On the Design of a Photonic Network-on-Chip

Assaf Shacham
Columbia University
Dept. of Electrical Engineering
500 W 120th St.,
New York, NY 10027
assaf@ee.columbia.edu

Keren Bergman
Columbia University
Dept. of Electrical Engineering
500 W 120th St.,
New York, NY 10027
bergman@ee.columbia.edu

Luca P. Carloni
Columbia University
Dept. of Computer Science
1214 Amsterdam Ave.,
New York, NY 10027
luca@cs.columbia.edu

Abstract

Recent remarkable advances in nanoscale silicon-photonic integrated circuitry specifically compatible with CMOS fabrication have generated new opportunities for leveraging the unique capabilities of optical technologies in the on-chip communications infrastructure. Based on these nano-photonic building blocks, we consider a photonic network-on-chip architecture designed to exploit the enormous transmission bandwidths, low latencies, and low power dissipation enabled by data exchange in the optical domain. The novel architectural approach employs a broadband photonic circuit-switched network driven in a distributed fashion by an electronic overlay control network which is also used for independent exchange of short messages. We address the critical network design issues for insertion in chip multiprocessors (CMP) applications, including topology, routing algorithms, path-setup and teardown procedures, and deadlock avoidance. Simulations show that this class of photonic networks-on-chip offers a significant leap in the performance for CMP intrachip communication systems delivering low-latencies and ultra-high throughputs per core while consuming minimal power.

1. Introduction

A major design paradigm shift is currently impacting high-performance microprocessors as critical technologies are simultaneously converging on fundamental performance limits. Essentially, the scaling in transistor speeds and integration densities can no longer drive the expected congruent multiples in computation performance [14]. Accelerated local processing frequencies have clearly reached a point of diminishing returns: further increases in speed lead to tighter bounds on the logic coherently accessed on-chip [3, 13] and the associated power dissipation is exacerbated in an exponential fashion [2, 22]. Evidence of this trend is unmistakable as practically every commercial manufacturer of high-performance processors is currently introducing products based on multi-core architectures: AMD Opteron, Intel Montecito, Sun Niagara, and IBM Cell and Power5. These systems aim to optimize performance per watt by operating multiple parallel processors at lower clock frequencies.

Clearly, within the next few years, performance gains will come from increases in the number of processor cores per chip [14, 22], leading to the emergence of a key bottleneck: the global intrachip communications infrastructure. Perhaps the most daunting challenge to future systems is to realize the enormous bandwidths capacities and stringent latency requirements when interconnecting a large number of processing cores in a power efficient fashion.

Low latency, high data-rate, on-chip interconnection networks have therefore become a key to relieving one of the main bottlenecks to CMP system performance. Significant research activity has recently focused on intra-chip global communication using packet-switched micro-networks [1,6,10,11,19]. These so-called networks-on-chip (NoC) are made of carefully-engineered links and represent a shared medium that is highly scalable and can provide enough bandwidth to replace many traditional bus-based and/or point-to-point links. However, with a fixed upper limit to the total chip power dissipation, and the communications infrastructure emerging as a major power consumer, performance-per-watt is becoming the most critical design metric for the scaling of NoCs and CMPs. It is not clear how electronic NoCs will continue to satisfy future communication bandwidths and latency requirements within the power dissipation budget.

Photonic interconnection networks offer a potentially disruptive technology solution with fundamentally low power dissipation that remains independent of capacity while providing ultra-high throughputs and minimal access latencies. One of the main drivers for considering photonic NoCs is the expected reduction in power expended on intrachip communications. The key power saving rises from the fact that once a photonic path is established, the data are transmitted end to end without the need for repeating, regeneration or buffering. In electronic NoCs, on the other hand, messages are buffered, regenerated and then transmitted on the inter-router links several times en route to their destination. In previous work [27] we provided a detailed power analysis of a photonic NoC, and compared it to an electronic NoC designed to provide the same bandwidth to the same number of cores. The compelling conclusion of the study was that the power expended on intrachip communications can be reduced by **two orders of magnitude** when **high-bandwidth communications** is required among the cores.

In this paper we explore the design and performance of an optical NoC that can capitalize on the enormous capacity, transparency, low latency, and fundamentally low power consumption of photonics. The construction of this optical NoC is based on remarkable advances made over the past several years in silicon photonics that have yielded unprecedented control over device optical properties. Fabrication capabilities and integration with commercial CMOS chip manufacturing that are now available open new exciting opportunities [8].

The optical NoC building blocks are nanoscale photonic integrated circuits (PICs) that employ optical microcavities, particularly those based on ring resonator structures shaped from photonic waveguides which can easily be fabricated on conventional silicon and silicon-on-insulator (SOI) substrates. This new class of small footprint PICs can realize extremely high interconnection bandwidths which consume less power and introduce less latency than their contemporary bulk counterparts. Compatibility with existing CMOS fabrication systems and the juxtaposition with silicon electronics enable direct driving, controllability, and the integration of these optical networks with processor cores and other silicon-based systems. By using photonic NoCs we exploit the unique advantages in terms of bandwidth, latency, and energy that have made photonics ubiquitous in long-haul transmission systems for on-chip interconnection networks.

High speed optical modulators, capable of performing switching operations, have been realized using these ring resonator structures [29, 30] (see Fig. 1) or the free carrier plasma dispersion effect in Mach-Zhender geometries [21]. The integration of modulators, waveguides and photodetectors with CMOS integrated circuits for off-chip communication has been reported and recently became commercially available [8]. On the receiver side, SiGe-based photodetectors and optical receivers were fabricated with reported high efficiencies [9]. Finally, low-loss waveguide technology with crossovers and a fairly aggressive turn radii has made some remarkable recent advances and the enabling technologies are currently available [4, 15].

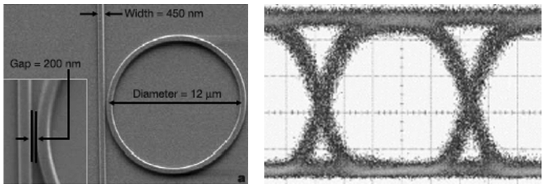

Figure 1. A silicon ring resonator [30] (left) and a 10 Gb/s eye-diagram from a silicon modulator [8] (right)

For the first time, the integration of a fully functional photonic system on a VLSI electronic die can be realistically envisioned, and, in particular, the photonic elements necessary to build a complete photonic NoC (dense waveguides, switches, modulators, and detectors) are now viable for integration on a single silicon chip. Fig. 1 shows published examples of some of the silicon electro-optic devices mentioned above.

In this paper we present a novel architecture for a photonic network-on-chip, and discuss its advantages while addressing critical design challenges. Network topology, routing algorithms, and flow control are discussed in detail. We developed an event-driven network simulator to quantitatively evaluate design aspects such as deadlock avoidance and recovery, path diversity and flow control. The conclusion of the quantitative study demonstrates the potential performance leap offered by the integration of a photonic micro-network within high-performance multi-core systems.

2 Architecture Overview

The photonic NoC architecture employs a hybrid design, synergistically combining an optical network for bulk message transmission and an electronic network, with the same topology, for distributed control and short message exchange.

While photonic technology offers unique advantages in terms of energy and bandwidth, two necessary functions for packet switching, namely buffering and processing, are very difficult to implement. Electronic NoCs which have many advantages in flexibility and abundant functionality tend to consume high power which scales up with the transmitted bandwidth [22]. The hybrid approach deals with this problem by employing two layers:

1. A photonic interconnection network, comprised of silicon broadband photonic switches interconnected by waveguides, is used to transmit high bandwidth messages.

2. An electronic control network, topologically identical to the photonic network, is used to control the photonic network and for the exchange of short control messages.

Every photonic message transmitted is preceded by an electronic control packet (a *path-setup* packet) which is routed in the electronic network, acquiring and setting-up a photonic path for the message. Buffering of messages is impossible in the photonic network, as there are no photonic equivalents for storage elements (e.g. flip-flops, registers, RAM). Hence, buffering, if necessary, only takes place for the electronic packets during the path-setup phase. The photonic messages are transmitted without buffering once the path has been acquired. This approach has many similarities with optical circuit switching, a technique used to establish long lasting connections between nodes in the optical internet core.

The main advantage of using photonic paths relies on a property of the photonic medium, known as *bit-rate transparency* [25]. Unlike routers based on CMOS technology that must switch with every bit of the transmitted data, leading to a dynamic power dissipation that scales with the bit rate [22], photonic switches switch on and off once per message, and their energy dissipation does not depend on the bit rate. This property facilitates the transmission of very high bandwidth messages while avoiding the power cost that is typically associated with them in traditional electronic networks. Another attractive feature of optical communications results from the *low loss in optical waveguides*: at the chip scale, the power dissipated on a photonic link is completely independent of the transmission distance. Energy dissipation remains essentially the same whether a message travels between two cores that are 2 mm or 2 cm apart. Furthermore, low loss off-chip interconnects such as optical fibers enable the *seamless scaling* of the optical communications infrastructure to multi-chip systems.

The photonic network is comprised of broadband 2×2 photonic switching elements which are capable of switching wavelength parallel messages (i.e. each message is simultaneously encoded on several wavelengths) as a single unit, with a sub-ns switching time. The switches are arranged as a two dimensional matrix and organized in groups of four. Each group is controlled by an electronic circuit termed electronic router to construct a 4×4 switch. This structure lends itself conveniently to the construction of planar 2-D topologies such as a mesh or a torus. A detailed explanation on the design of the photonic switching elements and the 4×4 switches is given in Section 3.

Two-dimensional topologies are the most suitable for the construction of the proposed hybrid network. The same reasons that made them popular in electronic NoCs, namely their appropriateness to handle a large variety of workloads and their good layout compatibility with a tiled CMP chip [6], still apply in the photonic case. Further, high-radix switches are very difficult to build with photonic switching elements so the low-radix switches, the building blocks of mesh/torus networks, are a better fit. Torus networks, which offer a lower network diameter, compared to meshes at the expense of having longer links [7], are the obvious choice since the transmission power on photonic links is independent of the length, unlike in copper lines.

Topological means can also be employed to overcome the lack of buffering in photonics. Since the photonic switching elements have small area and power consumption, many of them can be used to provision the network with additional paths on which circuits can be created, thus reducing the contention manifested as path-setup latency.

Electronic/Optical and Optical/Electronic (E/O and O/E) conversions are necessary for the exchange of photonic messages on the network. Each node therefore includes a network gateway serving as a photonic network interface. Small footprint microring-resonator-based silicon optical modulators with data rates up to 12.5 Gb/s [29] as well as 10 Gb/s Mach-Zehnder silicon modulators [8] and SiGe photodetectors [9] have been reported and have recently become commercially available [8], to be used in photonic chip-to-chip interconnect systems. The laser sources, as in many off-chip optical communication systems [8, 16] can be located off chip and coupled into the chip using optical fibers.

The network gateways should also include some circuitry for clock synchronization and recovery and serialization/deserialization. When traditional approaches are used, this circuitry can be expensive both in terms of power and latency. New technological opportunities enabled by the integration of photonics onto the silicon die may reduce these costs. An example of such an opportunity is an optical clock distribution network which can be used to provide a high-quality low-power clock to the entire chip thus alleviating the need for clock recovery in the gateways. In any case, the gateway design should account for these issues.

Since electronic signals are fundamentally limited in their bandwidth to a few GHz, larger data capacity is typically provided by increasing the number of parallel wires. The optical equivalent of this wire parallelism can be provided by a large number of simultaneously modulated wavelengths using wavelength division multiplexing (WDM) [4] at the network interfaces. The translating device, which can be implemented using microring resonator modulators, converts directly between space-parallel electronics and wavelength-parallel photonics in a manner that conserves chip space as the translator scales to very large data capacities [20, 31]. Optical time division multiplexing (OTDM) can additionally be used to multiplex the modulated data stream at each wavelength to achieve even higher transmission capacity [17].

The energy dissipated in these large parallel structures is not small, but it is still smaller then the energy consumed by the wide busses and buffers currently used in NoCs: the network gateway interface and corresponding E/O and O/E conversions occur once per node in the proposed system, compared to multiple ports at each router in electronic equivalent NoCs [27].

The employment of 4×4 switches places a unique constraint on the placement of the gateways. Since the photonic switches do not have a fifth port for injection/ejection the gateways require specially designed access point to permit injection and ejection of messages without interfering with pass-through traffic. The gateway placement policy and the design of access points will be described in Subsection 3.2.

2.1 Life of a Packet on the Photonic NoC

Finally, to illustrate the operation of the proposed NoC we describe the typical chain of events in the transmission of a message between two terminals. In this example, a *write* operation takes place from a processor in node A to a memory address located at node B. Both are arbitrary nodes connected through the photonic NoC.

As soon as the write address is known, possibly even before the contents of the message are ready, a path-setup packet is sent on the electronic control network. The packet includes information on the destination address of node B, and perhaps additional control information such as priority, flow id, or other. The control packet is routed in the electronic network, reserving the photonic switches along the path for the photonic message which will follow it. At every router in the path, a next-hop decision is made according to the routing algorithm used in the network.

When the path-setup packet reaches the destination node B, the photonic path is reserved and is ready to route the message. Since the photonic path is completely bidirectional a fast light pulse can then be transmitted onto the waveguide, in the opposite direction (from the destination node B to the source node A), signaling to the source that the path is open (using a technique presented similar to Ref. [26]). The photonic message transmission then begins and the message follows the path from switch to switch until it reaches its destination. Since special hardware and additional complexity are required to transmit and extract the counter-directional light pulses, an alternative approach can be used: This approach is based on transmitting the message when the path is assumed to be ready according to the maximum expected path reservation latency. While this scheme does not utilize the network resources as well as the first one, it requires less hardware resources.

After the message transmission is completed a path-teardown packet is finally sent to release the path for usage by other messages. Once the photonic message has been received and checked for errors, a small acknowledgement packet may be sent on the electronic control network, to support guaranteed-delivery protocols.

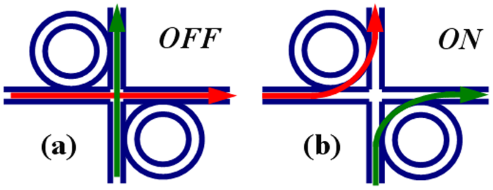

Figure 2. Photonic switching element: (a) *OFF* **state: a passive waveguide crossover. (b)** *ON* **state: light is coupled into rings and forced to turn**

In the case where a path-setup packet is dropped in the router due to congestion, a path-blocked packet is sent in the reverse direction, backtracking the path traveled by the path-setup packet. The path-blocked packet releases the reserved switches and notifies the node attempting transmission that its request was not served.

3 Network Design

In this section we describe in detail the proposed implementation of the photonic network and its electronic control layer, touching some key implementation issues.

3.1 Building Blocks

The fundamental building block of the photonic network is a *broadband photonic switching element (PSE)*, based on a ring-resonator structure. The switch is, in essence, a waveguide intersection, positioned between two ring resonators (Fig. 2). The rings have a certain resonance frequency, derived from material and structural properties. In the *OFF* state, when the resonant frequency of the rings is different from the wavelength (or wavelengths) on which the optical data stream is modulated, the light passes through the waveguide intersection uninterrupted, as if it is a passive waveguide crossover (Fig. 2a). When the switch is turned *ON*, by the injection of electrical current into p-n contacts surrounding the rings, the resonance of the rings shifts such that the transmitted light, now in resonance, is coupled into the rings making a right angle turn (Fig. 2b), thus creating a switching action.

Photonic switching elements and modulators based on the forementioned effect have been realized in silicon and a switching time of 30 ps has been experimentally demonstrated [29]. Their merit lies mainly in their extremely small footprint, approximately $12\mu m$ ring diameter and their low

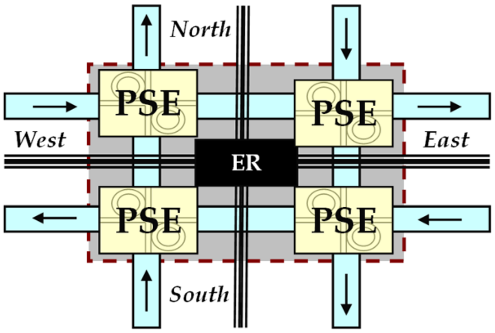

Figure 3. 4×4 switch. Four photonic switching elements (PSE) controlled by an electronic router (ER).

power consumption: less than 0.5 mW, when *ON*. When the switches are *OFF*, they act as passive devices and consume nearly no power. Ring-resonator based switches exhibit good crosstalk properties ($>$ 20 dB), and a low insertion loss, approximately 1.5 dB [28]. These switches are typically narrow-band, but advanced research efforts are now undergoing to fabricate wideband structures capable of switching several wavelengths simultaneously, each modulated at tens of Gb/s. It is also reasonable to assume that the loss figures can be improved with advances in fabrication techniques.

The PSEs are interconnected by silicon waveguides, carrying the photonic signals, and are organized in groups of four. Each quadruplet, controlled by an electronic circuit termed an *electronic router*, forms a 4×4 switch (Fig. 3). The 4×4 switches are, therefore, interconnected by the inter-PSE waveguides and by metal lines connecting the electronic routers. Control packets (e.g. path-setup) are received in the electronic router, processed and sent to their next hop, while the PSEs are switched *ON* and *OFF* accordingly. Once a packet completes its journey through a sequence of electronic routers, a chain of PSEs is ready to route the optical message. Owing to the small footprint of the PSEs and the simplicity of the electronic router, which only handles small control packets, the 4×4 switch can have a very small area. Based on the size of the microring resonator devices [29], and the minimal logic required to implement the electronic router, we estimate this area at 70 μm × 70 μm.

A keen observer will notice that the 4×4 switch in Fig. 3 is blocking. For example, a message routed from South to East will block message requests from West to South and from East to North. In general, every message which makes a **wide turn** (i.e. a turn involving 3 PSEs) may block two other message requests that attempt to make wide turns. Messages that make **narrow turns** (e.g. South to West) and messages that are routed straight through do not block other messages and cannot be blocked. To limit the blocking problem U-turns within the switches are forbidden. The blocking relationships between messages are summarized in Table 1.

Table 1. Inter-message blocking relationships in the 4×4 photonic switch

Current message	Blocked message I	Blocked message II
North→West	East→North	West→South
West→South	North→West	South→East
East→North	South→East	North→West
South→East	West→South	East→North

Being nonblocking is an important requirement from an atomic switch in an interconnection network. Nonblocking switches offer improved performance and simplify network management and routing. However, constructing a nonblocking 4×4 switch with the given photonic building blocks requires an exceedingly complex structure. This will have a negative impact on the area and, more importantly, the optical signal integrity, as each PSE hop can introduce additional loss and crosstalk. The design choice is, therefore, to use the blocking switch, because of its compactness, and bear its blocking properties in mind when designing the topology and a routing algorithm.

It is worth mentioning that different PSE grouping schemes can be used, where the directions of the waveguides are flipped, causing the blocking properties to slightly change. One possible scheme is to group the PSEs as a mirror-image of the current grouping scheme, where the directions of all waveguides are flipped. The analysis of this case is identical to the original grouping scheme. In yet another scheme, the direction of only one pair of waveguides is flipped (either the vertical or the horizontal). In this case each turning message may block one other message.

A related constraint, resulting from the switch structure, concerns the local injection/ejection port. Typically, 2-D mesh/torus NoCs use 5×5 switches, where one port is dedicated for local injection and ejection of packets. A 5×5 switch, while very simple to implement as an electronic transistor-based crossbar, is quite difficult to construct using 2×2 photonic switching elements. The injection and ejection of packets is, therefore, done through one of the 4 existing ports, thus blocking it for *through traffic*. This design decision places constraints on the topology, as described in the next subsection.

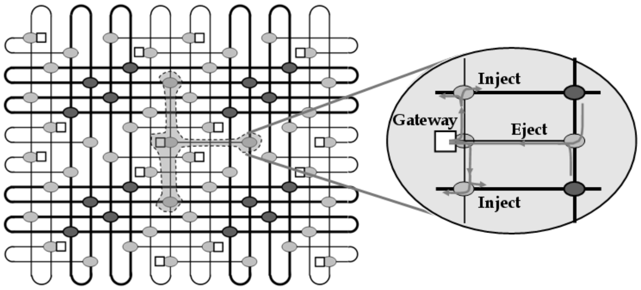

Figure 4. A 4-ary 2-D folded torus network (thick lines and dark ovals), access points (thin lines and light ovals), and 16 gateways (rectangles). One access point is shaded and enlarged.

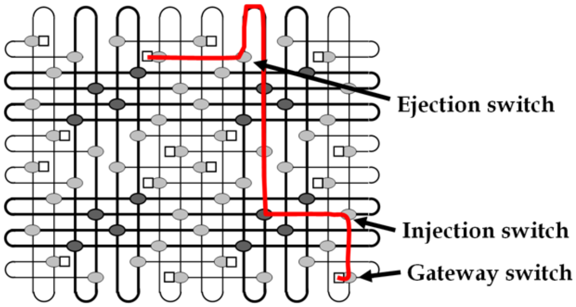

Figure 5. Example of a deadlock-avoiding path on the augmented folded torus network.

3.2 Topology

The topology of choice in our design reflects the characteristics of the entire system - a chip multiprocessor (CMP), where a number of homogeneous processing cores are integrated as tiles on a single die. The communication requirements of a CMP are best served by a 2-D regular topology such as a mesh or a torus [24]. These topologies match well the planar, regular layout of the CMP and the application-based nature of the traffic - any program running on the CMP may generate a different traffic pattern [7]. As mentioned above, a regular 2-D topology requires 5×5 switches which are overly complex to implement using photonic technology. We therefore use a folded torus topology as a base and augment it with access points for the gateways. An example of a 4×4 folded torus network, with the augmenting access points appears in Fig. 4.

The access points for the gateways are designed with two goals in mind: (1) to facilitate injection and ejection without interference with the through traffic on the torus, and (2) to avoid blocking between injected and ejected traffic which may be caused by the switches internal blocking. Injection-ejection blocking can be detrimental to the performance and may also cause deadlocks. The access points are designed such that gateways (i.e. the optical transmitters and receivers) are directly connected to a 4×4 switch (the gateway switch), through its West port (see Fig. 4). We assume, without loss of generality, that all the gateways are connected to the same port in their respective switches.

To avoid internal blocking a set of injection-ejection rules must be followed: injected messages make a turn at the gateway switch, according to their destination, and then enter the torus network through an injection switch. Messages are ejected from the torus network when they arrive to the ejection switch associated with their final destination. The ejection switches are located on the network, at the same row as the gateway switch, and this is the place where the ejecting messages turn. Finally, ejected messages pass through the gateway switch without making turns. An example of a path with the different kinds of switches is illustrated in Fig. 5.

Since torus networks are edge-symmetric [7], injection can be done at any port of the gateway switch, as long as the structure of the access point is rotated accordingly. An explanation on how this structure reduces contentions and avoids deadlocks is provided in Section 5.

The design of the access points contributes to a larger switch count in the network, as every access point requires 3 additional switches. However, each switch has rather small footprint and power dissipation, thus making the overall penalty minimal compared to the global power savings enabled by the photonic design [27].

A network designer may take advantage of the small footprint to improve the performance by increasing path diversity. Whenever the path-setup packet faces contention it is buffered in the electronic router until the blocking is cleared. The torus network can be augmented with additional paths, without changing the number of access points, so that the probability of blocking is lowered and the path-setup latency is, accordingly, reduced. Owing to the small footprint of the switches, the simplicity of the routers, and the fact that the PSEs only consume power when they cause messages to turn, the power and area cost of adding parallel paths is not large. The latency penalty that results from the increased hop-count should be balanced against the latency reduction achieved by mitigating contention. This study is performed in Section 7.

3.3 Routing

Dimension order routing is a simple routing algorithm for mesh and torus networks. It requires minimal logic in the routers and, being an oblivious algorithm, it does not require the routers to maintain a state or exchange additional information between them. We use XY dimension order routing on the torus network, with a slight modification required to accommodate the injection/ejection rules described in Subsection 3.1 above.

Each message is encoded with 3 addresses: 2 intermediate addresses and a final address, encapsulated within one another. The first intermediate address directs the message to the injection switch on torus network, thus causing the message to make the turn at the gateway switch, as required by the injection rules (see Fig. 5). The message is then routed on the torus, using plain XY dimension order routing, to the second intermediate address, the ejection switch (in the final destinations row, but one column away from it). Only then the final address is decapsulated and the message is forwarded to the destination gateway, where it arrives without having to turn, according to the ejection rules. The address encapsulation mechanism relieves the routers from processing system-scale considerations when setting up a path and preserves the simplicity of dimension order routing in the torus network.

When the torus network is path-diversified [7], (or *path-multiplied*, where several parallel lanes exist in each row/column), the address encapsulation mechanism can be used to take advantage of the path diversity while preserving the simplicity and obliviousness of dimension order routing. The encoding of the intermediate addresses can be done with the goal of balancing the load between parallel lanes, thus reducing the contention. According to this method the first intermediate address will be an injection switch on one of the lanes, as chosen by the gateway. The ejection, among the several parallel lanes, is also chosen by the gateway and encoded on the second intermediate address. The final address, of course, does not change. The selection of intermediate addresses is equivalent to choosing, at random, one among several *torus sub-networks* thus balancing the load among them.

In Section 7 we use the load-balancing approach when evaluating the effect of path diversity. Alternative intermediate address selection methods can be used such as restricting one lane to high priority traffic or allocating lanes to sources or designated flows.

3.4 Flow Control

The flow control technique in the network greatly differs from common NoC flow control methods. The dissimilarity stems from the fundamental differences between electronic and photonic technologies and mainly from the fact that memory elements (such as flip-flops and SRAM) cannot be used to buffer messages or even to delay them while processing is done. Electronic control packets are, therefore, exchanged to acquire photonic paths, and the data are only transmitted, with a very high bandwidth, once the path has been acquired.

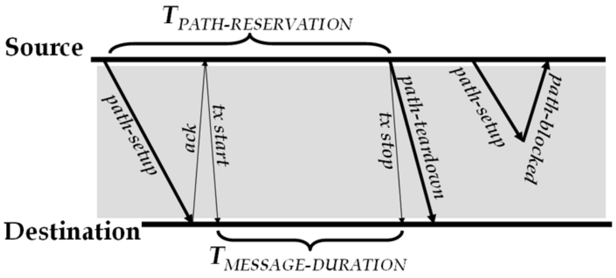

Figure 6. Qualitative timing diagram of a successful message setup (left) and a blocked setup request (right)

The path acquisition procedure requires the path-setup packet to travel a number of electronic routers and undergo some processing in each hop. Additionally, the packet may experience blocking at certain points in its path further contributing to the setup latency. Once the path is acquired, the transmission latency of the optical data is very short and depends only on the group velocity of light in a silicon waveguide: approximately 6.6×10^7 m/s, or 300 ps for a 2-cm path crossing a chip [15]. The network can, therefore, be perceived as a fast circuit-switched network, where the path-setup latency is much longer than the transmission latency. On the other hand, path-setup latency is still on the order of nanoseconds, a very short time compared to typical millisecond-setup-time circuit-switched networks. Hence, it can still be considered fast and can emulate packet switching traffic when packets are fairly large. The timing diagram in Fig. 6 illustrates the timing discrepancy.

The decision regarding the minimal size of the data unit exchanged on the photonic network must, therefore, take into account the path-setup latency. Exchange of small packets, such as memory read requests or cache-coherency snoop messages, for example, is clearly inefficient. The exchange of memory pages or long cache lines, instead, can utilize the photonic network much better. A good example is represented by the IBM/Toshiba/Sony Cell Broadband Engine processor where the bulk of the interconnection network traffic is made of DMA transactions [18]. For other applications, long lasting connections can be set up between processors that are expected to communicate frequently, providing a high-bandwidth link with minimal latency and low power consumption on which packets of any size can be transmitted.

The hybrid network architecture addresses the small packets exchange problem elegantly. Control messages that carry no data and are of a very small size, such as read requests, write acknowledgments, or cache snoops, can be exchanged on the control network which is, in essence, a low-bandwidth electronic NoC. These control messages are not expected to present a challenge for the control network because of their small size and will not require large resources in terms of additional circuitry or power.

Further, some applications, such as cryptanalysis for example, are characterized with exchange of very small data messages without any locality that can be exploited for grouping or speculative fetching. A CMP featuring the proposed hybrid architecture can utilize the electronic control network to exchange these massages at a reasonable performance.

In any case, it is of interest to study the optimal photonic message size in a given implementation of the network. This depends on the network size, on the latency of the individual components (routers, photonic links, electronic links, etc.), and on the bandwidth of the gateways. While one would want to minimize the setup overhead by using large messages, their size should be kept small enough to allow for good flexibility and link utilization. In Section 6 we analyze the optimal message size for the proposed network using an event-driven simulation model.

4 Simulation Setup

A key stage in the development of the ideas presented above is their functional validation using simulation. The correctness of the distributed path-setup procedure and the routing algorithm, for example, must be verified in a software environment that models accurately the network architecture. A quantitative performance study, using a variety of traffic loads, should also be carried out to evaluate algorithms, topologies and flow control techniques. This performance study also requires an accurate simulation model.

We developed POINTS (Photonic On-chip Interconnection Network Traffic Simulator), an event-driven simulator based on OMNeT++ [23]. OMNeT++ is an open source simulation environment that provides good support for modular structures, message-based communications between modules, and accurate modeling of physical layer factors such as delay, bandwidth and error rate.

The implemented model is highly parameterized to allow for a broad exploration of the design space. For the study in this paper the following design point is chosen: The system is a 36-core chip multiprocessor (CMP), organized in a 6×6 planar layout, built in a future 22 nm CMOS process technology. The chip size is assumed to be 20 mm along its edge, so each core is 3.3 mm × 3.3 mm in size. The network is a 6×6 folded-torus network augmented with 36 gateway access points (Fig. 4 presents a similar, albeit smaller, network), so it uses 144 switches, organized as 12×12. The electronic routers, each located at a center of a switch, are spaced by 1.67 mm and the PSEs (576 are used) are spaced by 0.83 mm.

The area and spacing considerations dictate the timing parameters of the network, as modeled in simulation. We assume a propagation velocity of 15.4 ps/mm in a silicon waveguide for the photonic signals [26] and 131 ps/mm in an optimally repeatered wire at 22 nm for the electronic signals traveling between electronic routers [12]. The inter-PSE delay and inter-router delay are, therefore, 13 ps and 220 ps respectively. The PSE setup time is assumed to be 1 ns and the router processing latency is 600 ps, or 3 cycles of a 5GHz clock.

Message injection processes in NoC simulation models are typically Bernoulli or modulated-Bernoulli processes, which work well with packet-switched slotted network. Since our architecture resembles circuit-switching more than packet- switching, we model the inter-message gap as an exponential random variable with a parameter μ_{IMG}. In the simulation reported in this paper we use uniform traffic. While this traffic pattern does not necessarily model the actual loads presented to the network in a CMP, it serves well as an initial measurement technique to demonstrate the capacity of the network and as a reference to use in future measurements.

In the following sections we describe three simulation-based studies performed using POINTS.

5 Dealing With Deadlock

Deadlock in torus networks has been studied extensively. When dimension order routing is used, no channel-dependency cycles are formed between dimensions, so deadlock involving messages traveling in different dimensions cannot occur [7]. Virtual channel flow control has been shown to be successful in eliminate intra-dimension deadlocks [5] and make dimension order routing deadlock free. Both these proofs assume that each router in the torus network is internally nonblocking.

As mentioned in Section 3, this is not the case in our network. Area and technology constraints lead us to use a 4×4 switch which has some internal blocking between messages. We recall that every wide turn in the switch may block two other wide turns. Messages that make narrow turns and messages that pass straight through do not block other messages and cannot be blocked, and U-turns are forbidden. Therefore, we must evaluate the topology, find when deadlocks may occur, and develop solutions to avoid them. The injection-ejection are explained in Section 3 and illustrated in Fig. 5. They include the separation of injection and ejection to different switches so that turns that may block other messages cannot occur in the same switch. To prove this we inspect each of the 3 switches comprising the access point:

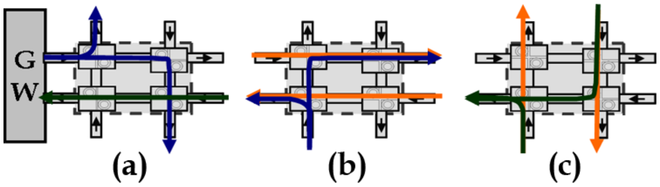

Figure 7. Gateway (a), injection (b), and ejection (c) switches. All possible message-paths are marked to demonstrate that no blocking interactions occur

- Gateway switch: Injected messages are required to make a turn towards the injection switches. Ejected messages arrive from the ejection message and pass straight through. Therefore, blocking cannot happen.
- Injection switch: messages already traveling on the torus network do not turn to the injection paths, so no blocking interactions exist between them and the injected messages.
- Ejection Switch: messages may arrive only from the torus network and they either turn for ejection or continue straight through. Since no messages arrive from the gateway switch, none of the blocking interactions may happen.

In Fig. 7 the three switches are shown with all the possible paths marked on them. The reader is invited to verify that none of the internal blocking scenarios, listed in Table 1, occur.

Even though injection-ejection blockings are completely avoided, and so are the associated performance penalty and possible deadlocks, the problem of intra-dimensional blocking of dimension order routing still remains. The accepted solution for this problem is virtual channel flow control [5] where the channel dependencies are removed by splitting the physical channel to several virtual channels that compete with each other for router bandwidth. This approach is difficult to implement in a circuit-switched network where the channel bandwidth cannot be divided between several circuits.

Hence, in our network we solve the intra-dimensional deadlock problem using path-setup timeouts. When a path-setup packet is sent, the gateway sets a timer to a pre-defined time. When the timer expires, a *terminate-on-timeout* packet is sent following the path-setup packet. The timeout packet follows the path acquired by the path-setup packet until it reaches the router where it is blocked. At that router, the path-setup packet is removed from the queue and a path-blocked packet is sent on the reverse path, notifying the routers that the packet was terminated and the path should be freed. If a deadlock has occurred, the system recovers from it at that point. While this method suffers from some inefficiency because paths and gateway injection ports are blocked for some time until they are terminated without transmitting, it guarantees deadlock-recovery.

In another possible scenario, the path-setup packet is not deadlocked but merely delayed and it reaches its destination while the timeout packet is en-route. In these cases the timeout packet reaches the destination gateway where it is ignored and discarded, and the path is acquired as if the timeout had not expired. This procedure has been tested in extensive simulations and has shown to be effective in resolving deadlocks.

6 Optimizing Message Size

In order to maintain the network efficiency as well as its flexibility and link utilization the message duration should be carefully picked. If too large messages are used, then link utilization is compromised as well as latency when messages are queued in the gateway for a long time while other long messages are transmitted. On the other hand, if messages are too small, then the relative overhead of the path-setup latency becomes too large and efficiency is degraded.

Of course, there is no technical reason preventing us from granting full freedom in message-sizing to each node, but this may cause starvation and unfairness. In this section we study the optimal size with respect to the overhead incurred in the path-setup process under the assumption that it is constant across all messages.

We define the *overhead ratio* as:

$$\rho = \frac{T_{path-reservation}}{T_{message-duration}}$$

where $T_{path-reservation}$ is the time between the transmission of the path-setup packet and the transmission of the path-teardown packet, and $T_{message-duration}$ is the time during which actual transmission takes place, corresponding to the size of the message (see Fig. 6). The smaller the value of ρ, the higher the network efficiency. In Fig. 8 we plot ρ as a function of the path length and $T_{message-duration}$, for a completely unloaded network. In this simulation messages of different sizes, addressed to all destination in the network, are injected at node (0,0), while all the other nodes do not transmit. No generality is lost when we inject from only a single node because of the edge-symmetry of the torus network.

The optimal message size is the smallest size which does not exceed a certain overhead ratio. As an arbitrary limit for an unloaded network we set the maximum allowed overhead to 20%. The maximum allowed overhead ratio is, therefore, ρ=1.25. In Fig. 8, where the 20% overhead line appears as a dashed line, we can see that the limit is met by

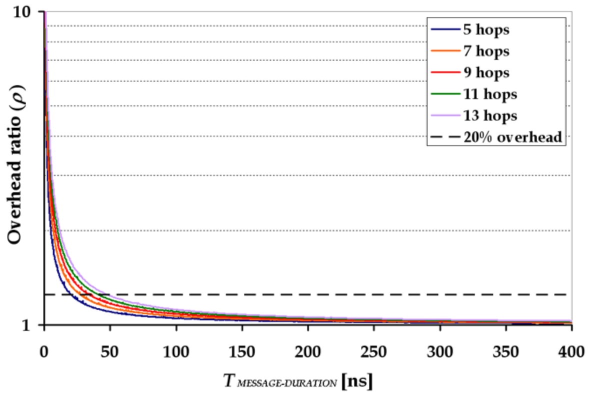

Figure 8. Overhead ratio as a function of path-length and message duration in a unloaded 12×12 torus network

messages with duration larger than 50 ns, for the longest path (13 hops). We therefore pick 50 ns to be the message duration in the network and use this duration in the rest of the simulations. It is worth mentioning that thanks to the huge bandwidth that can be transmitted in the photonic waveguides and the broadband switches (see Section 2), the amount of data that can be transmitted in 50 ns can be more than 2 KBytes, supporting the exchange of full memory pages or large DMA transactions.

Naturally the overhead will be larger when the network becomes loaded with traffic from other nodes, as path acquisition is expected to take longer due to blocking. To evaluate the effect of congestion on the message setup overhead we transmit 50-ns messages, from all nodes, with uniformly distributed addresses. The load on the network is managed by controlling the distribution parameter of the exponentially distributed inter-message gap (μ_{IMG}). The load offered (α) to the network is then given as:

$$\alpha = \frac{T_{message-duration}}{T_{message-duration} + \frac{1}{\mu_{IMG}}}$$

At the limit of constant transmission by all sources ($\frac{1}{\mu_{IMG}} \rightarrow 0$) the offered load approaches 1, and when the inter-message gap is very large ($\frac{1}{\mu_{IMG}} \rightarrow \infty$) the offered load approaches zero. The results of the congestion experiment are shown in Fig. 9.

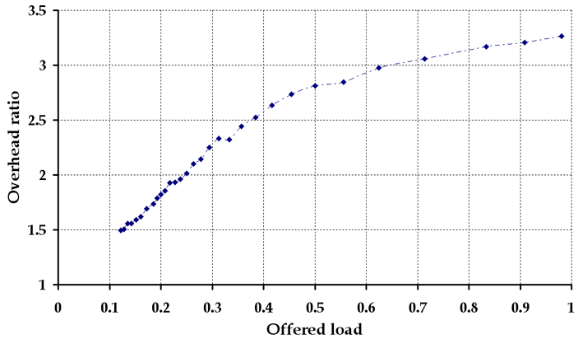

Figure 9. Overhead ratio for different path-lengths in an unloaded network

Fig. 9 reveals that the overhead in a loaded network, even lightly loaded, is larger, as was expected. The overhead ratio rises quickly to a value of 3 (or a path-setup latency of 100 ns) for loads exceeding a 0.6 value. Clearly the increased congestion and its detrimental effects on the latency must be dealt with. Adaptive routing algorithms, which use information about the availability of adjacent paths when making a routing decision, can be used to locate alternative paths for messages and reduce the blocking probability. Another technique is to increase the path diversity by augmenting the network with parallel lines. This approach is considered in the next section.

7 Increasing Path Diversity

One of the advantages of packet-switching networks lies in the statistical multiplexing of packets across channels and its extensive usage of buffers. These allow for distribution of loads across space and time. In a photonic circuit-switched network, there is no statistical multiplexing and buffering is impractical. Additional paths, however, can be provisioned, over which the load can be distributed using either random load-balancing techniques, or adaptive algorithms that use current information on the network load.

The topology chosen for the proposed network, a torus, can be easily augmented with additional parallel paths that provide path-diversity and facilitate this distribution of the load. The performance metric used to evaluate the improvement gained by adding the paths is again the path-setup overhead ratio, which is derived directly from the path-setup latency. Similarly to the previous experiment, we set $T_{message-duration}$ at 50 ns. T_{IMG} is exponentially distributed with a parameter μ_{IMG} which is, again, varied to control the offered load. Network with path diversity values of 1-4 are simulated, where a value of 1 represents the baseline 6×6 torus with 36 access points and a value of 4 represents a 24×24 torus, also with 36 access points. Naturally, path diversity has overheads in terms of hardware and increased zero-load latency as a result of the larger network diameter. Table 2 lists the numbers of switches required to implement each of the networks simulated. If we assume that the area of the 4×4 switch is about 5000 μm^2 then, theoretically, more than 80000 such switches can be integrated in the photonic layer of a 400 mm^2 die. The power dissi-

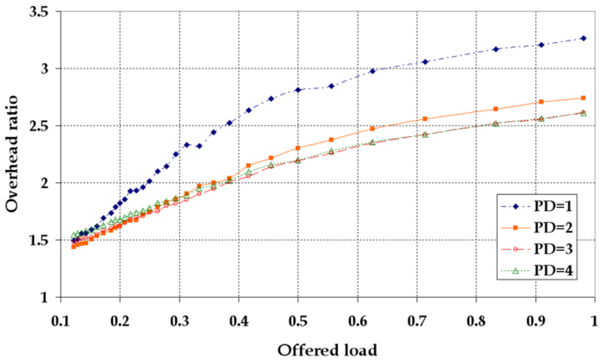

Figure 10. Overhead ratio vs. offered load for a 12×12 torus network with 36 gateway access points (324 switches).

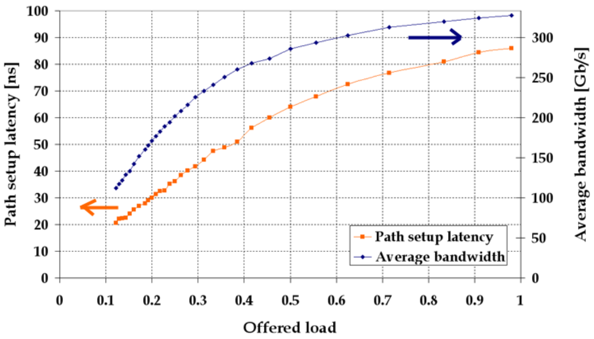

Figure 11. Latency and average bandwidth vs. offered load for a 12×12 torus network with 36 gateway access points (324 switches).

pated by the diversified network scales sub-linearly with the number of switches as switches only consume power when they cause a message-turn. The number of turns is fixed and independent of the number of switches, thereby setting a strict upper bound on the power expended in forwarding the photonic message regardless of the actual physical distance traveled [27].

Table 2. Switch counts for networks with different path-diversity values

PD value	Network	Gateway	Injection	Ejection	TOTAL
1	36	36	36	36	144
2	144	36	72	72	324
3	324	36	108	108	576
4	576	36	144	144	900

The simulation results are given in Fig.10. First, as expected, the increased network diameter caused by the provisioning of paths actually increases the latency when the network is lightly loaded and blocking is not frequent. As the network becomes congested, message blocking starts to dominate the path-setup latency and the additional paths, which reduce the blocking, contribute to a lower latency and a more efficient network. Second, a path-diversity degree of 4 is certainly unnecessary in the test case and in fact it performs poorly compared to a network with a path diversity of 3 because it does not reduce the blocking probability significantly, while, on the other hand, it increases the network diameter.

As a final measurement, Fig. 11 illustrates the motivation for the integration of a photonic network on chip, i.e. the immense bandwidth that can be routed using wavelength division multiplexing and broadband switches. Fig 11 shows the path-setup latency and the average bandwidth available per port for a network with a path diversity value of 2 as a function of the offered load, when 50-ns messages are injected, assuming a peak bandwidth of 960 Gb/s using OTDM and WDM. It can be observed that even under low loads and relatively low average latencies of 30 ns an average bandwidth of 230 Gb/s can be provided to each core, while using simple routing algorithms and circuit-setup methods. The bisection bandwidth corresponding to this operating point is 2.07 Tb/s.

8 Conclusions

Recent extraordinary advances in the fabrication of silicon photonic devices and their integration with CMOS electronics on the same die open a new field of opportunities for on-chip and off-chip interconnection networks designers. In this paper we gave a detailed presentation of a new architectural approach based on these opportunities: a network-on-chip combining a photonic circuit-switched network for high-bandwidth bulk data transmission and an electronic network which controls the photonic network while providing a medium for the exchange of short messages. The presentation covered critical design issues such as topology, routing algorithms, path-setup/teardown procedures, and deadlock avoidance rules and recovery procedures.

We developed POINTS, an event-driven simulator, and used it to validate the architecture and to complete a set of studies showing that low-latency low-power photonic links can be set up within tens of nanoseconds. By limiting the message duration in the circuit-switched photonic network to a short time, one can emulate a high-bandwidth packet-switched network, with an average bandwidth on the order of hundreds of Gb/s per core.

While the exchange of data at such bandwidths leads to exceedingly high power dissipation in electronic NoCs, in a photonic NoC, end-to-end paths are formed across a chain of low-power transparent switching elements. Power consumed in routing the high-bandwidth messages can therefore be dramatically reduced. Photonic NoCs can present a true leap in the sheer performance of intrachip interconnection networks and, more importantly, in their performance per watt.

References

[1] L. Benini and G. D. Micheli. Networks on chip: A new SoC paradigm. *IEEE Computer*, 49(2/3):70–71, Jan. 2002.

[2] R. W. Brodersen, M. A. Horowitz, D. Marković, B. Nikolić, and V. Stojanović. Methods for true power minimization. In *Intl. Conf. on Computer-Aided Design*, pages 35–42, Nov. 2002.

[3] L. P. Carloni and A. L. Sangiovanni-Vincentelli. Coping with latency in SoC design. *IEEE Micro*, 22(5):24–35, Sept./Oct. 2002.

[4] X. Chen, B. G. Lee, X. Liu, B. A. Small, I.-W. Hsieh, K. Bergman, J. Richard M. Osgood, and Y. A. Vlasov. Demonstration of 300 Gbps error-free transmission of WDM data stream in silicon nanowires. In *Conference on Lasers and Electro-Optics (CLEO'07)*, May 2007.

[5] W. J. Dally and C. L. Seitz. Deadlock-free message routing in multiprocessor interconnection networks. *IEEE Trans. Comput.*, 36(5):547–553, May 1987.

[6] W. J. Dally and B. Towles. Route packets, not wires: On-chip interconnection networks. In *Design Automation Conf.*, pages 684–689, June 2001.

[7] W. J. Dally and B. Towles. *Principles and Practices of Interconnection Networks*. Morgan Kaufmann, San Francisco, CA, 2004.

[8] C. Gunn. CMOS photonics for high-speed interconnects. *IEEE Micro*, 26(2):58–66, Mar./Apr. 2006.

[9] A. Gupta, S. P. Levitan, L. Selavo, and D. M. Chiarulli. High-speed optoelectronics receivers in SiGe. In *17th Intl. Conf. on VLSI Design*, pages 957–960, Jan. 2004.

[10] A. Hemani, A. Jantsch, S. Kumar, A. Postula, J. Oberg, M. Millberg, and D. Lindqvist. Network on chip: An architecture for billion transistor era. In *18th IEEE NorChip Conference*, Nov. 2000.

[11] S. Heo and K. Asanović. Replacing global wires with an on-chip network: a power analysis. In *Intl. Symp. on Low Power Elect. and Design (ISLPED 2005)*, pages 369–374, Aug. 2005.

[12] R. Ho. Wire scaling and trends. a presentation at MTO DARPA meeting, Aug. 2006.

[13] R. Ho, K. W. Mai, and M. A. Horowitz. The future of wires. *Proc. IEEE*, 89(4):490–504, Apr. 2001.

[14] M. A. Horowitz, E. Alon, D. Patil, S. Naffziger, R. Kumar, and K. Bernstein. Scaling, power, and the future of CMOS. In *IEEE Intl. Electron Devices Meeting*, Dec. 2005.

[15] I.-W. Hsieh, X. Chen, J. I. Dadap, N. C. Panoiu, J. Richard M. Osgood, S. J. McNab, and Y. A. Vlasov. Ultrafast-pulse self-phase modulation and third-order dispersion in si photonic wire-waveguides. *Optics Express*, 14(25):12380–12387, Dec. 2006.

[16] L. A. Johansson, Z. Hu, D. J. Blumenthal, L. A. Coldren, Y. A. Akulova, and G. A. Fish. 40-GHz dual-mode-locked widely tunable sampled-grating DBR laser. *IEEE Photon. Technol. Lett.*, 17(2):285–287, Feb. 2005.

[17] S. Kawanishi, H. Takara, K. Uchiyama, I. Shake, and K. Mori. 3 Tbit/s (160 Gbit/s×19 channel) optical TDM and WDM transmission experiment. *Electronic Letters*, 35(10):826–827, 13 May 1999.

[18] M. Kistler, M. Perrone, and F. Petrini. Cell multiprocessor communication network: Built for speed. *IEEE Micro*, 26(3):10–23, May/June 2006.

[19] R. Kumar, V. Zyuban, and D. M. Tullsen. Interconnections in multi-core architectures: Understanding mechanism, overheads, scaling. In *ISCA '05: 32nd annual international symposium on Computer architecture*, June 2005.

[20] B. G. Lee, B. A. Small, Q. Xu, M. Lipson, and K. Bergman. Demonstrated 4×4 Gbps silicon photonic integrated parallel electronic to WDM interface. In *Optical Fiber Communications Conf. (OFC)*, Mar. 2007.

[21] L. Liao, D. Samara-Rubio, M. Morse, A. Liu, D. Hodge, D. Rubin, U. D. Keil, and T. Franck. High speed silicon mach-zehnder modulator. *Optics Express*, 13(8):3129–3135, 18 Apr. 2005.

[22] T. Mudge. Power: A first-class architectural design constraint. *IEEE Computer*, 34(4):52–58, 2001.

[23] OMNeT++ discrete event simulation system. available online at http://www.omnetpp.org/.

[24] T. M. Pinkston and J. Shin. Trends toward on-chip networked microsystems. *Intl. J. High Performance Computing and Networking*, 3(1):3–18, 2001.

[25] R. Ramaswami and K. N. Sivarajan. *Optical Networks: A Practical Perspective*. Morgan Kaufmann, San Francisco, CA, second edition, 2002.

[26] A. Shacham and K. Bergman. Building ultra low latency interconnection networks using photonic integration. *IEEE Micro*, Mar./Apr. 2007. to be published.

[27] A. Shacham, K. Bergman, and L. P. Carloni. Maximizing GFLOPS-per-Watt: High-bandwidth, low power photonic on-chip networks. In *P=ac^2 Conference*, pages 12–21, Oct. 2006.

[28] F. Xia, L. Sekaric, and Y. A. Vlasov. Ultracompact optical buffers on a silicon chip. *Nature Photonics*, 1:65–71, Jan. 2007.

[29] Q. Xu, S. Manipatruni, B. Schmidt, J. Shakya, and M. Lipson. 12.5 Gbit/s carrier-injection-based silicon microring silicon modulators. *Optics Express*, 15(2):430–436, 22 Jan. 2007.

[30] Q. Xu, B. Schmidt, S. Pradhan, and M. Lipson. Micrometre-scale silicon electro-optic modulator. *Nature*, 435:325–327, 19 May 2005.

[31] Q. Xu, B. Schmidt, J. Shakya, and M. Lipson. Cascaded silicon micro-ring modulators for WDM optical interconnection. *Optics Express*, 14(20):9430–9435, 02 Oct. 2006.