as a way of improving interconnect density.

Integrating optical networking technology into computing systems can take several different paths. One such direction would widely separate discrete processor/memory chips and interconnect them using fibers, thus using optical links to create physically large but logically dense systems. This would offer simpler packaging and lower power and heat requirements yet leverage the increased bandwidth from wavelength multiplexing. However, chips generally connect to fibers at a relatively large 250 μm core pitch, not the 20 μm pitch of optical proximity couplers, so chip-to-chip bandwidth over fibers would not offer much density advantage over areal solder balls connected to package routes. To truly exploit the bandwidth advantages of silicon photonics, a high-performance system should instead employ dense silicon waveguides with fine-pitch connectors and tightly packed processors.

One such tightly coupled system design is the Oracle macrochip [9], which is a technology platform that integrates multiple processor die with a silicon photonic interconnection network. The network is embedded in a silicon-on-insulator (SOI) substrate, and the processor die are connected to the network using optical proximity communication; together, these make inter-die and intra-die communication bandwidths nearly equal. This approach provides a single-package compute block much larger than a single processor, but requiring neither large, low-yield chips nor area- and power-hungry soldered I/O pins.

The area efficiency offered by wavelength division multiplexed silicon photonics allows a system network that

Manuscript received January 19, 2011; revised May 30, 2011; accepted June 6, 2011; published June 24, 2011 (Doc. ID 141351).

Xuezhe Zheng (e-mail: xuezhe.zheng@oracle.com), Jin Yao, Kannan Raj, and Ashok V. Krishnamoorthy are with Oracle Labs, San Diego, California, USA.

Pranay Koka, Michael O. McCracken, and Herb Schwetman are with Oracle Labs, Austin, Texas, USA.

James G. Mitchell and Ron Ho are with Oracle Labs, Redwood Shores, California, USA.

Digital Object Identifier 10.1364/JOCN.3.000A21

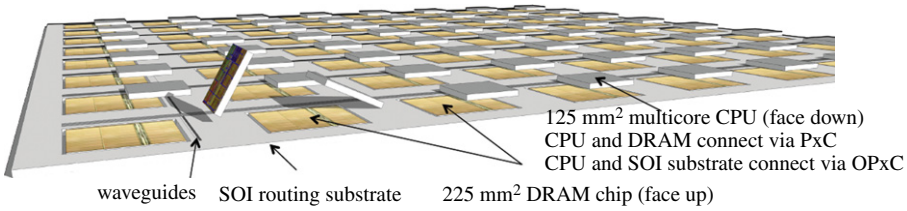

Fig. 1. (Color online) Diagram of an 8 × 8 macrochip.

can be as rich as a fully connected point-to-point topology, which minimizes communication latency for uniform traffic patterns [10]. Such an architecture would require on the order of 10,000 optical links, giving a total aggregated bandwidth of tens of terabytes per second. While this high bandwidth enables large quantities of data to be moved within a short amount of time, the number of errors occurring within the same time for the whole interconnection network increases significantly with the amount of data being moved assuming non-zero individual link bit error rates (BERs).

Ideally, a computer system will experience no undetected errors while it is in operation [11]; studying practice, a goal of fewer than one undetected bit error in the system over a ten-year period is a reasonable target. For our envisaged macrochip, this appears to require an effective link BER of less than 10^{-23}. Unfortunately, to pursue optical links with optimized energy efficiency, even achieving a BER of 10^{-12} at a high data rate of 20 Gb/s is a challenging goal, given device parasitics, complementary metal-oxide semiconductor (CMOS) circuit and system noise sources, and the low wall-plug efficiency (WPE) of wavelength division multiplexing (WDM) laser sources.

A more realistic approach is to use error control schemes to improve the link fidelity to a residual BER of 10^{-23}. The critical question is: what is the appropriate error control scheme for a macrochip interconnection network that can be both robust and power efficient?

Error control schemes have been widely applied in various optical networks especially in telecom metro and long-haul networks. In these networks, link BERs are relatively high, and, because of the distances involved, retransmitting faulty data is expensive. As a result, the focus has been on forward error correction (FEC). By contrast, for systems such as the macrochip with significantly different operating characteristics, the energy and performance costs of adding FEC to every packet become significant (as will be seen in Section IV below). Thus, other techniques must be considered. As will be seen below, using error detection at the packet level, coupled with an ACK/NACK protocol and retransmission of faulty packets, leads to a robust error control scheme that meets the goals for undetected errors and also meets the requirements for performance and energy use. We show below that these standard techniques can be adapted, in a straightforward manner, for use on the macrochip.

We begin by describing the macrochip, a tightly coupled multi-chip system based on a silicon photonic interconnection network, in Section II. We describe an energy-efficient photonic link in Section III, and in Sections IV and V we discuss the requirements of an energy-efficient error control scheme and the features of an optimized error detection scheme for the macrochip. Also in Section V, we present simulation results that show the impact of the error control protocol on performance. The contributions of this paper are summarized as follows:

1. A suitability analysis of error correction versus error detection schemes for energy-efficient interconnects.

2. Evaluation of error detection schemes for macrochip-like systems.

3. An approximate performance analysis showing the impacts of the error detection scheme.

The main result of this paper is the demonstration that existing error detection and correction techniques can be adapted to give the required levels of accuracy, energy efficiency, and performance for the macrochip.

II. The Macrochip System

The macrochip technology enables integration of multiple conventional die, each about 225 mm^2 in size, using silicon photonics to achieve performance similar to that of a large single die. In the current design, the macrochip can host 64 conventional die in an 8 × 8 array as shown in Fig. 1. This design bypasses die size limits imposed by technology yields and makes possible dramatically more cores on a virtual "chip." Multiple conventional die are integrated through a large SOI substrate with place holders to support the individual die. These place holders are called sites. The substrate contains two layers of silicon optical waveguides; the layers run in orthogonal directions much like on-chip electrical wiring, with via-like connections between the layers built using low-loss optical proximity connectors (OPxCs). By using two optical routing layers, orthogonal waveguides avoid physical intersections and the ensuing signal crosstalk. The substrate layers are SOI because the silicon waveguides require a buried oxide for light confinement [9], although photonics-enabled bulk silicon may in the future eliminate the need for SOI [12].

The sites on the macrochip can be a combination of logic and memory. Apart from compute logic and memory, each site includes optical transmitters, receivers, and waveguides positioned to overlap the SOI routing substrate, and uses OPxCs to connect its waveguides to those in the SOI routing substrate [13]. Power is delivered to each site from a top plate and the die are connected to the substrate using solderless spring contacts that allow chip replacement for higher system yield [8].

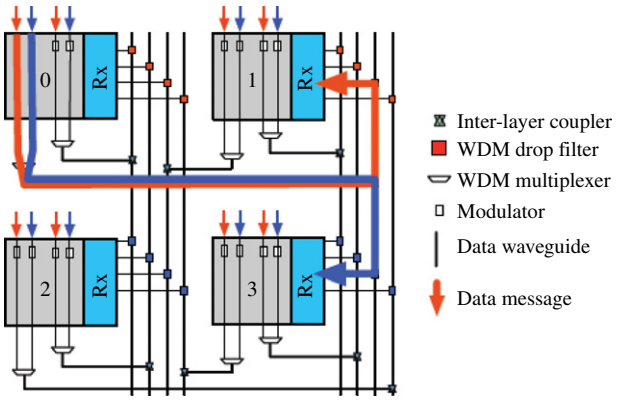

Fig. 2. (Color online) 2 × 2 static WDM point-to-point network.

The macrochip can be viewed as a large-scale shared memory multiprocessor or a message passing cluster, whose performance is not restricted by the limited pin-counts on processor die because all cores are interconnected through dense silicon photonics. Due to the applicability of the macrochip to different programing models, the macrochip needs to support different packet sizes, ranging from a few bytes to multiple kilobytes. In this paper we use three different packet sizes to represent shared memory and message passing applications. We used 8-byte and 64-byte packets to represent traffic in a shared memory system and 4096-byte packets to represent the commonly used MTU (maximum transferable unit) in message passing clusters.

The network used in the macrochip is a statically routed WDM point-to-point network, shown in Fig. 2. In this network each site has a dedicated optical data path to every other site. Such a network does not impose any connection setup or arbitration overheads. The network layout consists of horizontal and vertical waveguides between the rows and columns of the macrochip, respectively. The horizontal waveguides are laid on the bottom layer and the vertical waveguides are laid on the top layer of the SOI routing substrate. The horizontal waveguides connect to vertical waveguides using inter-layer couplers. Each vertical waveguide drops one wavelength at each site in the column. A transmitting site can communicate with any receiving site A by choosing the waveguides leading to the column of site A and the wavelength that is then dropped at site A.

Each site on the macrochip supports 320 GB/s input bandwidth and 320 GB/s output bandwidth. The system consists of a total of 8192 optical links, each running at 20 Gb/s. This yields an aggregate peak bandwidth for the macrochip of more than 20 TB/s. The parameters of the macrochip system, discussed in this paper, are summarized in Table I.

TABLE I
MACROCHIP SYSTEM PARAMETERS

Parameter	Value
Number of sites	64
Bandwidth per site	320 GB/s
Total number of optical channels	8192
Wavelengths per waveguide	8
Bit rate of a wavelength	20 Gb/s
Total peak bandwidth	20 TB/s
Latency	0.1 ns/cm

III. ENERGY-EFFICIENT INTERCONNECTS

Interconnects with very high energy efficiency are expected for the macrochip. A useful system interconnect metric must take into account not only the energy efficiency of the link technology, but also its capacity and its link utilization. We believe that one such metric is energy per useful bit, which is defined as [14]

$$\text{Energy/Useful Bit} = \frac{\text{Power}}{\text{Payload/Delivery Time}}. \quad (1)$$

Clearly, efficient interconnection systems would require not only energy-efficient photonic links, but also an energy-efficient networking protocol.

A. Energy-Efficient Photonic Links

Very-low-power photonic links are expected for macrochip-like applications. As depicted in Fig. 3, a typical photonic link consists of a laser source, a transmitter, a receiver, and passive WDM channel including a WDM mux/demux (multiplexer/demultiplexer), optical proximity couplers, and waveguides. The total dissipated power of an optical link for a given link bit rate can be measured as follows:

$$\text{Power}_{\text{Link}} = P_{\text{receiver}} + P_{\text{transmitter}} + P_{\text{WDM}} + P_{\text{Laser}}. \quad (2)$$

Here, the total power represents the effects of the receiver, transmitter, WDM mux/demux, and the source laser. Assuming that the photonic link is thermal noise limited, the source laser power is further determined by

$$P_{\text{Laser}} = \frac{2P_{\text{sens}}}{\text{IL}_{\text{Link}}} \cdot \frac{(E_r + 1)}{(E_r - 1)} \cdot \frac{Q}{Q_0} \cdot \frac{1}{\eta_{\text{Laser}}}, \quad (3)$$

where P_{sens} is the sensitivity of the receiver, IL_{Link} is the total link loss, E_r is the extinction ratio of the transmitter modulation, Q is the required link quality factor derived from the required undetected BER without error control, Q_0 is the sensitivity quality factor (e.g., a Q_0 of 7 for a BER of 10^{-12}), and η_{Laser} is the (wall-plug) laser efficiency defined as the ratio of optical power to d.c. electrical power.

To achieve the desired energy-efficient photonic links, we have to not only develop low-power transmitters and receivers using high-bandwidth, low-parasitic opto-electronic devices, but also reduce the laser power. Because η_{Laser} is typically low, less than 10% for currently available WDM sources, laser power becomes a significant portion of the total power dissipation of optimized photonic links. In addition, optimizing for energy efficiency, receiver design has to trade off power versus sensitivity, while transmitter design has to trade off power versus extinction ratio. With transmitter and receiver design, and link loss fixed, the laser power depends solely on the required channel quality factor. If the required link BER is significantly lower than the sensitivity BER for reliable

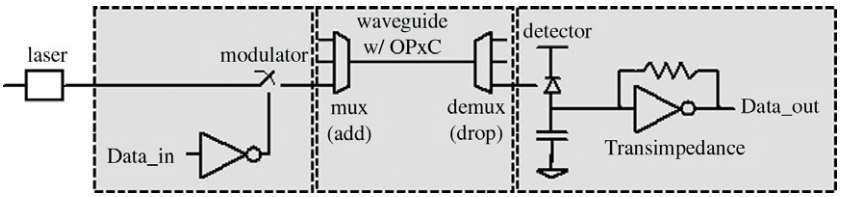

Fig. 3. Simplified block diagram of a macrochip WDM photonic link with an off-chip laser source. The receiver is represented by a photodetector and a transimpedance amplifier. Limiting amplifier stages and the clock and data recovery circuit that typically follows the receiver are not shown.

system operation with many photonic links, 10^{-23} for example, a high required link quality factor Q, 10 in this case, would be needed without error control. Consequently as Eq. (2) indicates, significantly more laser power will be needed to achieve such a high required link quality factor because of the relatively low laser efficiency.

B. Energy-Efficient Error Control

The macrochip expects to use photonic links with a BER of 10^{-12} [9]. With 8192 photonic links in the macrochip system, the effective BER for the system becomes very high. Our target undetected error rate for the system is 10^{-23}, which gives a very low probability for an undetected error in 10 years. Obviously there is a gap between the real optical link performance and the expected desirable link fidelity. One way to fill this gap is to improve the link signal-to-noise ratio (SNR) by increasing the laser power. But, as discussed above, this may significantly increase the total link power due to the low laser efficiency, and hence is not an energy-efficient approach. In addition, high laser optical power may manifest other impairments like waveguide nonlinearity. A better alternative approach is to use a robust error detection/correction method in conjunction with a link layer protocol to achieve the target undetected error rate. Unfortunately, error detection/correction will also add additional power and latency to the photonic link, consequently degrading the energy efficiency. In this section we discuss the requirements for such a scheme resulting in a minimum increase in the energy/useful bit metric for photonic links with undesirable BER.

To manage errors in a system, the error control mechanism should have two components: a component for detecting errors in received packets and a component for correcting the errors or retransmitting the packets. This is traditionally done in one of two ways: automatic repeat-request (ARQ) using error detection codes or FEC using error-correcting codes (ECC).

Powerful ECCs such as convolution codes could achieve error correction capability close to the Shannon limit. But the associated complexity, high power, and high latency prohibit its application for the macrochip. Relatively simple block codes, like Turbo codes [15] and Reed–Solomon codes [16–18], can also correct errors efficiently, but the power and latency required for hardware implementation are still too high. We estimate that implementing Turbo codes will consume power on the order of a few watts and that implementing Reed–Solomon codes will consume power on the order of a few tens of milliwatts. On the other hand, a simple ECC, such as

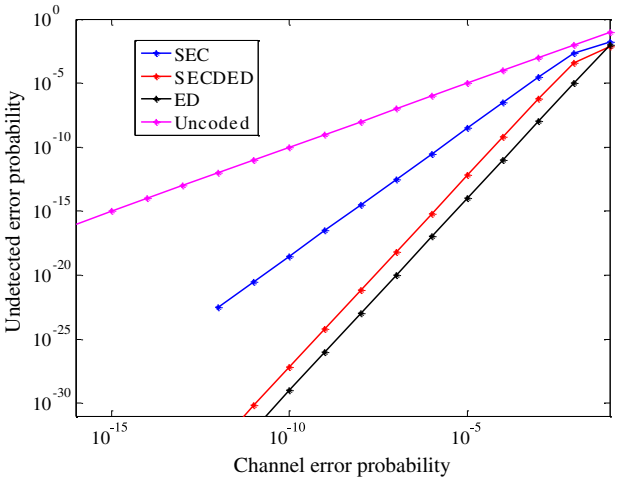

Fig. 4. (Color online) Undetected error probability for Hamming codes.

the Hamming code [19,20], does not have a significant impact on overall link energy efficiency, but it does have very limited error correction capabilities.

Preliminary error detection analysis for Hamming codes for 8-byte message transmission is shown in Fig. 4. As depicted in Fig. 4, Hamming codes have limited capability when used as an ECC, but with a Hamming distance of 3, an ECC can detect more errors than it can correct. If all of the detected errors can be corrected by retransmission, even with a high 10^{-8} raw physical channel BER, better than 10^{-25} residual BER can be achieved when it is used for error detection only.

Cyclic redundancy check (CRC) codes, on the other hand, can create larger Hamming distances [21–23], and therefore are more powerful in error detection. The undetected error probabilities at various packet sizes and CRC code sizes are shown in Figs. 5, 6, and 7. For the same 8-byte message length, CRC 8, 16, and 32 codes can produce code distances of 5, 6, and 10, respectively. With 10^{-8} raw physical channel BER, a better than 10^{-30} residual BER can be achieved using just the CRC 8 code.

As shown in Figs. 5, 6, and 7, a target undetected error rate of 10^{-23} for different packet sizes can be achieved even for physical links with a BER much higher than 10^{-12}. Furthermore, for a given optical link BER and the target undetected error rate, different CRC sizes can be used for different message sizes. The macrochip system uses message sizes from 8 bytes to 4096 bytes. In this case using a 32-bit CRC for all packet sizes will cause a large overhead for small

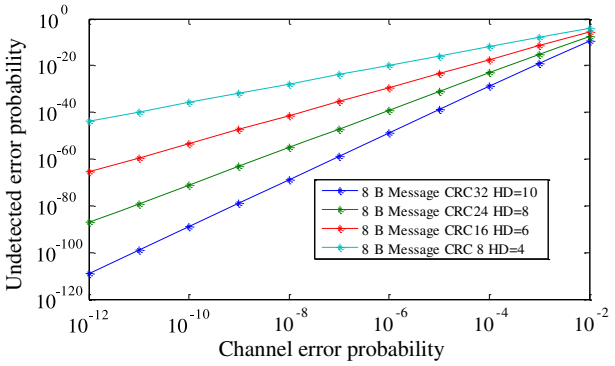

Fig. 5. (Color online) Undetected error probability of CRC codes for 8-byte packets.

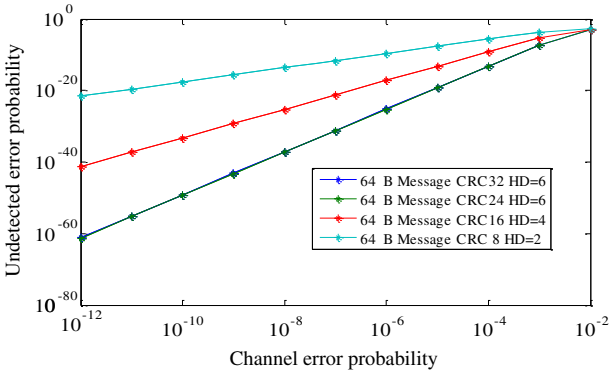

Fig. 6. (Color online) Undetected error probability of CRC codes for 64-byte packets.

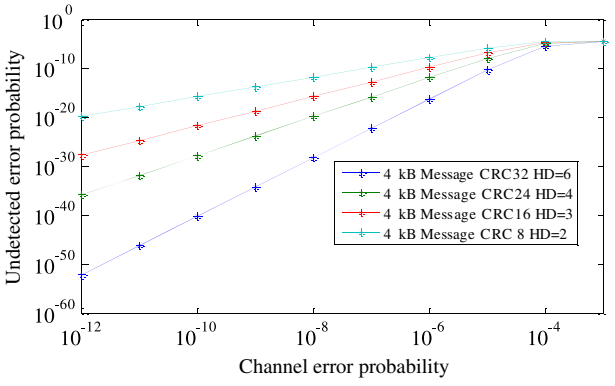

Fig. 7. (Color online) Undetected error probability of CRC codes for 4096-byte packets.

packets. Hence a variable CRC scheme that adds a CRC code with just enough bits to achieve the target undetected error rate, for that packet size, is required. From Figs. 5, 6, and 7, CRC lengths of 8, 16, and 32 bits for packet sizes of 8, 64, and 4096 bytes are appropriate.

Another important thing to consider for the error control scheme is its capability for handling the burst errors from the optical links. A burst error here is defined as contiguous sequence of error bits. Most of the optical link impairments may degrade the Q of the link, resulting in a higher BER, but they do not necessarily generate burst errors. Other sources of burst errors are associated with receiver design, optical crosstalk, and laser mode hops. Since decision feedback equalizer circuits are not used in our receiver design [24], there is no associated error propagation mechanism in our link. Optically, we avoid waveguide crossing completely by using inter-layer optical proximity couplers as described above. We do not consider laser mode hops in this paper and assume a stable optical source. With these assumptions, sequential errors become a probabilistic occurrence for a given BER. For example, for a link with a BER of 10^{-7}, the probabilities of 2-, 3-, and 4-bit errors in a row are 10^{-14}, 10^{-21}, and 10^{-28}, respectively. When CRC 8, 16, and 32 are used for packet sizes of 8, 64, and 4096 bytes, maximum Hamming distances of 5, 5, and 6 can be achieved, respectively, with certain codes, which means a minimum of up to 4-bit burst errors can be detected.

When using an error detection mechanism instead of an error correction mechanism, a protocol for retransmissions of corrupted packets is required. Such protocols are currently used in all forms of computer networks. These protocols rely on acknowledgments (ACKs) for correctly transmitted packets, negative acknowledgments (NACKs) for error packets, and timeouts for retransmission. ACK packets in the macrochip would be approximately 4 bytes in size; this is 50% of the size of the smallest packet. Hence transmitting one ACK for every correctly transmitted packet will waste bandwidth. Protocols such as sliding-window [25] do not ACK every correct packet; instead they ACK the receipt of multiple packets. The number of packets combined per ACK is adapted using timeouts. In the following sections we describe the average case behavior of such a protocol for the macrochip and quantify the performance impacts.

The properties of the error control protocol selected for the macrochip can be summarized as follows:

1. Error detection using CRC codes for energy-efficient error detection.

2. A variable CRC scheme for different packet sizes to reduce the CRC overheads.

3. A retransmission protocol that batches acknowledgments.

These mechanisms already exist in widely used error control techniques.

IV. Performance Simulation

A protocol that implements the properties laid out in the previous section will require an adaptive timeout-based protocol for retransmissions. In order to study the performance impact of different components of the error control protocol, we simulated the macrochip system with an approximate average case behavior of a protocol that meets the requirements.

The goals for the simulation modeling are the following:

1. Quantify the performance impact of CRC and protocol overheads (ACKs and NACKs) on the bandwidth usage and latency.

Preamble (01111110)	Type (00)	SequenceID (8 bits)	Size (14 bits)	Payload (8 to 4096 Bytes)	CRC (1, 2 or 4 Bytes)

Fig. 8. Data packet format used in the simulation.

2. Quantify the impact of BER on performance.

3. Quantify the send-buffer requirements for the system.

We simulated an 8×8 (64-site) macrochip system as described in Section II. The simulation model is an event driven simulator developed using the CSIM [26] discrete-event simulation engine. The model simulates the WDM point-to-point network on the macrochip at the specified link bandwidths. We used a packet generator at each site as the network traffic driver. The driver generates packets to random destinations with a uniform probability. These packets are of a fixed size, preset for every simulation run. The following sub-sections discuss the specifics of the protocol simulated.

A. The Simulated Packet Format

The data packet format used in the simulation is as shown in Fig. 8. The packet format used is not a strict requirement for the macrochip but is a reasonable approximation based on the requirements of the basic error detection mechanism.

The Preamble Field

Each packet in the macrochip starts with a specific bit pattern called the preamble. In order to maintain frame integrity we use a bit stuffing scheme as in the HDLC protocol [27] to make sure that no random data can be mistaken for the preamble. Because of bit stuffing, even if there is an error in a preamble, the recipient will resynchronize at the beginning of the next packet with a valid preamble. Any packets lost as a result of the error in the preamble will be retransmitted just as they would have been if there had been an error in the rest of the packet.

The Type Field

There are two major packet types supported by the macrochip protocol:

Positive acknowledgment (Type = 01): Acknowledgment packets are sent by the recipient to the sender to confirm receipt of a sequence of data packets. The received PktID contained in an acknowledgment packet identifies the SequenceID of the last correct packet that was received.

Negative acknowledgment (Type = 10): A negative acknowledgment packet is sent from recipient to source when an arriving packet is determined to have one or more bit errors. The Received PktID identifies the last correct packet that was received.

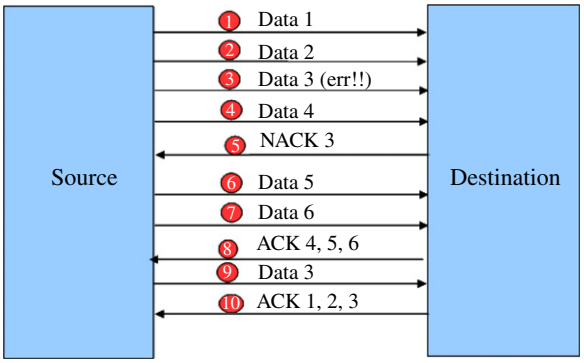

Fig. 9. (Color online) Simulated error control protocol.

The SequenceID Field

SequenceID is used to enable retransmission. When a data packet is received with an error, a NACK is sent to the sender with the same sequenceID as the data packet. A retransmission of the data packet will contain the same sequenceID.

The CRC Field

The CRC field is computed over all the bits of a data packet except the preamble. In this way any bit errors in the sequenceID, size, or the payload fields will be detected. To minimize the overhead for short packets, we used a variable CRC scheme with 8-, 16-, and 32-bit CRC codes for 8-, 64-, and 4096-byte packets.

B. The Simulated Error Control Protocol

The simulation model introduces errors into the packets at a specified BER. The model detects all error packets at the destination and follows the protocol shown in Fig. 9 for retransmissions. The protocol shown in Fig. 9 is not the exact protocol that would be used in a real system. In order to reduce the simulation complexity, we simulated an approximate version of an error control protocol that would meet the requirements stated in Section IV for the average case.

Steps 1 through 4, in Fig. 9, show the transmission of four data packets from a source on the macrochip to a destination, where the third packet has an error. A NACK packet is sent from the destination to the source in step 5, prompting a retransmission of the third data packet in step 9. The simulated protocol acknowledges every set of n

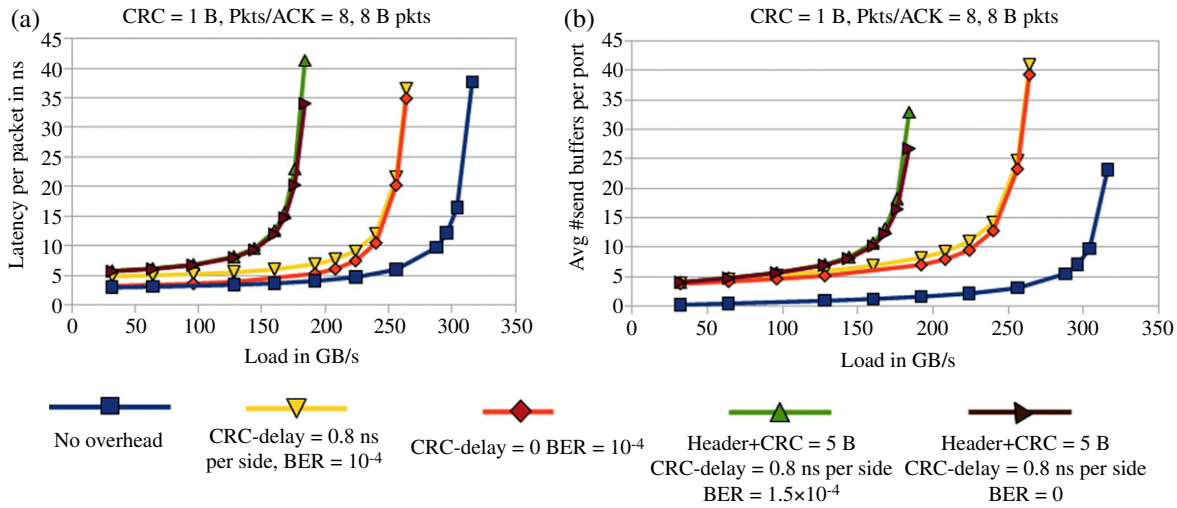

Fig. 10. (Color online) 8-byte performance and buffer requirement plots: (a) latency versus load plot; (b) buffers per port versus load plot.

packets transmitted from the source. In this example n = 3. Hence in step 8, an ACK of packets 4, 5, and 6 is transmitted from the destination to the source. After the correct retransmission of packet 3 in step 9, packets 1, 2, and 3 are acknowledged in step 10. The optimal value for n depends on the packet size. We performed simulations to determine the optimal n value for different packet sizes. The results will be discussed below. A real error control protocol will not generate an ACK for a fixed set of packets but instead will group a variable number of packets for an acknowledgment based on a timeout mechanism. The simulated retransmission mechanism approximates the average case behavior of a real protocol that on average generates one ACK for every n packets.

The following are the assumptions made by the simulated model:

1. The ACK/NACK and retransmit mechanism is a simplified average case behavior of a protocol that uses an adaptive timeout mechanism to group packets for an ACK.

2. The ACK/NACK packets are error free.

3. The bit errors are introduced using a uniform random probability.

4. Burst errors are not included as part of the simulation.

C. Simulation Results and Discussion

We simulated the WDM point-to-point network on the macrochip with the protocol described in the previous section with three packet sizes: 8 bytes, 64 bytes, and 4096 bytes. We define the term "input load" in a system as the rate (GB/s) at which traffic is generated into the network at each site. The peak bandwidth out of each site is 320 GB/s. For each message size we varied the input load on the system and measured the latency of each correctly transmitted packet. We define the latency as the time elapsed from the generation of the packet at the sending site to the error-free reception of the packet at the destination site. Hence the latency of the packet includes the NACK and retransmission latency if any.

As shown in Figs. (10a) through (12a), we plot the latency versus load for each message size. Each plot consists of five curves representing the following:

1. The blue curve shows latency versus load for a perfect system without any CRC and protocol overheads with BER = 0.

2. The orange curve shows the latency with a batched ACK scheme, a non-zero CRC, non-zero BER and a 0 ns CRC compute delay.

3. The yellow curve is the same as the orange curve but includes a 0.8 ns CRC compute delay.

4. The brown curve shows the latency with a batched ACK scheme, a header, and CRC; CRC compute latency = 0.8 ns, BER = 0.

5. The green curve shows the actual performance of the system with all overheads included and a non-zero BER.

The Impact of BER on Performance

The green and brown curves in Fig. 10(a) show that a BER = 10^{-4} has almost no performance impact on 8-byte packets. Figures 10(a), 11(a), and 12(a) show that better BERs are required for larger packets to reduce the performance impact. BER = 10^{-7} has almost no impact on performance of 4096-byte packets. This shows that a target BER = 10^{-7} for the point-to-point links in the macrochip is adequate for up to 4096-byte packets.

The Impact of CRC and Protocol Overheads

Comparing the orange and yellow curves in Figs. 10(a), 11(a), and 12(a) we see that the performance impact of CRC computation overheads on latency is negligible. The impact of the CRC calculation and the batched acknowledgments is

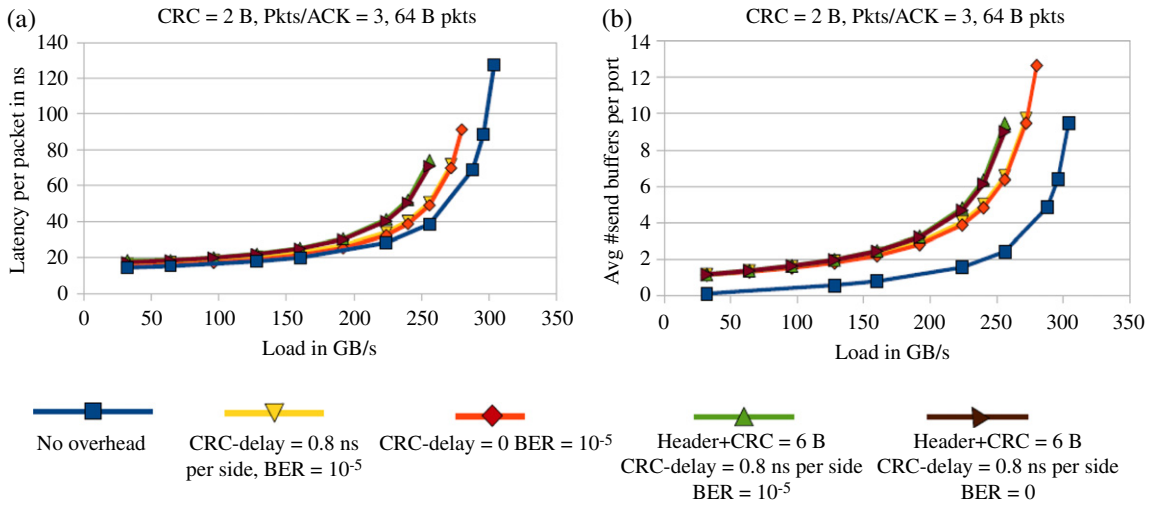

Fig. 11. (Color online) 64-byte performance and buffer requirement plots: (a) latency versus load plot; (b) buffers per port versus load plot.

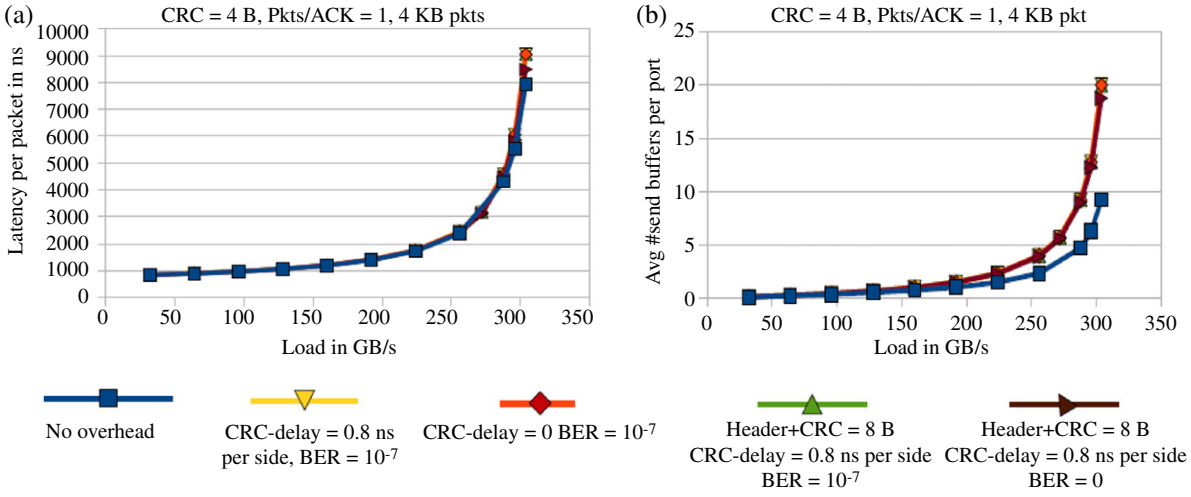

Fig. 12. (Color online) 4096-byte performance and buffer requirement plots: (a) latency versus load plot; (b) buffers per port versus load plot.

seen by comparing the yellow curve to the blue curve. The performance drops to 84% and 94% of that of the zero-overhead curve (blue curve) for 8-byte and 64-byte packets, respectively. The performance impact on 4096-byte packets is negligible.

In the case of 8-byte packets, theoretically, the overhead due to the 1-byte CRC should be 12.5%. The remaining overhead is due the ACK/NACK scheme. We performed a detailed simulation analysis (not shown) to determine the degree of batching required for the ACKs. The more ACKs we batch, the greater the buffer size requirements at the sender. This is because a send-buffer for a packet is not deallocated until the packet is acknowledged. In our analysis we found that an average of 8 packets/ACK and 3 packets/ACK seemed the most appropriate degree for batching for 8- and 64-byte packets, respectively.

By comparing the green and yellow curves we can estimate the impact of the packet headers on performance. In the case of 8-byte packets, the header overheads reduce the sustainable bandwidth to about 170 GB/s, which is 53% of the peak. In the case of 64-byte packets, the header overheads reduce the bandwidth to about 240 GB/s, which is 75% of the peak. The performance impacts of the protocol and CRC overheads are negligible for the 4096-byte packets.

Buffer Requirements

Figures 10(b) through 12(b) show the buffer size requirements for the three packet sizes. In the above figures, we fix the maximum allowed latency to be 3 times the minimum latency at low loads. On the basis of this requirement, we need about 12, 6, and 4 buffers per port for 8-, 64-, and 4096-byte messages, respectively. If we have to support all of the message sizes, we need a sufficient number of buffers for short messages and buffers of sufficient size for long messages. This leads to a

TABLE II
SUMMARY OF VALUES USED FOR ERROR CONTROL

Parameter	Value
Link BER (to meet system level goal)	10^{-7}
Batch length for ACKs	
For 8-byte packets	8
For 64-byte packets	3
For 4096-byte packets	1
Buffer requirements	
Number of 4 kB buffers/port	12
Space per site (simple allocation)	3 MB
Space per site (chained allocation)	1 MB

requirement of twelve 4 kB buffers per port and hence 3 MB of send-buffer space per site on the macrochip. By using smarter buffering mechanisms, such as chained buffers where a large buffer is made up of multiple small buffers, we can reduce the total number of buffers per port to four 4 kB buffers. This will reduce the send-buffer requirements per site to 1 MB. However, such schemes do add buffer allocation overheads.

Summary of the Simulation Results

On the basis of the results from the runs of the simulation model, we are able to select values of the key parameters of the error control scheme which allow us to meet the goals for accuracy, energy, and performance for the proposed system. These selected values are summarized in Table II.

V. OPTIMIZED INTERCONNECT ENERGY EFFICIENCY

As discussed above, photonic links in a tightly integrated large-scale macrochip system, unlike those in traditional long-reach wide-area networks, are well suited to a low-overhead CRC-based error detection scheme with retry. Using a simplified analysis, we suggest an adaptive protocol to minimize overheads and maximize effective throughput over a range of different message sizes. To quantify the energy costs of the resulting interconnect system, we can take the ratio of power to derated total network throughput:

$$\text{Energy/Useful Bit} = \frac{\text{Total Power}}{\text{Network Throughput}}. \quad (4)$$

On the basis of the network level simulation results shown above, we can combine our physical link costs with the energy cost of CRC circuits and the protocol overhead. We simulated a basic parallel 32-bit CRC implementation, synthesized, placed, and routed using 1.3 V, 130 nm CMOS technology, and scaled the results to 0.85 V, 28 nm technology; in doing so we conservatively assumed that while transistor capacitances scale, routing capacitance does not, due to increasingly restrictive design rules. The result was that we predict that a 32-bit CRC circuit will cost about 25 fJ/bit for each photonic link.

Figure 13 shows the energy per useful bit as a function of the average data length of packets in the system. With a physical channel energy of 160 fJ/bit [9], the network energy

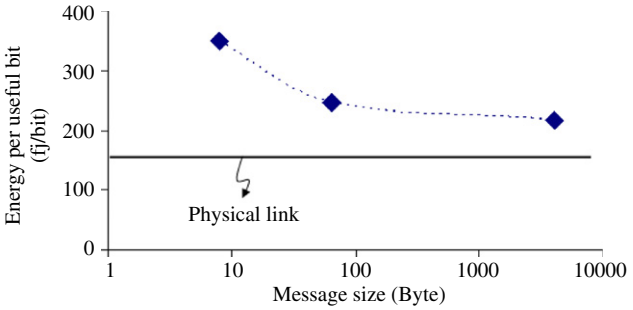

Fig. 13. (Color online) Energy per useful bit versus payload size.

goes to about 348 fJ/bit for short (8-byte) messages, 247 fJ/bit for 64-byte messages, and 215 fJ/bit for 4096-byte messages. The buffer memory power is not included in the calculation because buffers are needed even for networks using perfect links. As the simulation results above indicate, using links with increased raw channel BER does not significantly change the requirement for buffers needed. The number of buffers needed is primarily determined by the maximum allowed message latency.

VI. SUMMARY AND CONCLUSIONS

Future computer systems will require a large number of physical communication links to satisfy the bandwidth needs of increasingly dense and powerful processors. Silicon photonic links are a promising technology for building these systems in an energy-efficient manner. In the context of a 64-processor-site system with a fully connected point-to-point network consisting of 8192 silicon photonic links, we have analyzed the requirements for error control in these links. We also showed that error detection is preferable to error correction because of performance and energy overheads and presented an error detection scheme based on variable-length CRC codes and a protocol with batched acknowledgments. We presented performance simulation results that show that the scheme incurs an acceptable overhead while boosting the residual BER to 10^{-23}. We demonstrated that existing techniques can be adapted for the macrochip to achieve the BER goal, while minimizing the impact on energy use and performance.

ACKNOWLEDGMENTS

This material is based upon work supported, in part, by DARPA (the Defense Advanced Research Projects Agency) under Agreement No. HR0011-08-09-0001. The authors thank Dr. Jag Shah of DARPA MTO for his inspiration and support of this program. The views, opinions, and/or findings contained in this paper are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense. Approved for public release. Distribution unlimited.

REFERENCES

[1] K. Asanovic, R. Bodik, B. C. Catanzaro, J. J. Gebis, P. Husbands, K. Keutzer, D. A. Patterson, W. L. Plishker, J. Shalf, S. W. Williams, and K. A. Yelick, "The landscape of parallel computing research: a view from Berkeley," *Tech. Rep. UCB/EECS-2006-183*, EECS, UC Berkeley, 2006.

[2] S. Bell, B. Edwards, J. Amann, R. Conlin, K. Joyce, V. Leung, J. MacKay, M. Reif, L. Bao, J. Brown, M. Mattina, C.-C. Miao, C. Ramey, D. Wentzlaff, W. Anderson, E. Berger, N. Fairbanks, D. Khan, F. Montenegro, J. Stickney, and J. Zook, "Tile64-processor: a 64-core SoC with mesh interconnect," in *ISSCC*, 2008, pp. 88–598.

[3] S. Vangal, J. Howard, G. Ruhl, S. Dighe, H. Wilson, J. Tschanz, D. Finan, A. Singh, T. Jacob, S. Jain, V. Erraguntla, C. Roberts, Y. Hoskote, N. Borkar, and S. Borkar, "An 80- tile sub-100-W TeraFLOPS processor in 65-nm CMOS," *IEEE J. Solid-State Circuits*, vol. 43, no. 1, pp. 29–41, 2008.

[4] P. Kogge, K. Bergman, S. Borkar, D. Campbell, W. Carlson, W. Dally, M. Denneau, P. Franzon, W. Harrod, K. Hill, J. Hiller, S. Karp, S. Keckler, D. Klein, R. Lucas, M. Richards, A. Scarpelli, S. Scott, A. Snavely, T. Sterling, R. Williams, and K. Yelick, "Exascale computing study: technology challenges in achieving exascale systems," *DARPA IPTO Report*, 2008.

[5] Semiconductor Industries Association, *International Technology Roadmap for Semiconductors*, 2008 [Online]. Available: http://www.itrs.net/Links/2008ITRS/Home2008.htm.

[6] R. Drost, R. Hopkins, R. Ho, and I. Sutherland, "Proximity communication," *IEEE J. Solid-State Circuits*, vol. 39, no. 9, pp. 1529–1535, 2004.

[7] K. Kanda, D. Antono, K. Ishida, H. Kawaguchi, T. Kuroda, and T. Sakurai, "1.27 Gb/s/pin 3 mW/pin wireless superconnect (WSC) interface scheme," in *ISSCC*, 2003, vol. 1, pp. 186–487.

[8] J. Mitchell, J. Cunningham, A. V. Krishnamoorthy, R. Drost, and R. Ho, "Integrating novel packaging technologies for large scale computer systems," in *ASME/Pacific Rim Technical Conf. and Exhibition on Packaging and Integration of Electronic and Photonic Systems, MEMS, and NEMS (InterPACK 2009)*, June 2009, pp. 57–66.

[9] A. V. Krishnamoorthy, R. Ho, X. Zheng, H. Schwetman, J. Lexau, P. Koka, G. Li, I. Shubin, and J. Cunningham, "Computer systems based on silicon photonic interconnects," *Proc. IEEE*, vol. 97, no. 7, pp. 1337–1361, 2009.

[10] P. Koka, M. O. McCracken, H. Schwetman, X. Zheng, R. Ho, and A. V. Krishnamoorthy, "Silicon–photonic network architectures for scalable, power-efficient multi-chip systems," in *Proc. 37th Annu. Int. Symp. Computer Architecture*, 2010, pp. 117–128.

[11] W. Dally and B. Towles, *Principles and Practices of Interconnection Networks*. Morgan Kaufmann, 2004, p. 411.

[12] J. Orcutt, A. Khilo, M. Popovic, C. Holzwarth, B. Moss, H. Li, M. Dahlem, T. Bonifield, F. Kartner, E. Ippen, J. Hoyt, R. Ram, and V. Stojanovic, "Demonstration of an electronic photonic integrated circuit in a commercial scaled bulk CMOS process," in *Conf. on Lasers and Electro-Optics (CLEO)*, 2008, CTuBB3.

[13] A. V. Krishnamoorthy, J. E. Cunningham, X. Zheng, I. Shubin, J. Simons, D. Feng, H. Liang, C.-C. Kung, and M. Asghari, "Optical proximity communication with passively aligned silicon photonic chips," *IEEE J. Quantum Electron.*, vol. 45, no. 4, pp. 409–414, 2009.

[14] A. V. Krishnamoorthy, R. Ho, B. O'Krafka, J. E. Cunningham, J. Lexau, and X. Zheng, "Potentials of group IV photonics interconnects for 'red-shift' computing applications," in *4th IEEE Int. Conf. on Group IV Photonics*, 2007, pp. 180–182, PLE2.1.

[15] A. Burr, "Turbo-codes: the ultimate error control codes?" *Electron. Commun. Eng. J.*, vol. 13, no. 4, pp. 155–165, Aug. 2001.

[16] S. Gao, *Communications, Information and Network Security*. V. K. Bhargava, H. V. Poor, V. Tarokh, and S. Yoon, Eds., Kluwer Academic, 2003, ch. 5.

[17] I. S. Reed and G. Solomon, "Polynomial codes over certain finite fields," *J. Soc. Indust. Appl. Math.*, vol. 8, no. 2, pp. 300–304, 1960.

[18] L. Song, M. Yu, and M. S. Shafter, "A 10 Gb/s and 40 Gb/s forward-error-correction device for optical communications," in *ISSCC*, 2002, pp. 415–416.

[19] D. Bertozzi, L. Benini, and G. D Micheli, "Low power error resilient encoding for on-chip data bus," in *Proc. 2002 Design, Automation and Test in Europe Conf. and Exhibition (DATE '02)*, 4–8 Mar. 2002, pp. 102–109.

[20] S. Roman, *Coding and Information Theory*. J. H. Ewing, F. W. Gehring, and P. R. Halmos, Eds., Springer-Verlag, 1992, pp. 253–278, ch. 6.

[21] G. Castagnoli, S. Brauer, and M. Herrmann, "Optimization of cyclic redundancy-check codes with 24 and 32 parity bits," *IEEE Trans. Commun.*, vol. 41, no. 6, pp. 883–892, 1993.

[22] G. Castagnoli, J. Ganz, and P. Graber, "Optimum cyclic redundancy-check codes with 16-bit redundancy," *IEEE Trans. Commun.*, vol. 38, no. 1, pp. 111–114, 1990.

[23] P. Koopman and T. Chakravarty, "Cyclic redundancy code (CRC) polynomial selection for embedded networks," in *Proc. 2004 Int. Conf. Dependable Systems and Networks*, 2004, pp. 145–154.

[24] R. Ho, J. Lexau, F. Liu, D. Patil, R. Hopkins, E. Alon, N. Pinckney, P. Amberg, X. Zheng, J. E. Cunningham, and A. V. Krishnamoorthy, "Circuits for silicon photonics on a 'macrochip'," in *IEEE Asian Solid-State Circuits Conf.*, Nov. 2009.

[25] "Transmission Control Protocol," *RFC-793*, 1981.

[26] Mesquite Software, Inc., *CSIM 19 C++ Users' Guide*. 2005.

[27] "PPP in HDLC-like framing," *RFC-1662*, 1994.

Xuezhe Zheng (M'03–SM'03) received the B.S., M.S., and Ph.D. degrees in optical instruments from Tsinghua University, Beijing, China, in 1993 and 1997, respectively. He is currently a Consulting Hardware Engineer at Oracle Labs. He has extensive experience in photonic switching and optical cross-connects, fiber optic components, dense wavelength division multiplexing (DWDM) optical networks, and optical interconnections. His current research interests are in WDM silicon photonics for advanced inter/intra-chip interconnects.

Dr. Zheng is a recipient of the Science and Technology Development Award from the National Education Committee of China. He has author/co-authored more than 100 papers in technical journals and conferences and holds 12 US patents.

Pranay Koka received the M.S. degree in electrical and computer engineering from the University of Wisconsin, Madison, in 2005 and the M.S. degree in electrical engineering from Southern Illinois University, Edwardsville, in 2002. He joined Oracle Labs, a division of Oracle, in 2005 with the Computer Architecture and Performance group. At Oracle he developed architectures and simulation models for HPC systems and interconnects. His current research interests include photonic interconnect architectures, shared memory architectures, system simulation, and tracing methodologies.

Michael O. McCracken received the Ph.D. degree in computer science from the University of California, San Diego, in 2010. He is a Principal Member of Technical Staff at Oracle Labs, Oracle,

Austin. As a graduate student at UCSD, he worked closely with staff at the San Diego Supercomputer Center on performance modeling and optimization of large-scale scientific programs. He was a major contributor to a Gordon Bell Prize finalist entry at the Supercomputing 2007 conference. His research interests include performance modeling and optimization and performance tools for highly parallel systems and programs.

Herb Schwetman is a Principal Member of Technical Staff in Oracle Labs, a division of Oracle. He has been a member of the architecture/performance team in the Oracle–DARPA UNIC project since 2007. Prior to joining Sun in 2001, he was founder and CEO of Mesquite Software, Inc. He served as an Adjunct Professor in the Department of Computer Sciences at The University of Texas 1984–2009. He also worked for MCC, an R&D consortium located in Austin. Prior to 1984, he was a member of the faculty in the Department of Computer Sciences at Purdue University. He received his Ph.D. in computer sciences from The University of Texas.

James G. Mitchell has a B.S. (Honors) from the University of Waterloo (1966) and a Ph.D. from Carnegie-Mellon University (1971). He has been a Senior Fellow at Xerox PARC (1971–1984), a Senior Visiting Fellow at Cambridge University (1980–81), and Director of R&D at Acorn Computer, plc, UK (1984–1988). At Sun Microsystems (1988–2009) he was a Sun Fellow and Vice President of Sun Labs (1998–2003). He is currently Vice President, Photonics, Interconnects and Packaging, at Oracle Labs.

Dr. Mitchell has worked on programing language design and implementation (Mesa, Euclid, C++, Java), interactive programing systems, dynamic interpretation and compilation, document preparation systems, user interface design, distributed transactional file systems, distributed object-oriented operating systems, high-performance computer architecture, microprocessor design, and microelectronics packaging and interconnects.

Jin Yao received the Ph.D. degree in electrical engineering and computer sciences from the University of California, Berkeley in 2007. He is currently in the Photonics, Interconnects, and Packaging Group, Oracle Labs, San Diego, CA. His research areas include photonic components, links, networking systems, and interconnects, with a focus on innovations of device and system performance. His research interests also include photonic integrated circuits, MEMS/NEMS, and their system applications.

He has co-authored more than 40 papers in refereed journals and conference proceedings. He was also a California NanoSystems Institute (CNSI) Fellowship Award winner in 2002–2003.

Ron Ho (S'92–M'93–SM'08) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Stanford University. His background includes ten years at Intel Corporation. He is currently an Architect at Oracle Labs, Redwood Shores, CA, where he is engaged in a series of research programs centered around VLSI design, communications circuits, and advanced computer systems.

Kannan Raj (M'90, SM'04) has led many successful advanced engineering, product development, and R&D teams. Currently he serves as a Hardware Research and Operations Director in Oracle Labs, where he manages research teams in photonics, interconnects, and packaging and operations of the UNIC silicon photonics program. He has a Ph.D. from George Mason University, M.S.E.E. from Virginia Tech, and M.E. from the Indian Institute of Science, Bangalore. He holds 40 issued patents and has co-authored over 60 conference and journal publications.

Ashok V. Krishnamoorthy is a Hardware Architect at Sun Labs, Oracle, and Principal Investigator for the DARPA UNIC initiative on silicon "photonics-to-the-processor."

Previously, he was a Distinguished Engineer and Director at Sun Microsystems responsible for advanced optical interconnect and silicon photonics development. He also spent several years as CTO and President of AraLight, a Lucent technologies spinout developing high-density parallel optical products and technologies. Prior to that he was an entrepreneur-in-residence at Lucent New Ventures group, and before that a member of the technical staff in the Advanced Photonics research department at Bell Labs, in Holmdel, NJ. He has worked over two decades on the integration of photonic devices with silicon CMOS circuits and on building switching and computing sub-systems based on these components—including electro-optic modulators on silicon, quantum well devices, VCSELs, and, most recently, Si/Ge photonics. He has published over 200 technical papers and 8 book chapters and he holds 52 US patents. He has served as member or chair of over 30 international conferences and has been guest editor for several technical journals. His honors include the IEEE Distinguished Lecturer award, the ICO international prize in Optics, and the Chairman's Award from Sun Microsystems. He is a member of Tau Beta Pi, Eta Kappa Nu, and is a Fellow of the OSA.