Photonic Networks-on-Chip for Future Generations of Chip Multiprocessors

Assaf Shacham, *Member*, *IEEE*, Keren Bergman, *Senior Member*, *IEEE*, and Luca P. Carloni, *Member*, *IEEE*

Abstract—The design and performance of next-generation chip multiprocessors (CMPs) will be bound by the limited amount of power that can be dissipated on a single die. We present photonic networks-on-chip (NoC) as a solution to reduce the impact of intrachip and off-chip communication on the overall power budget. The low loss properties of optical waveguides, combined with bit-rate transparency, allow for a photonic interconnection network that can deliver considerably higher bandwidth and lower latencies with significantly lower power dissipation than an interconnection network based only on electronic signaling. We explain why on-chip photonic communication has recently become a feasible opportunity and explore the challenges that need to be addressed to realize its implementation. We introduce a novel hybrid microarchitecture for NoCs that combines a broadband photonic circuit-switched network with an electronic overlay packet-switched control network. This design leverages the strength of each technology and represents a flexible solution for the different types of messages that are exchanged on the chip; large messages are communicated more efficiently through the photonic network, while short messages are delivered electronically with minimal power consumption. We address the critical design issues including topology, routing algorithms, deadlock avoidance, and path-setup/teardown procedures. We present experimental results obtained with POINTS, an event-driven simulator specifically developed to analyze the proposed design idea, as well as a comparative power analysis of a photonic versus an electronic NoC. Overall, these results confirm the unique benefits for future generations of CMPs that can be achieved by bringing optics into the chip in the form of photonic NoCs.

Index Terms—On-chip communication, chip multiprocessors, photonics, emerging technologies.

1 Introduction

In the continual drive toward improved microprocessor performance, power efficiency has emerged as a prime design consideration. In fact, the limitations on power dissipation imposed by packaging constraints have become so paramount that performance metrics are now typically measured *per unit power* [1]. At the chip scale, the trend toward multicore architectures and chip multiprocessors (CMPs) for driving performance-per-watt by increasing the number of parallel computational cores is dominating new commercial releases [2], [3], [4], [5], [6]. With the future path clearly toward further multiplication of the on-chip processing cores, CMPs have begun to essentially resemble highly parallel computing systems integrated on a single chip. In this context, the role of the interconnect and associated global communication infrastructure is becoming central to the chip performance. As with highly parallel systems, performance is increasingly tied to how efficiently information is exchanged and how well the growing number of computational resources are utilized. Thus, global on-chip communications will play a central role in the overall performance of future CMPs.

The realization of a scalable on-chip communication infrastructure faces critical challenges in meeting the large bandwidth capacities and stringent latency requirements demanded by CMPs in a power-efficient fashion [7], [8]. Recent research on packet-switched networks-on-chip (NoC) [9], [10], [11], [12] has shown that carefully engineered shared links can provide enough bandwidth to replace many traditional bus-based communication media and point-to-point links. However, NoCs do not directly address the power dissipation challenge. With vastly increasing on-chip and off-chip communication bandwidths, the interconnect power consumption is widely seen as an acutely growing problem. It is unclear how electronic NoCs will continue to satisfy future bandwidths and latency requirements within the CMP power budget [13].

The insertion of *photonics* in the on-chip global interconnect structures for CMP can potentially leverage the unique advantages of optical communication and capitalize on the capacity, transparency, and fundamentally low energy consumption that have made photonics ubiquitous in long-haul transmission systems. The construction of photonic NoC could deliver performance-per-watt scaling that is simply not possible to reach with all-electronic interconnects. The photonics opportunity is made possible now by recent advances in nanoscale silicon photonics and considerably improved photonic integration with commercial CMOS chip manufacturing [14]. Unlike prior generations of photonic technologies, the remarkable capabilities of nanoscale silicon photonics offer the possibility of

- *A. Shacham is with Aprius Inc., 440 N. Wolfe Rd., Sunnyvale, CA 94085. E-mail: assaf@ee.columbia.edu.*
- *K. Bergman is with the Department of Electrical Engineering, Columbia University, 500 W. 120th St., 1300 Mudd, New York, NY 10027. E-mail: bergman@ee.columbia.edu.*
- *L.P. Carloni is with the Department of Computer Science, Columbia University, 466 Computer Science Building, 1214 Amsterdam Avenue, Mail Code: 0401, New York, NY 10027-7003. E-mail: luca@cs.columbia.edu.*

Manuscript received 5 July 2007; revised 3 Mar. 2008; accepted 13 Mar. 2008; published online 28 Apr. 2008.
Recommended for acceptance by R. Marculescu.
For information on obtaining reprints of this article, please send e-mail to: tc@computer.org, and reference IEEECS Log Number TCSI-2007-07-0305.
Digital Object Identifier no. 10.1109/TC.2008.78.

creating highly integrated photonic platforms for generating and receiving optical signals with fundamentally superior power efficiencies. These tremendous gains in power efficiencies for optical modulators and receivers are driven by the nanoscale device footprints and corresponding capacitances, as well as by the tight proximity of electronic drivers enabled by the monolithic CMOS platform integration [15], [16], [17], [18], [19]. Photonic elements have recently become available as library cells in standard CMOS processes. For the first time, we can practically consider basing the communication infrastructure of a CMP on a photonic interconnection network.

In particular, photonic NoCs can deliver a dramatic reduction in power expended on intrachip global communications while satisfying the high bandwidths requirements of CMPs. Photonic NoCs change the rules of power scaling: As a result of low loss optical waveguides, once a photonic path is established, the data is transmitted end-to-end without the need for repeating, regenerating, or buffering. In electronic NoCs, on the other hand, a message is buffered, regenerated, and then transmitted on the interrouter links multiple times en route to its destination. Furthermore, the switching and regenerating elements in CMOS consume dynamic power that grows with the data rate. The power consumption of optical switching elements, conversely, is independent of the bit rate, so, once generated, high-bandwidth messages do not consume additional dynamic power when routed.[1]

In this paper, we present *photonic* NoC as a solution for high-performance CMP design which leverages the remarkable progress in silicon photonics to offer a major reduction in the power dissipated on intrachip communications. The intrachip photonic infrastructure also offers seamless off-chip communications. Specifically, we propose a hybrid NoC microarchitecture that combines a photonic circuit-switched network with an electronic packet-switched network. We envision that, in the span of three or four CMOS process generations, a similar photonic NoC will be implemented as an additional layer of optical and optoelectronic devices grown on top of the silicon die and the metal layers comprising the CMP and possibly with multiple memory planes in between. This will likely be realized using 3D Integration (3DI) based on through-silicon via technology [20] in order to separately optimize logic, memory, and Si photonics planes. Further, current trends in multicore architectures suggest that CMPs will soon host a few dozen complex cores, each containing multiple logic blocks including one or more processing units, a local memory, a direct memory access (DMA) memory controller, and a network interface. In our vision, the photonic NoC will be the global communication medium connecting these cores among themselves and with off-chip memories and devices.

1.1 Paper Organization and Contribution

Early versions of the work presented here were reported in previous conference publications [21], [22], [23]. In this paper, we collect the main results presented in these papers and further extend the work with an improved power dissipation estimation model and new performance simulation results. This paper is organized as follows:

Section 2 briefly reviews prior work done by researchers on the integration of optical communication elements in electronic integrated circuits and, specifically, in microprocessors. In Section 3, we give an overview of the hybrid microarchitecture, we explain the rationale behind its choice and we describe the most important photonic components that characterize it. In Section 4, we discuss in detail the critical design issues for the photonic NoC including: technology building blocks, network topology, routing algorithms, deadlock avoidance, and path-setup/teardown procedures.

We developed POINTS, an event-driven network traffic simulator, to quantitatively evaluate critical design aspects such as deadlock avoidance/recovery, optimal message size, path multiplicity (PM), and alternative flow control mechanisms. In Section 5, we report a series of simulation-based experimental results that broadly confirm the potential performance leap offered by the integration of a photonic NoC in future high-performance CMPs. In Section 6, we present a comparative power analysis of a photonic NoC versus an electronic NoC that is designed to provide the same bandwidth to the same number of cores. The compelling conclusion of the study is that the power expended on intrachip communications can be reduced by nearly two orders of magnitude when *high-bandwidth communications* is required among a large number of cores.

Last, we comment on future research avenues.

2 Related Work

Optical communication is widely accepted as an interconnection medium for long and medium-reach distances, typically above 10 m [24]. A large body of research work exists on the design, fabrication, and performance analysis of optical interconnects for short-reach applications such as chip-to-chip interconnection. Studies about intrachip applications for optical interconnects are not as widely available because copper interconnects, until recently, have performed sufficiently well in addressing intrachip communication needs within power constraints.

Collet et al. [25] have studied the relative performance of optical and electrical on-chip interconnects for CMOS processes between 0.7 and 0.05 μm. They have concluded that the penetration of on-chip optical interconnects can be envisioned in lengths larger than 1,000 times the wavelength (e.g., 45 μm in a 45 nm process) where they can have lower power and latency than electronic interconnects. The work assumes the lasers are integrated into the silicon die and are directly modulated, thus consuming the bulk of the power of the optical system.

A multicore processor architecture where remote memory accesses are implemented as transactions on a global on-chip optical bus is suggested by Kirman et al. [26]. The work shows a latency reduction as high as 50 percent for some applications and a power reduction of about 30 percent over a baseline electrical bus. Because this design is based on bus topology, it suffers from obvious scalability limits. The simulated design connects 64 processing cores organized in

1. While this is true for the photonic network, the power consumption of other network components (e.g., E/O and O/E conversion) does scale with the bit rate, but it is still significantly lower than that of an electronic NoC. A power analysis follows in Section 6.

four supernodes. It is expected that bus contention will limit performance when a larger number of nodes are connected in the bus. Additionally, optical buses are limited in the number of terminals due to the finite launching power and coupling losses incurred by each terminal.

An optical NoC based on a wavelength-routed crossbar is presented by Briére et al. [27]. The crossbar, comprised of passive resonator devices and routing between an input-output pair, is achieved by selecting the appropriate wavelength. This approach, however, requires either widely tunable laser sources or large arrays of fixed-wavelength sources with fast wavelength-selection switches. The performance of such a system will strongly depend on the ability to select a wavelength quickly and accurately and its scalability will be limited by the number of fixed sources (or the tuning range, if tunable lasers are used).

Intel's Technology and Manufacturing Group performed a study evaluating the benefits of optical intrachip interconnects [28]. Their conclusion is that, while optical clock distribution networks are not especially attractive, wavelength division multiplexing (WDM) does offer interesting advantages for intrachip optical interconnects over copper in deep-submicron processes.

Our work builds on these projects and suggests a system where optical interconnects are used for intercore communication, thus replacing the global interconnects which are generally long and stretch across the chip. The penetration length is reduced by using on-chip modulators and simple off-chip constant-wave laser sources [14]. The off-chip lasers are cooled separately, thus dramatically reducing the chip's power and heat density. The topology used is of a distributed network, which is scalable to a large number of terminals. Current silicon technology is leveraged to design a system which both consumes low power and is feasible for fabrication in today's or near-term silicon-photonic technology.

3 Hybrid NoC Microarchitecture

The photonic NoC microarchitecture employs a hybrid design synergistically combining an optical circuit-switched network for bulk message transmission and an electronic packet-switched network for distributed control and short message exchange. Hence, the term *hybrid* has a twofold meaning: It denotes both the concept of combining a circuit-switched network and a packet-switched network as well as the idea of combining electronic and photonic technologies.

While photonic technology offers unique advantages in terms of energy and bandwidth, two necessary functions for packet switching, namely, buffering and header processing, are very difficult to implement with optical devices. On the other hand, electronic NoCs do have many advantages in flexibility and abundant functionality, but tend to consume high power, which scales up with the transmitted bandwidth [29]. The hybrid approach that we propose 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 large messages.
2. An electronic control network, topologically identical to the photonic network, is "folded" within the photonic network to control its operations and execute the exchange of short 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 not currently feasible 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 can be seen as *optical circuit switching:* The established paths are, in essence, optical circuits (or transparent lightpaths) between processing cores, thus enabling low-power, low-latency, high-bandwidth communications.

The main advantage of using photonic paths relies on a property of the photonic medium, known as *bit-rate transparency* [24]: 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 [29], 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 or between two chips that are tens of centimeters apart—low loss off-chip interconnects enable the seamless scaling of the optical communication infrastructure to multichip systems.

3.1 Exploiting Photonics in NoC Design

The proposed NoC is comprised of broadband 2×2 photonic switching elements (PSEs) interconnected by optical waveguides. The PSEs can switch wavelength parallel messages (i.e., each message is simultaneously encoded on several wavelengths) as a single unit, with a subnanosecond switching time. The switches are arranged as a 2D matrix and organized in groups of four. Each group is controlled by an electronic circuit termed *electronic router* (ER) to construct a 4×4 switch. This structure lends itself conveniently to the construction of planar 2D topologies such as a mesh or a torus. A detailed explanation on the design of the PSEs and the 4×4 switches is given in Section 4.

Two-dimensional topologies are the most suitable for the construction of the proposed hybrid microarchitecture. The same reasons that made them popular in electronic NoCs, namely, their appropriateness for handling a large variety of workloads and their good layout compatibility with a tiled CMP chip [10], still apply in the photonic case. Further, large-radix switches are very difficult to construct using PSEs, so

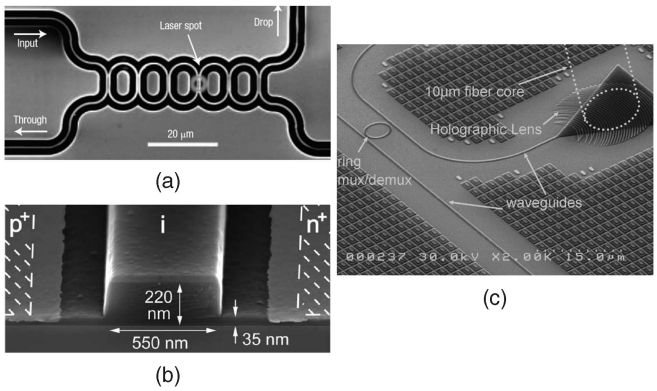


Fig. 1. Building blocks examples. (a) A silicon nanophotonic wavelength-insensitive switch [19]. (b) An ultracompact 10 Gbps silicon modulator [16]. (c) CMOS-compatible waveguides and holographic fiber-coupling lens for off-chip access [14].

the low-radix switches, the building blocks of mesh/torus networks, are a better fit. A key advantage of photonic implementations of meshes and tori is related to the nature of the guided waves. When two waveguides intersect at a right angle, as they do many times in mesh and torus networks, the waves continue propagating in their original direction and the crosstalk is negligible. This property enables the construction of the photonic NoC in *a single layer*, above the metal stack, thus reducing the fabrication complexity, the chip dimensions, and the total cost.

Torus networks offer a lower network diameter compared to meshes, at the expense of having longer links [30]. Hence, they are a better choice for photonic NoCs since the transmission power on photonic links is independent of the length, unlike in copper lines. Topology can also be employed to address issues caused by the lack of buffering in photonics. Since the PSEs 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 NoC. Each core in the CMP, 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 Gbps [31] as well as 10 Gbps Mach-Zehnder silicon modulators [14], [16] and SiGe photodetectors [32] have been reported and have recently become commercially available [14] for photonic chip-to-chip interconnect systems (see Fig. 1). The laser sources, as in many off-chip optical communication systems [14], can be located off chip and coupled into the chip using optical fibers or, alternatively, can be bonded to the silicon die, constructing hybrid-evanescent laser sources [33].

The network gateways 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 of 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 provide a high-quality low-power clock to the entire chip, simplifying the clock recovery in the gateways.

Since electronic signals are fundamentally limited in their bandwidth to a few gigahertz, larger data capacity is typically obtained 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 WDM [34] 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 [35], [36]. The energy dissipated in these large parallel structures is not small, but it is still smaller then the energy consumed by the wide buses and buffers currently used in NoCs. The network gateway interface and corresponding E/O and O/E conversions occur once per core in the proposed system, compared to multiple ports at each router in electronic equivalent NoCs. A study of the power dissipated by the proposed hybrid NoC and a comparison with an all-electronic NoC architecture is given in Section 6.

3.2 Life of a Message in the Photonic NoC

To illustrate the operation of the proposed NoC, we describe the typical chain of events in the transmission of a message between two ports placed on different cores in the CMP, for example, a *write* operation that takes place from a processing unit in a core to a memory that is located in another core. 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 destination address information and, perhaps, additional control information such as priority or flow ID. 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 port, the photonic path is reserved and is ready to route the message. Since the photonic path is completely bidirectional, a short light pulse can then be transmitted onto the waveguide in the opposite direction (from the destination to the source), signaling to the source that the path is open. This technique is similar to the one described in detail by Shacham and Bergman in [37]. When the optical pulse is received at the message source, the optical link is established. The photonic message transmission then begins and the message follows the path from switch to switch until it reaches its destination.

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

In the case where a path-setup packet is dropped in the router due to congestion, a *path-blocked* packet is transmitted

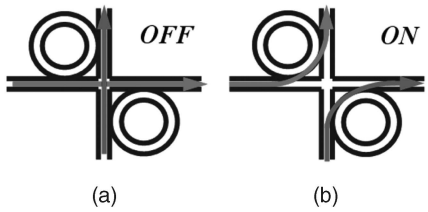


Fig. 2. PSE: (a) *OFF* state: A passive waveguide crossover. (b) *ON* state: Light is coupled into rings and forced to turn.

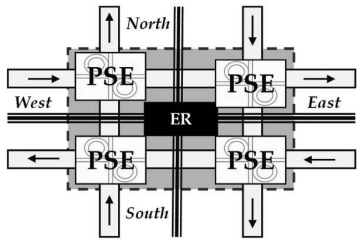


Fig. 3. A 4×4 switch. Four PSE controlled by an electronic router (ER).

by the dropping router to the source, backtracking the path traveled by the path-setup packet. The path-blocked packet releases the reserved switches and notifies the core attempting transmission that its request was not served. The source may then attempt transmission again and take advantage of PM in the network.

4 Network Design

The design of the photonic NoC requires an approach fundamentally different, in many aspects, from electronic NoCs. In this section, we describe in detail the proposed implementation, including the network's electronic and photonic building blocks, topology, routing algorithms, and flow control.

4.1 Building Blocks

The main building block of the photonic NoC is a broadband *PSE*, based on a microring-resonator structure. A similar device, although optically pumped, was recently reported in [19]. The switch is, in essence, a waveguide intersection, positioned between two ring-shaped waveguide structures (i.e., microring resonators). The rings have a certain resonance frequency, derived from material and structural properties. The PSE can be in one of two possible states:

- *OFF state:* The resonant frequency of the rings is different from the wavelength (or wavelengths) on which the optical data stream is modulated. Hence, the light passes through the waveguide intersection uninterrupted, as if it is a passive waveguide crossover (Fig. 2a).
- *ON state:* The switch is turned on by the injection of electrical current into p-n contacts surrounding the rings; the resonance of the rings shifts so that the light, now on resonance, is coupled into the rings, making a right angle turn, thus causing a switching action (Fig. 2b).

Photonic switching elements and modulators based on these effects have been realized in silicon and a switching time of 30 ps has been experimentally demonstrated [31]. Their merit lies mainly in their extremely small footprint, with ring diameters of approximately 12 μm, and their low power consumption of less than 0.5 mW of DC power when *ON* and approximately 1 pJ for modulating narrowband single-wavelength signals. For switching multiwavelength broadband signals, the ring resonators are designed as comb-pass filters with somewhat larger footprints, consuming 10 mW when *ON* [15], [18]. When the switches are *OFF*, they act as passive devices consuming nearly no power. Ring-resonator-based switches exhibit good crosstalk properties (> 20 dB), and a low insertion loss, approximately 1.5 dB [38].

Recent results reported in [19] (see Fig. 1a) demonstrate an optically pumped PSE with a measured insertion loss of 2.5 dB in the pass-band, capable of simultaneously switching nine 40 Gbps wavelengths. The switch is compact $(40 \times 12\mu\text{m})$ and has a switching time < 2 ns. It is reasonable to assume that the loss figures can be improved with advances in fabrication techniques and that electrically pumped devices, necessary to enable fabrication and electronic control will be developed.

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 *ER*, forms a 4×4 switch (Fig. 3). The 4×4 switches are therefore interconnected by the inter-PSE waveguides and by the metal lines connecting the ERs. Control packets (e.g., path-setup) are received in the ER, 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 ERs, a chain of PSEs is ready to route the optical message. Owing to the small footprint of the PSEs and the simplicity of the ER, which only processes small control packets, the 4×4 switch can have a very small area. Based on the size of the microring resonator devices [19], [31] and the minimal logic required to implement the ER, this area is estimated to be about $70 \times 70\mu\text{m}$.

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 that makes a *wide turn* (i.e., a turn involving three 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.

It is an important requirement for an atomic switch to have a nonblocking property 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 has a negative impact on the area and, more importantly, the optical signal integrity, as each PSE hop introduces additional loss and crosstalk. The design choice is, therefore, to use the blocking switch because of

TABLE 1
Intermessage Blocking Relations 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

its compactness and to bear its blocking properties in mind when designing the network topology and 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, 2D mesh/torus NoCs use 5×5 switches, where one port is dedicated for local injection and ejection of packets. A 5×5 switch is very simple to implement as an electronic transistor-based crossbar, but it is quite difficult to build using 2×2 PSEs. The injection and ejection of packets is therefore done through one of the four existing ports, blocking it for through traffic. This design decision constrains the topology, as described in Section 4.2.

4.2 Topology

The topology of choice in our design reflects the characteristics of the entire system—a 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 2D regular topology such as a mesh or a torus [39]. These topologies well match 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 [30]. As explained above, a regular 2D 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. Fig. 4 illustrates an example of a 4×4 folded torus network with the augmenting access points.

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 of the

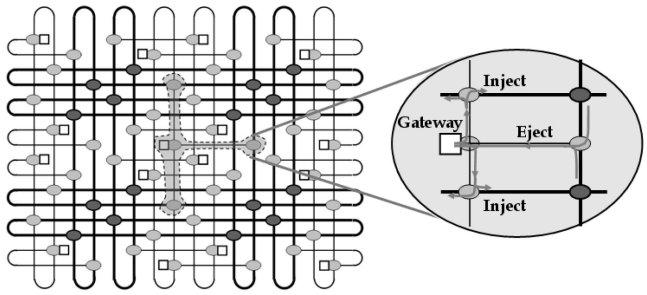


Fig. 4. A 4-ary 2D folded torus network (thick lines and dark ovals), access points (thin lines and light ovals), and 16 gateways (rectangles). One access point is shaded and enlarged.

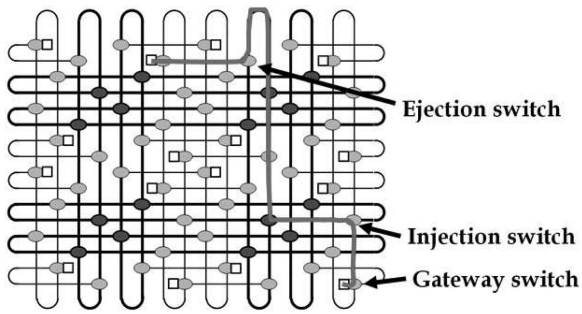


Fig. 5. Deadlock-avoiding path on the augmented folded torus network.

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 at the ejection switch associated with their final destination. The ejection switches are located on the network, in 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. In Fig. 5, the switches are marked, along with an example of a path.

Since torus networks are edge-symmetric [30], 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 of how this structure can be exploited to reduce contentions and avoid deadlocks is given in Section 5.2.

The design of the access points contributes to a larger switch count in the network because every access point requires three additional switches. However, each switch is rather small in footprint and power consumption. Consequently, as shown in Section 6, the overall penalty is minimal compared to the global power savings enabled by the photonic technology.

Topological means can also be exploited to reduce contention-generated latency. A network designer, however, may take advantage of the small footprint to improve the performance by increasing the *PM* in the network: Specifically, 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. Due 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 such that an optimal latency point is found. This issue is studied in detail in Section 5.4.

4.3 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 simplifies the router design in terms of area and power consumption. We use XY dimension-order routing on the torus network, with a slight modification required to accommodate the injection/ejection rules described in Section 4.2 above.

Each message is encoded with three addresses: two intermediate addresses and a final address, encapsulated within one another. The first address directs the message to the injection switch on the torus network, 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, i.e., the ejection switch, in the final destination's row, one column away from it. Only then is the final address "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-multiplied* such that several parallel lanes exist in each row/column, the address encapsulation mechanism can be used to take advantage of the PM while preserving the simplicity and obliviousness of dimension-order routing [30]. The encoding of 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 subnetworks*, thus balancing the load among them. In Section 5.4, we use the load-balancing approach when evaluating the effect of PM. Alternative methods to select an intermediate address can be used such as restricting one lane to high priority traffic or allocating lanes to sources or designated flows.

4.4 Flow Control

The flow control technique in our NoC differs remarkably from flow control methods that have been previously proposed for NoCs. The dissimilarity stems from the fundamental differences between electronic and photonic technologies and, particularly, from the fact that memory elements (such as flip-flops and SRAM) cannot be used to buffer messages or even to delay them while contention resolution and header processing are 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.

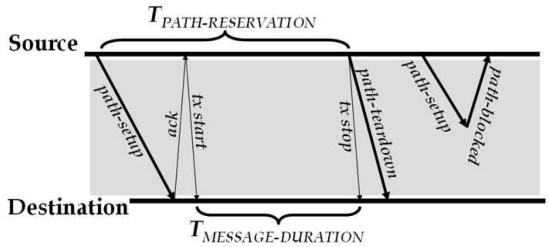


Fig. 6. Qualitative timing diagram of (left) a successful message setup and (right) a blocked setup request.

The path acquisition procedure requires the path-setup packet to travel a number of ERs 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 [40] or 300 ps for a 2 cm-path crossing a chip. Hence, the overall NoC essentially becomes a fast circuit-switched network, where the path-setup latency is much longer than the transmission latency. Still, path-setup latency is on the order of nanoseconds, a very short time compared to conventional circuit-switched networks where the typical setup time is in the millisecond range. Therefore, when packet sizes are fairly large, the setup time in our NoC can be considered fast and the network can handle packet-switched traffic with reasonable latency. The timing diagram in Fig. 6 illustrates the timing discrepancy.

For short messages, in a lightly loaded network, the largest latency component is the zero-load path-setup latency. If, conversely, the NoC is heavily loaded, the path-setup latency is contention dominated. Namely, most of the latency is the time spent by the path-setup packets in the routers' internal buffers. Path-setup packets are buffered when they are blocked by contention and are only released when the blocking message has been cleared. This approach, however, suffers from increased latency, especially for large message sizes. An alternative solution is to reduce the buffering depth in the router while relying on path multiplicity. In the extreme case, the buffer depths in the routers are reduced to zero, path-setup packets are dropped on contention, and the originating sources are immediately notified by a *packet-dropped* packet. The sources can then exploit the network's PM to attempt transmission on a different path and throughput can be improved significantly. Experimental results on a study comparing these methods are reported in Section 5.5.

The optimal size of a photonic message in a given implementation of the photonic NoC 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. Further, it is critical to account for the path-setup latency. While one would want to

minimize the setup time overhead by using large messages, their size should be kept small enough to allow for good flexibility and link utilization and to avoid excessive serialization latencies. Clearly, exchanges of small messages, such as memory read requests, write acknowledgments, and cache-coherency snoop messages, pose a challenge in terms of efficiency. When large memory pages or long cache lines are exchanged, instead, the photonic network is utilized much more efficiently. Our hybrid NoC microarchitecture represents a communication medium with unique performance features for the latter case (i.e., large messages) while making it possible to address the problem of exchanging small packets in an elegant way. In fact, small messages can be exchanged on the control network, which is essentially a low-bandwidth electronic NoC, while not requiring large resources in terms of additional circuitry or power dissipation. Long-lasting connections can be set up between processing cores that are expected to communicate frequently, thus providing a high-bandwidth link with minimal latency and low power consumption on which packets of any size can be transmitted.

Another favorable communication model for communications between processing cores is DMA, where large blocks of data are exchanged between memory modules with minimal CPU overhead. Considering the path-setup overhead in the photonic NoC, DMA can be configured to use memory transactions that are fairly large and are planned in advance. The DMA overhead messages can be transmitted over the control network while the optical path is being set up. Hence, some of the path-setup latency can overlap with the DMA overhead to reduce the total latency. In Section 5.3, we present experimental results for DMA communications on the proposed photonic NoC and look at optimal block sizes.

5 Design Analysis and Optimization

A key stage in the development of the ideas presented above is their functional validation using simulation. Further, a quantitative performance study, using a variety of traffic loads, must be carried out to evaluate alternative topologies, routine algorithms, and flow control techniques.

We developed a new event-driven simulator that is specifically tailored to provide support for the design exploration of the proposed photonic NoC. After describing the simulation setup, in the next sections, we report the results of several studies performed using the simulator. The first study is about avoiding deadlock, while the others explore various performance optimization techniques to increase PM, limit the path-setup procedure overhead, and determine the optimal message size.

5.1 Simulation Setup

We developed POINTS (Photonic On-chip Interconnection Network Traffic Simulator), an event-driven simulator based on OMNET++ [41]. 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.

We implemented a highly parameterized model, which enables a broad exploration of the design space, and we use it to analyze the case study of a 36-core 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 3.3×3.3 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—for clarity purposes), so it uses a matrix of 12×12 switches. The ERs, each located at the 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 optical signals [40] and 131 ps/mm in an optimally repeated wire at 22 nm for the electronic signals traveling between ERs [42]. The inter-PSE delay and interrouter delay are, therefore, 13 and 220 ps, respectively. The PSE setup time is assumed to be 1 ns and the router processing latency is 600 ps, or three cycle times of a 5 GHz 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 microarchitecture resembles circuit switching more than packet switching, we model the intermessage gap as an exponential random variable with a parameter μ_{IMG}. In the simulations 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.

5.2 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 [30]. *Virtual channel flow control* has been proven successful in eliminating intradimension deadlocks [43] and make dimension-order routing deadlock free. These results assume that each router in the torus network is internally nonblocking. As described in Section 4, 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. U-turns are forbidden. Hence, we are required to 1) evaluate the topology, 2) find when deadlocks may occur, and 3) develop solutions to avoid them. In Section 4, we explained the injection-ejection mechanisms that are 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 three switches comprising the access point and show that, by design, deadlock cannot occur in any of them:

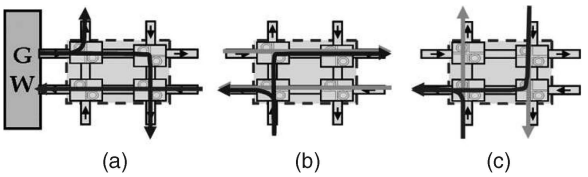


Fig. 7. (a) Gateway, (b) injection, and (c) ejection switches. All possible message-paths are marked to demonstrate that no blocking interactions occur.

- *Gateway switch.* Injected messages are required to make a turn toward the injection switches. Ejected messages arrive from the ejection message and pass straight through. Therefore, blocking cannot happen.
- *Injection switch.* Messages 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 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 possible paths marked on them. One could verify that none of the internal blocking scenarios (listed in Table 1) can occur.

Even though injection-ejection blocking situations are completely avoided and so are the associated performance penalty and possible deadlocks, the problem of intradimensional blocking of dimension-order routing still remains. The accepted solution for this problem is virtual channel flow control where the channel dependencies are removed by splitting the physical channel into several virtual channels that compete with each other for router bandwidth [43]. This approach is difficult to implement in a circuit-switched network where the channel bandwidth cannot be divided between several circuits.

One way to solve the intradimensional deadlock problem is to use path-setup timeouts. When a path-setup packet is sent, the gateway sets a timer to a predefined time. When the timer expires, a *terminate-on-timeout* packet is sent after 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. This allows the system to recover from a potential deadlock. While this method suffers from some inefficiency because paths and gateway injection ports are blocked for some time until they are terminated without transmission, it guarantees deadlock-recovery.

In an alternative method, 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 has not expired. This method has been tested in extensive simulations and has been shown to be effective in resolving deadlocks.

5.3 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 and serialization latency is increased. 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 core, 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 defined as 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). Obviously, $\rho \geq 1$ and the smaller the value of ρ, the higher the network efficiency.

We pick the message size by setting a desired overhead ratio and finding the smallest message size for which the selected ratio is not exceeded. In an unloaded network, if the maximum allowed overhead is 20 percent, then the maximum overhead ratio is $\rho = 1.25$. This limit is met by messages with duration larger than 50 ns for the longest path, which consists of 13 hops.

The next step is to simulate a loaded network when the global message duration is 50 ns. Naturally, the overhead will be larger when the network becomes loaded with traffic from other cores 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 cores with uniformly distributed addresses. The load on the network is regulated by controlling the distribution parameter of the exponentially distributed intermessage 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 intermessage gap is very large ($\frac{1}{\mu_{IMG}} \rightarrow \infty$), the offered load approaches zero. The results of the congestion experiment are reported in Fig. 8, showing 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

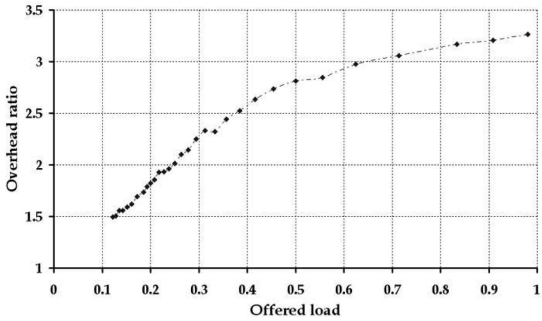


Fig. 8. Overhead ratio as a function of offered load for 50 ns messages in a 36-core photonic NoC (6×6 torus, no PM).

probability. One must also remember that the network is simulated under uniform traffic. Typical application in CMP environments is expected to generate more localized traffic patterns which can be routed more efficiently by the network.

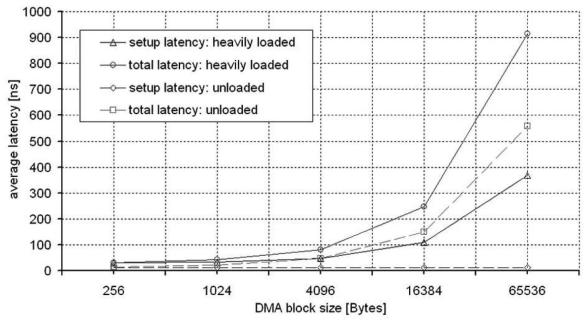
To allow better understanding of the effect of the message size on the network's performance and to confirm the choice of the overhead threshold (20 percent), we completed additional simulations using a more realistic DMA model. We note that, with a peak transmission rate of 960 Gbps (using WDM, see Section 6.2), 50 ns can be used to transmit a 6 Kbyte message, e.g., a DMA block. Accurate modeling of a DMA transaction requires the knowledge of the specific implementation of the DMA hardware [7]. However, interesting data points can be obtained by simulating the effects of the block size on the latency and on the average bandwidth of the photonic NoC for the cases of an unloaded network and a heavily loaded network $(\text{offered load} = 0.85)$. Using the POINTS simulator, we obtained the results reported in Fig. 9 for a peak transmission bandwidth equal to 960 Gbps.

As predicted, for small block sizes $(\leq 1\ \text{Kbyte})$, the overall latency is dominated by the path-setup overhead, which is greater than the serialization latency, because of the extremely large transmission bandwidth. DMA blocks of this size will clearly be inefficient. On the other hand, whenever very large blocks $(\geq 65\ \text{Kbytes})$ are used, the increased serialization and contention latencies overshadow the gain in bandwidth, which is diminishing for large blocks. Therefore, in the presence of this trade-off, the optimal DMA block size for the transactions over the photonic NoC ranges between 4 and 16 Kbytes. This result is consistent with the 50 ns result obtained earlier.

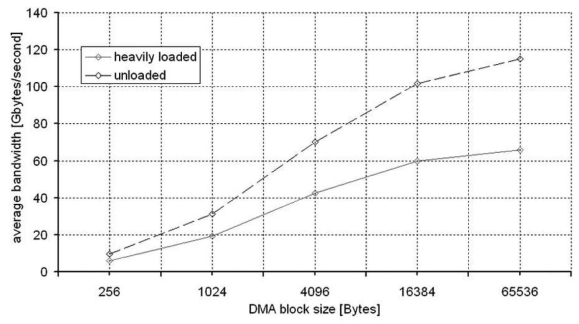
An alternative technique to reduce the path-setup latency by relieving contention is increasing the PM by augmenting the network with parallel lines. This approach is considered in the next section.

5.4 Increasing Path Multiplicity

One of the advantages of packet-switched 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



(a)



(b)

Fig. 9. (a) Average latency and (b) bandwidth for network transaction of different sizes in 36-core photonic NoC (12×12 torus).

TABLE 2
NoC Switch Counts as Function of Path Multiplicity Values

PM 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

algorithms that use current information on the network load. As discussed in Section 4, the topology chosen for the proposed network, a torus, can be easily augmented with additional parallel paths that provide path multiplicity 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.

Like in 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. Networks with path multiplicity 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, PM presents an overhead 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 5,000 μm^2, then, theoretically, more than 80,000 such switches can be integrated in the photonic layer of a $400\ \text{mm}^2$ die. The power dissipated by the diversified network scales sublinearly with the number of switches as switches only consume power when they cause

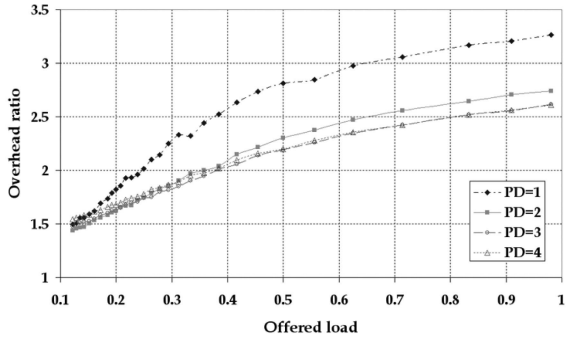


Fig. 10. Overhead ratio versus offered load for varying values of path multiplicity (PM). The corresponding network sizes are given in Table 2.

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. A more detailed power analysis is given in Section 6.

The simulation results are reported 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, dramatically reduce the latency, thus contributing to a more efficient network.

Second, path multiplicity clearly has a diminishing return beyond a $\times 3$ factor. When the path multiplicity is too large, the additional paths lead to a large network diameter and zero-load latency, while making only a minor reduction in network blocking. In fact, for the simulated test case, it can be argued that the performance gain achieved by a $\times 3$ path multiplicity is too small to justify the increase in the optical layer density and fabrication complexity. In any case, a $\times 2$ path multiplicity increase does offer dramatic performance gains. Clearly, for any system, these design issues should be modeled, studied, and evaluated carefully against physical design constraints (e.g., waveguide routing space limits) in the preliminary design and microarchitecture phases.

5.5 Evaluating Path-Setup Procedures

In Section 5.3, we showed how the network throughput and performance is determined by the path-setup overhead. Reductions in path-setup latency translate to improved efficiency of the network interfaces and to higher average bandwidth.

For a given source-destination pair, the setup latency can be expressed as $D = (H-1)\cdot t_p + t_q$, where H is the number of hops in the packet's path, t_p is the processing latency in each router, and t_q is the total additional latency due to contentions. As discussed in Section 4.4, contention in the path-setup phase can be handled by blocking the path-setup packet until the path is cleared. Simulations show that t_q is a major contributor to the overall setup latency, especially when the network is heavily loaded. Notice, however, that the actual processing latency in the path-setup phase, which is equal to $(H-1)\cdot t_p$, is typically much lower than the contention-based latency. Hence, in order to decrease the contention-based setup latency t_q, one can use an alternative method that consists of immediately dropping any path-setup packet that is blocked instead of buffering it. This allows reducing the buffering depth in the ER down to zero, thus simplifying its circuitry. On the other hand, it requires that a *packet-dropped* packet be sent on the control network in the opposite direction to notify the sender. Then, the sender can immediately attempt to set up an alternative path, exploiting the network's path multiplicity. With an adequate level of path multiplicity, it is reasonable to assume that an alternative path can be found faster than it would take for the message obstructing the original path to be torn down.

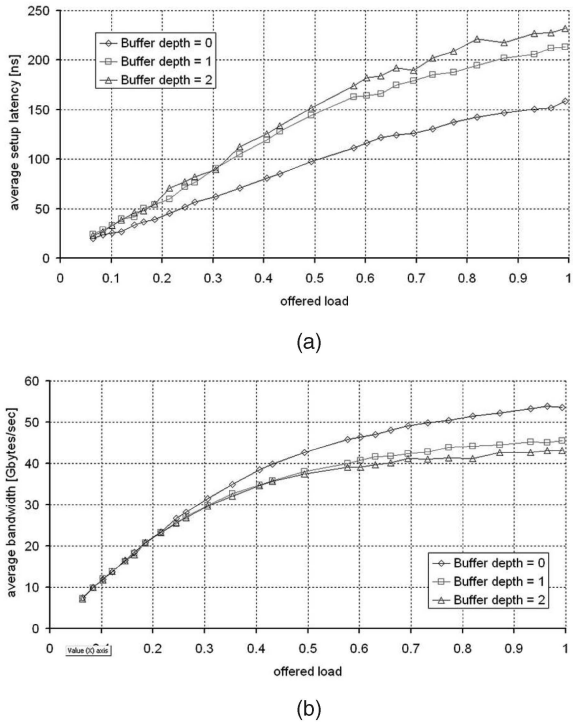


Fig. 11. (a) Average path-setup latency and (b) bandwidth as a function of buffer depth in a 36-core photonic NoC (12×12 torus).

We used the POINTS simulator to evaluate this idea in the case of the same 36-core CMP system discussed above and assuming an optical message size equal to 16 Kbytes and a path multiplicity factor equal to $\times 2$. The simulation results are reported in Fig. 11: By setting the buffer depth to 0, i.e., dropping every blocked packet and immediately notifying the sender, the path-setup latency can be reduced by as much as 30 percent when compared to the case where path-setup packets are not dropped on contention (buffer depth of 2). When a buffer depth of 1 is simulated, i.e., a single path-setup packet may be queued in each direction in each electronic router, the latency reduction is smaller.

The peak optical bandwidth per port in the simulations, using WDM, is set at 960 Gbps. The average bandwidth is calculated as the product of the peak bandwidth and the fractional time, in steady state, that can be allocated for actual transmission of the optical messages, after messages have been set up. The average bandwidth results are also shown in Fig. 11, demonstrating a maximum sustained

TABLE 3
Predictions for Future Technology Nodes

	65 nm	45 nm	32 nm
Clock Frequency [GHz]	3.2	4	5
E_{LINK} [pJ/mm/bit]	0.58	0.46	0.34
E_{BUFFER} [pJ/bit]	0.16	0.13	0.12
$E_{CROSSBAR}$ [pJ/bit]	0.93	0.63	0.36
E_{STATIC} [pJ/bit]	0.06	0.11	0.35

TABLE 4
Power Consumption of Electronic NoC

	65 nm	45 nm	32 nm
Flit width	256	208	168
Link length [mm]	3.33	2.33	1.67
$E_{FLIT-HOP}$ [pJ]	788	406	235
P_{E-NoC} **[W]**	**227**	**146**	**106**

bandwidth, or throughput, of approximately 53 Gbytes/s. This result, 45 percent of the peak possible bandwidth, is considered quite good for an interconnection network and can even be improved with better routing algorithms and when more realistic and localized traffic patterns are applied onto the network.

6 Comparative Power Analysis

The main motivation for the design of a photonic NoC is the potential dramatic reduction in the power dissipated on high-bandwidth communications. To evaluate this power reduction, we perform a comparative high-level power analysis between two equivalent on-chip interconnection networks for CMPs: a photonic NoC and a reference electronic NoC. They are equivalent in the sense that both networks must provide the same bandwidth to the same number of processing cores. For our case study, we assume a CMP implemented in a future 22 nm CMOS technology and hosting 36 processing cores, each requiring a peak bandwidth of 800 Gbps and an average bandwidth of 512 Gbps. These numbers match widely accepted predictions on future on-chip bandwidth requirements in high-performance CMPs. We will see that, in this high-bandwidth realm, photonic technologies can offer a significant reduction in the interconnect power. We assume a uniform traffic model, a mesh topology, and XY dimension-order routing. Of course, different conditions can be used, but, as our goal is to provide an equal comparison plane, this choice provides a simple "apples-to-apples" comparison.

6.1 Reference Electronic NoC

The reference electronic network is a 6×6 mesh, where each router is integrated in one processor tile and is connected to four (or fewer) neighboring tiles. A router microarchitecture that is based on an input-queued crossbar with a 4-flit buffer[2] on each input port has been widely proposed in the NoC literature [10], [39]. The router has five I/O ports: one for the local processor and one for each of the four network connections with a neighbor tile (N, S, E, and W). We estimate the power expended in an electronic NoC under a given load using the method developed by Eisley and Peh in [44]: This assumes that, whenever a flit traverses a link and the subsequent router, five operations are performed:

1. reading from a buffer,
2. traversing the routers' internal crossbar,
3. transmission across the interrouter link,
4. writing to a buffer in the subsequent router, and
5. triggering an arbitration decision.

The energy required for a single hop through a link and a router ($E_{FLIT-HOP}$) is the sum of the energies spent in these operations. Table 3 reports the values of the energy spent in these operations (buffer reading and writing energies are combined, arbiter energy is neglected) that were obtained with the ORION NoC simulator [45]. ORION accounts for the static energy dissipated in the router and converts it to a per-bit scale. $E_{FLIT-HOP}$, the energy expended to transmit one flit across a link and a subsequent router, is computed based on the energy estimates in Table 3 as well as the link length and flit-width, which vary for different technology nodes. The total energy expended in a clock cycle can be computed as

$$E_{NETWORK-CYCLE} = \sum_{j=1}^{N_L} U_{Lj} \cdot E_{FLIT-HOP},$$

where U_{Lj} is the average number of flits traversing link j per clock cycle, an estimate on the utilization of link j. Then, the power dissipated in the network is equal to

$$P_N = E_{NETWORK-CYCLE} \cdot f,$$

where f is the clock frequency. For a 6×6 mesh under uniform traffic using XY routing and an injection rate of $\alpha = 0.625$, the global average link utilization is $\bar{U} = 0.75$. Hence, the energy expended in a clock cycle in the reference electronic NoC (which has 120 links) is

$$E_{NETWORK-CYCLE} = 0.75 \cdot 120 \cdot E_{FLIT-HOP}$$

and the total power dissipated is estimated as

$$P_{E-NoC} = E_{NETWORK-CYCLE} \cdot f.$$

The results are given in Table 4. The main conclusion that can be drawn from this analysis is that, when a truly high communication bandwidth is required for on-chip data exchange, even a dedicated, carefully designed NoC may not be able to provide it within reasonable power constraints. Since the electronic transmission is limited in bandwidth to a few gigahertz at most, high transmission capacity requires the use of many parallel lines and wide buffers [10], which lead to high power dissipation for transmission and buffering. Admittedly, the above analysis is based on a simple circuit implementation, but, even if aggressive electronic circuit techniques such as low-swing current mode signaling are employed, the overall NoC power consumption that is necessary to meet the communication bandwidth requirements in future CMPs will likely be too high to manage within the tight packaging constraints [1].

2. A flit is the minimal flow control unit, equal to the number of bits that cross the link in a clock cycle, i.e., the link width.

6.2 Proposed Photonic NoC

Since our NoC is based on a hybrid microarchitecture, its power dissipation can be estimated as the sum of three components: 1) the photonic data-transfer network, 2) the electronic control network, and 3) the O/E and E/O interfaces.

6.2.1 Photonic Data-Transmission Network

Path multiplicity is a low-power cost-effective solution to compensate for the lack of buffers in the photonic network. In this design, we assume a path multiplicity factor of 2, meaning a 12×12 photonic mesh, comprised of 576 PSEs (144 4×4 switches), serves the 6×6 CMP. The power analysis of a photonic NoC is fundamentally different from the electronic network analysis since it mainly depends on the state of the PSEs: In the *ON* state, when the multi-wavelength message is forced to turn, the power dissipated is approximately 10 mW [15], [18], while there is no dissipation in the *OFF* state when a message proceeds undisturbed or when no message is forwarded.

Hence, the total power consumption in the network depends on the number of switches in the *ON* state, which can be estimated based on network statistics and traffic dynamics. We assume that, in the photonic NoC, each message makes, at most, four turns, based on the 3-stage routing algorithm described in Section 4.3. In the photonic network, we assume a peak bandwidth of 960 Gbps, exceeding the 800 Gbps requirement, and an injection rate of 0.6, so the average bandwidth is 576 Gbps. The average number of messages in the network at any given time is calculated as $36 \times 0.6 = 21.6$. The average number of PSEs in the *ON* state is about 86 in a 576-PSE NoC. Hence, the total power consumption is estimated as

$$P_{P-NoC,transmission} = 86 \cdot 10 \text{ mW} = 860 \text{ mW},$$

dramatically lower than anything that can be approached by an electronic NoC.

6.2.2 Electronic Control Network

The power analysis of the electronic control network is based on the fact that this is essentially an electronic packet-switched NoC, i.e., similar to the reference electronic NoC that we discussed in Section 6.1 except for the larger dimensions (12×12 compared to 6×6). We assume that each photonic message is accompanied by two 32-bit control packets and the typical size of a message is 2 Kbytes. Then, the total power consumed by the electronic control network can be approximated as

$$P_{P-NoC,control} = P_{E-NoC} \cdot 2 \cdot \frac{32}{16,384} \cdot 2 = 0.82 \text{ W}.$$

If the electronic control network is utilized lightly, the impact of static power becomes more dominant in the overall NoC power budget. However, recent technological breakthroughs in semiconductor processes, namely, Intel's 45 nm process leveraging high-K dielectrics, have been shown to reduce the gate leakage more than 10-fold [46]. Having dramatically reduced the gate leakage, channel leakage remains the major challenge. Given past trends in semiconductor technology, it is reasonable to expect that a solution will be found.

6.2.3 Network Interfaces

To generate the 960 Gbps peak bandwidth, we assume a modulation rate of 40 Gbps on 24 wavelengths, as was demonstrated in [34]. The modulated data streams are grouped using passive WDM multiplexers, so power is dissipated mainly in the 24 modulators and 24 receiver circuits in each gateway. Since there is presently no equivalent to the International Technology Roadmap for Semiconductors (ITRS) [1] for the photonic technology, predictions on the power consumption of photonic elements vary greatly. A reasonable estimate for the energy dissipated by a modulator/detector pair, at 10 Gbps, today is about 2 pJ/bit, based on recent results reported by IBM [16]. We estimate that, using silicon ring-resonator modulators and SiGe photodetectors, the energy will decrease to about 0.2 pJ/bit in the next 8-10 years. Consequently, the total power dissipated by 36 interfaces under the conditions described above is

$$P_{P-NoC,gateways} = 0.2 \text{ pJ/bit} \times 36 \times 576 \text{ Gbps} = 4.2 \text{ W}.$$

Supplementary circuits that are usually required for the implementation of optical receivers (e.g., clock-data recovery, serializer/deserializer, and dispersion compensation), are not needed in an ultrashort link in which the modulation rate is equal to the chip clock rate [13]. As most of the power consumed by optical receivers is usually due to these circuits [24], the power saving potential is large. The off-chip laser sources consume an estimated power of 10 mW per wavelength. Although a large number of lasers are required to exploit the bandwidth potential of the optical NoC, their power is dissipated off-chip and does not contribute to the chip power density.

Putting things together, the estimated power consumed by the photonic NoC to exchange data between 36 cores at an average bandwidth of 576 Gbps is given by the sum of the three components and is equal to ~ 6 W. Although the power analysis used here is rather simplistic and uses many assumptions to ease the calculation and work around missing data, its broader conclusion is clear. The potential power difference between photonics-based NoCs and their purely electronic counterparts is significant. Importantly, once generated, the power consumed by propagating the optical signals off-chip across the system is essentially negligible and can enable true scaling for off-chip CMP high-bandwidth communications. Even when one accounts for inaccuracies in our analysis and considers predicted future trends, the advantages offered by photonics represent a clear leap in terms of *bandwidth-per-watt* performance.

7 Concluding Remarks

We have proposed the idea of building a photonic NoC for future generations of CMPs. The motivation behind our work is rooted in the intersection of several technological trajectories from different fields. First, multicore processors step into an era where high-bandwidth communications between large numbers of cores is a key driver of computing performance. Second, power dissipation has clearly become the limiting factor in the design of high-performance microprocessors.

Moreover, the power dissipated on intrachip communication is a large and growing fraction of the total power budget. Third, recent breakthroughs in the field of silicon photonics suggest that the integration of optical elements with CMOS electronics is likely to become viable in the near future.

The intersection of these three factors suggests that silicon photonic technology can be used to construct NoCs, offering a promising low-power solution for high-performance on-chip communication. The design of photonic NoC presents interesting and challenging problems. To address these problems, we proposed a hybrid NoC microarchitecture that combines a photonic circuit-switched network with an electronic packet-switched network so that each technology is used advantageously: photonics for bulk-data transmission and electronics for network control. Electronic packets are used to establish transparent lightpaths that carry high-bandwidth optical messages across a network of broadband optical switches.

The proposed microarchitecture has been analyzed and optimized through extensive simulations: a torus topology, augmented with multiple paths and gateway access points, has been shown to provide large average transmission bandwidth and low latency while avoiding injection-triggered deadlocks. Techniques of recovering from inter-dimension deadlocks were also suggested. Several performance-related parameters were modeled and simulated and their effects on the bandwidths and the message latency were quantified. A power analysis was conducted, demonstrating the potential power reduction of the proposed design over traditional NoCs. When very large bandwidths are required, the power dissipated on intrachip communication can easily exceed 100 W in regular electronic NoCs. The proposed photonic NoC can potentially reduce this figure to a few watts.

From the photonic NoC design and optimization viewpoint, there is still much work that can be done. The POINTS simulator can be used to further explore the vast photonic NoC design space and evaluate modification to the flow control, topology, switch design, and routing algorithms.

Most importantly, real computing applications need to be mapped on the network model to generate traffic patterns that accurately represent real cases. Scientific benchmarks or real applications should be used to validate the network design and assist in exploring routing algorithms, topology, flow control, and other design decisions. Much of this is ongoing work as we plan to address many of these challenges in follow-up publications.

The technology required to implement the photonic devices (PSEs and 4×4 switches) and their integration in large-scale NoCs is still immature. We have carefully reviewed the recent major progress made in both academia and industry and we expect that, within a small number of years, the enabling technologies will gradually become available to the designers of silicon-integrated circuits. Detailed study of other design issues such as process integration, design complexity, and area overhead is also important to evaluate the feasibility of this project and is an interesting area for future research. Based on our continuous interactions with researchers working on silicon photonic devices, we believe that, to have a system-level perspective on how they can be composed in a future NoC is of critical importance to their design. This paper aims at laying the groundwork for future research progress by providing a complete discussion of the fundamental issues that need to be addressed to design a photonic NoC for CMPs.

Acknowledgments

A. Shacham and K. Bergman acknowledge the support of the US National Science Foundation (NSF) under Grant CCF-0523771 and the US Department of Defense under subcontract B-12-664. L.P. Carloni acknowledges the support of the NSF under Grant 0541278.

References

[1] ITRS, "The International Technology Roadmap for Semiconductors—2006 Edition," http://public.itrs.net, 2006.
[2] R. Kalla, B. Sinharoy, and J.M. Tendler, "IBM Power5 Chip: A Dual-Core Multithreaded Processor," *IEEE Micro,* vol. 24, no. 2, pp. 40-47, Mar./Apr. 2004.
[3] P. Kongetira, K. Aingaran, and K. Olukotun, "Niagara: A 32-Way Multithreaded SPARC Processor," *IEEE Micro,* vol. 25, no. 2, pp. 21-29, Mar./Apr. 2005.
[4] S. Naffziger, B. Stackhouse, and T. Grutkowski, "The Implementation of a 2-Core Multi-Threaded Itanium-Family Processor," *Proc. IEEE Int'l Solid State Circuits Conf.,* pp. 182-183, Feb. 2005.
[5] A. Kahle, M.N. Day, H.P. Hofsteee, C.R. Johns, T.R. Maeurer, and D. Shippy, "Introduction to the CELL Multiprocessor," *IBM J. Research and Development,* vol. 49, nos. 4/5, pp. 589-604, Sept. 2005.
[6] D. Pham et al., "The Design and Implementation of a First-Generation CELL Processor," *Proc. IEEE Int'l Solid State Circuits Conf.,* pp. 184-185, Feb. 2005.
[7] M. Kistler, M. Perrone, and F. Petrini, "Cell Multiprocessor Communication Network: Built for Speed," *IEEE Micro,* vol. 26, no. 3, pp. 10-23, May/June 2006.
[8] S. Vangal et al., "An 80-Tile 1.28 TFLOPS Network-on-Chip in 65 nm CMOS," *Proc. IEEE Int'l Solid State Circuits Conf.,* paper 5.2, Feb. 2007.
[9] A. Hemani, A. Jantsch, S. Kumar, A. Postula, J. Oberg, M. Millberg, and D. Lindqvist, "Network on Chip: An Architecture for Billion Transistor Era," *Proc. 18th IEEE NorChip Conf.,* Nov. 2000.
[10] W.J. Dally and B. Towles, "Route Packets, Not Wires: On-Chip Interconnection Networks," *Proc. 38th Design Automation Conf.,* pp. 684-689, June 2001.
[11] L. Benini and G. De-Micheli, "Networks on Chip: A New SoC Paradigm," *Computer,* vol. 49, no. 1, pp. 70-71, Jan. 2002.
[12] R. Kumar, V. Zyuban, and D.M. Tullsen, "Interconnections in Multi-Core Architectures: Understanding Mechanism, Overheads, Scaling," *Proc. 32nd Ann. Int'l Symp. Computer Architecture,* June 2005.
[13] J.D. Owens, W.J. Dally, R. Ho, D.J. Jayasimha, S.W. Keckler, and L.-S. Peh, "Research Challenges for On-Chip Interconnection Networks," *IEEE Micro,* vol. 27, no. 5, pp. 96-108, Sept./Oct. 2007.
[14] C. Gunn, "CMOS Photonics for High-Speed Interconnects," *IEEE Micro,* vol. 26, no. 2, pp. 58-66, Mar./Apr. 2006.
[15] F. Xia, M.J. Rooks, L. Sekaric, and Y.A. Vlasov, "Ultra-Compact High Order Ring Resonator Filters Using Submicron Silicon Photonic Wires for On-Chip Optical Interconnects," *Optics Express,* vol. 15, no. 19, pp. 11934-11941, Sept. 2007.
[16] W.M.J. Green, M.J. Rooks, L. Sekaric, and Y.A. Vlasov, "Ultra-Compact, Low RF Power, 10 Gb/s Silicon Mach-Zehnder Modulator," *Optics Express,* vol. 15, no. 25, pp. 17106-17113, Dec. 2007.
[17] C.L. Schow, F. Doany, O. Liboiron-Ladouceur, C. Baks, D.M. Kuchta, L. Schares, R. John, and J.A. Kash, "160-Gb/s, 16-Channel Full-Duplex, Single-Chip CMOS Optical Transceiver," *Proc. Optical Fiber Comm. Conf.,* paper OThG4, Mar. 2007.

[18] A. Biberman, B.G. Lee, K. Bergman, P. Dong, and M. Lipson, "Demonstration of All-Optical Multi-Wavelength Message Routing for Silicon Photonic Networks," *Proc. Optical Fiber Comm. Conf.*, paper OTuF6, Mar. 2008.
[19] Y.A. Vlasov, W.M.J. Green, and F. Xia, "High-Throughput Silicon Nanophotonic Wavelength-Insensitive Switch for On-Chip Optical Networks," *Nature Photonics*, vol. 2, no. 4, Apr. 2008.
[20] K. Bernstein et al., "Interconnects in the Third Dimension: Design Challenges for 3D ICs," *Proc. 44th Design Automation Conf.*, pp. 562-567, June 2007.
[21] A. Shacham, K. Bergman, and L.P. Carloni, "Maximizing GFLOPS-per-Watt: High-Bandwidth, Low Power Photonic On-Chip Networks," *Proc. Third Watson Conf. Interaction between Architecture, Circuits, and Compilers*, pp. 12-21, Oct. 2006.
[22] A. Shacham, K. Bergman, and L.P. Carloni, "The Case for Low-Power Photonic Networks on Chip," *Proc. 44th Design Automation Conf.*, pp. 132-135, June 2007.
[23] A. Shacham, K. Bergman, and L.P. Carloni, "On the Design of a Photonic Network on Chip," *Proc. First IEEE Int'l Symp. Networks-on-Chips*, pp. 53-64, May 2007.
[24] R. Ramaswami and K.N. Sivarajan, *Optical Networks: A Practical Perspective*, second ed. Morgan Kaufmann, 2002.
[25] J.H. Collet, F. Caignet, F. Sellaye, and D. Litaize, "Performance Constraints for Onchip Optical Interconnects," *IEEE J. Selected Topics in Quantum Electronics*, vol. 9, no. 2, pp. 425-432, Mar./Apr. 2003.
[26] N. Kirman, M. Kirman, R.K. Dokania, J. Martínez, A.B. Apsel, M.A. Watkins, and D.H. Albonesi, "On-Chip Optical Technology in Future Bus-Based Multicore Designs," *IEEE Micro*, vol. 27, no. 1, pp. 56-66, Jan./Feb. 2007.
[27] M. Briére, B. Girodias, Y. Bouchebaba, G. Nicolescu, F. Mieyeville, F. Gaffiot, and I. OConnor, "System Level Assessment of an Optical NoC in an MPSoC Platform," *Proc. Design, Automation and Test in Europe*, Mar. 2007.
[28] M.J. Kobrinsky et al., "On-Chip Optical Interconnects," *Intel Technology J.*, vol. 8, no. 2, pp. 129-142, May 2004.
[29] T. Mudge, "Power: A First-Class Architectural Design Constraint," *Computer*, vol. 34, no. 4, pp. 52-58, Apr. 2001.
[30] W.J. Dally and B. Towles, *Principles and Practices of Interconnection Networks*. Morgan Kaufmann, 2004.
[31] Q. Xu, S. Manipatruni, B. Schmidt, J. Shakya, and M. Lipson, "12.5 Gbit/s Carrier-Injection-Based Silicon Microring Silicon Modulators," *Optics Express*, vol. 15, no. 2, pp. 430-436, Jan. 2007.
[32] A. Gupta, S.P. Levitan, L. Selavo, and D.M. Chiarulli, "High-Speed Optoelectronics Receivers in SiGe," *Proc. 17th Int'l Conf. VLSI Design*, pp. 957-960, Jan. 2004.
[33] A.W. Fang, H. Park, O. Cohen, R. Jones, M.J. Paniccia, and J.E. Bowers, "Electrically Pumped Hybrid AlGaInAs-Silicon Evanescent Laser," *Optics Express*, vol. 14, no. 20, pp. 9203-9210, Oct. 2006.
[34] B.G. Lee et al., "Ultrahigh-Bandwidth Silicon Photonic Nanowire Waveguides for On-Chip Networks," *IEEE Photonics Technology Letters*, vol. 20, no. 6, pp. 398-400, Mar. 2008.
[35] B.G. Lee, B.A. Small, Q. Xu, M. Lipson, and K. Bergman, "Characterization of a 4×4 Gb/s Parallel Electronic Bus to WDM Optical Link Silicon Photonic Translator," *IEEE Photonics Technology Letters*, vol. 19, no. 7, pp. 456-458, Apr. 2007.
[36] Q. Xu, B. Schmidt, J. Shakya, and M. Lipson, "Cascaded Silicon Micro-Ring Modulators for WDM Optical Interconnection," *Optics Express*, vol. 14, no. 20, pp. 9430-9435, Oct. 2006.
[37] A. Shacham and K. Bergman, "Building Ultralow Latency Interconnection Networks Using Photonic Integration," *IEEE Micro*, vol. 27, no. 4, pp. 6-20, July/Aug. 2007.
[38] F. Xia, L. Sekaric, and Y.A. Vlasov, "Ultracompact Optical Buffers on a Silicon Chip," *Nature Photonics*, vol. 1, no. 1, pp. 65-71, Jan. 2007.
[39] T.M. Pinkston and J. Shin, "Trends toward On-Chip Networked Microsystems," *Int'l J. High Performance Computing and Networking*, vol. 3, no. 1, pp. 3-18, 2001.
[40] I.-W. Hsieh, X. Chen, J.I. Dadap, N.C. Panoiu, J.R.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*, vol. 14, no. 25, pp. 12380-12387, Dec. 2006.
[41] "OMNeT++ Discrete Event Simulation System," http://www.omnetpp.org/, 2008.
[42] R. Ho, "Wire Scaling and Trends," *MTO DARPA Meeting*, Sun Microsystems Laboratories, Aug. 2006.
[43] W.J. Dally and C.L. Seitz, "Deadlock-Free Message Routing in Multiprocessor Interconnection Networks," *IEEE Trans. Computers*, vol. 36, no. 5, pp. 547-553, May 1987.
[44] N. Eisley and L.-S. Peh, "High-Level Power Analysis for On-Chip Networks," *Proc. Int'l Conf. Compilers, Architecture, and Synthesis for Embedded Systems*, Sept. 2004.
[45] H.-S. Wang, X. Zhu, L.-S. Peh, and S. Malik, "Orion: A Power-Performance Simulator for Interconnection Networks," *Proc. 35th Ann. IEEE/ACM Int'l Symp. Microarchitecture*, Nov. 2002.
[46] "Under the Hood: Intel's 45-nm High-k Metal-Gate Process," http://www.eetimes.com/showArticle.jhtml?articleID=202806020, Nov. 2007.

Assaf Shacham received the BSc degree (cum laude) in computer engineering from the Technion, Israel Institute of Technology, in 2002 and the MS and PhD degrees in electrical engineering from Columbia University, New York, in 2004 and 2007, respectively. His doctoral dissertation was titled "Architectures of Optical Interconnection Networks for High Performance Computing." He has authored and coauthored more than 25 papers in peer-reviewed journals and major international conferences and has four patents pending. He is currently employed by Aprius Inc., where he is engaged in developing high-performance computer interconnection systems. He is a member of the IEEE and the IEEE Computer Society.

Keren Bergman received the BS degree from Bucknell University in 1988 and the MS and PhD degrees from the Massachusetts Institute of Technology in 1991 and 1994, respectively, all in electrical engineering. She is a professor of electrical engineering at Columbia University, where she also directs the Lightwave Research Laboratory. At Columbia, she leads multiple research projects in optical packet-switched networks, distributed grid computing over optical networks, photonic interconnection networks, nanophotonic networks-on-chip, and the applications of optical networking in high-performance computing systems. She is a recipient of the US National Science Foundation CAREER award in 1995 and the US Office of Naval Research Young Investigator in 1996. In 1997, she received the CalTech President's Award for joint work with JPL on optical packet networks. She is currently an associate editor for *IEEE Photonic Technology Letters* and the editor-in-chief for the *OSA Journal of Optical Networking*. She is a senior member of the IEEE, a member of the IEEE Computer Society, and a fellow of the Optical Society of America.

Luca P. Carloni received the Laurea degree (summa cum laude) in electrical engineering from the Università di Bologna, Bologna, Italy, in 1995 and the MS and PhD degrees in electrical engineering and computer sciences from the University of California, Berkeley, in 1997 and 2004, respectively. He is currently an assistant professor in the Department of Computer Science at Columbia University, New York. He has authored more than 50 publications and is the holder of one patent. His research interests include the area of design tools and methodologies for integrated circuits and systems, distributed embedded systems design, and design of high-performance computer systems. He received the Faculty Early Career Development (CAREER) Award from the US National Science Foundation in 2006 and was selected as an Alfred P. Sloan Research Fellow in 2008. He is the recipient of the 2002 Demetri Angelakos Memorial Achievement Award "in recognition of altruistic attitude towards fellow graduate students." In 2002, one of his papers was selected for "The Best of ICCAD: A collection of the best IEEE International Conference on Computer-Aided Design Papers of the past 20 years." He is a member of the IEEE and the IEEE Computer Society.