Silicon-photonic clos networks for global on-chip communication

Citation: Joshi, A. et al. "Silicon-photonic clos networks for global on-chip communication." Networks-on-Chip, 2009. NoCS 2009. 3rd ACM/IEEE International Symposium on. 2009. 124-133. © 2009 IEEE

As Published: http://dx.doi.org/10.1109/NOCS.2009.5071460

Publisher: Institute of Electrical and Electronics Engineers

Persistent URL: http://hdl.handle.net/1721.1/54685

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Silicon-Photonic Clos Networks for Global On-Chip Communication

Ajay Joshi[*], Christopher Batten[*], Yong-Jin Kwon[†], Scott Beamer[†], Imran Shamim[*]
Krste Asanović[†], Vladimir Stojanović[*]

[*] *Department of EECS, Massachusetts Institute of Technology, Cambridge, MA*
[†] *Department of EECS, University of California, Berkeley, CA*

Abstract

Future manycore processors will require energy-efficient, high-throughput on-chip networks. Silicon-photonics is a promising new interconnect technology which offers lower power, higher bandwidth density, and shorter latencies than electrical interconnects. In this paper we explore using photonics to implement low-diameter non-blocking crossbar and Clos networks. We use analytical modeling to show that a 64-tile photonic Clos network consumes significantly less optical power, thermal tuning power, and area compared to global photonic crossbars over a range of photonic device parameters. Compared to various electrical on-chip networks, our simulation results indicate that a photonic Clos network can provide more uniform latency and throughput across a range of traffic patterns while consuming less power. These properties will help simplify parallel programming by allowing the programmer to ignore network topology during optimization.

1. Introduction

Today's graphics, network, embedded and server processors already contain many processor cores on one chip and this number is expected to increase with future scaling. The on-chip communication network is becoming a critical component, affecting not only performance and power consumption, but also programmer productivity. From a software perspective, an ideal network would have uniformly low latency and uniformly high bandwidth. The electrical on-chip networks used in today's multicore systems (e.g., crossbars [8], meshes [3], and rings [11]) will either be difficult to scale to higher core counts with reasonable power and area overheads or introduce significant bandwidth and latency non-uniformities. In this paper we explore the use of silicon-photonic technology to build on-chip networks that scale well, and provide uniformly low latency and uniformly high bandwidth.

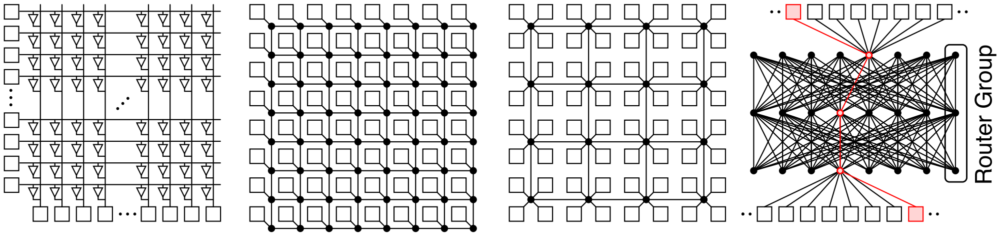
Various photonic materials and integration approaches have been proposed to enable efficient global on-chip communication, and several network architectures (e.g., crossbars [7, 15] and meshes [13]) have been developed bottom-up using fixed device technology parameters as drivers. In this paper, we take a top-down approach by driving the photonic device requirements based on the projected network and system needs. This allows quick design-space exploration at the network level, and provides insight into which network topologies can best harness the advantages of photonics at different stages of the technology roadmap.

This paper begins by identifying our target system and briefly reviewing the electrical on-chip networks which will serve as a baseline for our photonic network proposals. We then use analytical models to investigate the tradeoffs between various implementations of global photonic crossbars found in the literature and our own implementations of photonic Clos networks. We also use simulations to compare the photonic Clos network to electrical mesh and Clos networks. Our results show that photonic Clos networks consume significantly less optical laser power, thermal tuning power, and area as compared to photonic crossbar networks, and offer better energy-efficiency than electrical networks while providing more uniform performance across various traffic patterns.

2. Target System

Silicon-photonic technology for on-chip communication is still in its formative stages, but with recent technology advances we project that photonics might be viable in the late 2010's. This makes the 22 nm node a reasonable target process technology for our work. By then it will be possible to integrate hundreds of cores onto a single die. To simplify design and verification complexity, these cores and/or memory will most likely be clustered into tiles which are then replicated across the chip and interconnected with a well-structured on-chip network. The exact nature of the tiles and the inter-tile communication paradigm are still active areas of research. The tiles might be homogeneous with each tile including both some number of cores and a slice of the on-chip memory, or the tiles might be heterogeneous with a mix of compute and memory tiles. The global on-chip network might be used to implement shared memory, message passing, or both. Regardless of their exact configuration, however, all future systems will require some form of on-chip network which provides low-latency and high-throughput commu-

978-1-4244-4143-3/09/$25.00 ©2009 IEEE



(a) Crossbar (b) Mesh (c) CMesh (d) Clos

Figure 1: Logical View of 64 Tile Network Topologies – (a) 64x64 distributed tristate global crossbar, (b) 2D 8x8 mesh, (c) concentrated mesh (cmesh) with 4x concentration, (d) 8-ary, 3-stage Clos network with eight middle routers. In all four figures: squares = tiles, dots = routers, triangles = tristate buffers. In (b) and (c) inter-dot lines = two opposite direction channels. In (a) and (d) inter-dot lines = uni-directional channels.

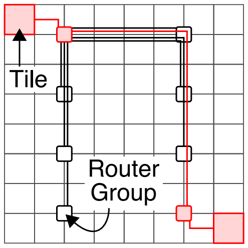


Figure 2: Clos Layout – Router group is three routers. Only a subset of the channels are shown.

nication at low energy and small area.

For this paper we assume a target system with 64 square tiles operating at 5 GHz on a 400 mm^2 chip. Figure 1 illustrates some of the topologies available for implementing on-chip networks. They range from high-radix, low-diameter crossbar networks to low-radix, high-diameter mesh networks. We examine networks sized for low (LTBw), medium (MTBw), and high (HTBw) bandwidth which correspond to ideal throughputs of 64, 128, and 256 b/cycle per tile under uniform random traffic. Although we primarily focus on a single on-chip network, our exploration approach is also applicable to future systems with multiple physical networks.

3. Electrical On-Chip Networks

In this section, we explore the qualitative trade-offs between various network architectures that use traditional electrical interconnect. This will provide an electrical baseline for comparison, and also yield insight into the best way to leverage silicon photonics.

3.1. Electrical Technology

The performance and cost of on-chip networks depend heavily on various technology parameters. For this work we use the 22 nm predictive technology models [16] and interconnect projections from [6] and the ITRS.

All of our inter-router channels are implemented in semi-global metal layers with standard repeated wires. For medium length wires (2–3 mm or approximately the width of a tile) the repeater sizing and spacing are chosen so as to minimize the energy for the target cycle-time. Longer wires are energy optimized as well as pipelined to maintain throughput. The average energy to transmit a bit transition over a distance of 2.5 mm in 200 ps is roughly 160 fJ, while the fixed link cost due to leakage and clocking is ≈20 fJ per cycle. The wire pitch is only 500 nm, which means that ten thousand wires can be supported across the bisection of our target chip even with extra space for power distribution and vias. Given the abundance of on-chip wiring resources, interconnect power dissipation will likely be a more serious constraint than bisection bandwidth for most network topologies.

We assume a relatively simple router microarchitecture which includes input queues, round-robin arbitration, a distributed tristate crossbar, and output buffers. The routers in our multihop networks have similar radices, so we fix the router latency to be two cycles. For a 5×5 router with 128 b flits of uniformly random data, we estimate the energy to be 16 pJ/flit. Notice that sending a 128 b flit across a 2.5 mm channel consumes roughly 13 pJ, which is comparable to the energy required to move this flit through a simple router. Future on-chip network designs must therefore carefully consider *both* channel and router energy, and to a lesser extent area.

3.2. Electrical On-chip Networks

Figure 1 illustrates four topologies that we will be discussing in this section and throughout the paper: global crossbars, two-dimensional meshes, concentrated meshes, and Clos networks. Table 1 shows some key parameters for these topologies assuming a MTBw system.

For systems with few tiles, a simple global crossbar is one of the most efficient network topologies and presents a simple performance model to software [8]. Such crossbars are strictly non-blocking; as long as an output is not oversubscribed every input can send messages to its desired output without contention. Small crossbars can have very low-latency and high-throughput but are difficult to scale to tens or hundreds of tiles.

Figure 1a illustrates a 64×64 crossbar network implemented with distributed tristate buses. Although such a network provides strictly non-blocking connectivity, it also requires a large number of global buses across the length of the chip. These buses are challenging to layout and must be pipelined for good throughput. Global arbitration can add significant latency and also needs to be pipelined. These global control and data wires result in significant power consumption even for communication

	Channels				Routers		Latency					
Topology	N_C	b_C	N_{BC}	$N_{BC} \cdot b_C$	N_R	radix	H	T_R	T_C	T_{TC}	T_S	T_0
Crossbar	*64	*128	*64	8,192	1	64x64	1	10	n/a	0	4	14
Mesh	224	256	16	4,096	64	5x5	2-15	2	1	0	2	7-46
CMesh	48	512	8	4,096	16	8x8	1-7	2	2	0	1	3-25
Clos	128	128	64	8,192	24	8x8	3	2	2-10	0-1	4	14-32

Table 1: Example MTBw Network Configurations – Networks sized to support 128 b/cycle per tile under uniform random traffic. N_c = number of channels, b_C = bits/channel, N_{BC} = number of bisection channels, N_R = number of routers, H = number of routers along data paths, T_R = router latency, T_C = channel latency, T_{TC} = latency from tile to first router, T_S = serialization latency, T_0 = zero load latency. *Crossbar "channels" are the shared crossbar buses.

between neighboring tiles. Thus global electrical crossbars are unlikely choices for future manycore on-chip networks, despite the fact that they might be the easiest to program.

Two-dimensional mesh networks (Figure 1b) are popular in systems with more tiles due to their simplicity in terms of design, wire routing, and decentralized flow-control [3, 14]. Unfortunately, high hop counts result in long latencies and significant energy consumption in both routers and channels. Because network latency and throughput are critically dependent on application mapping, low-dimensional mesh networks also impact programmer productivity by requiring careful optimization of task and data placement.

Moving from low-dimensional to high-dimensional mesh networks (e.g., 4-ary 3-cubes) reduces the network diameter, but requires long channels when mapped to a planar substrate. Also, higher-radix routers are required, resulting in more area and higher router energy. Instead of adding network dimensions, researchers have proposed using concentration to help reduce hop count [1]. Figure 1c illustrates a two-dimensional mesh with a concentration factor of four (cmesh). One of the disadvantages of cmesh topologies is that, for the same theoretical throughput, channels are wider than an equivalent mesh topology as shown in Table 1. One option to improve channel utilization for shorter messages is to divide resources among multiple parallel cmesh networks with narrower channels. The cmesh topology should achieve similar throughput as a standard mesh with half the latency at the cost of longer channels and higher-radix routers. CMesh topologies still require careful application mappings for good performance.

Clos networks offer an interesting intermediate point between the high-radix, low-diameter crossbar topology and the low-radix, high-diameter mesh topology [4]. Figure 1d illustrates an 8-ary 3-stage Clos topology which reduces the hop count but requires longer point-to-point channels. Figure 2 shows one possible layout of this topology. Clos networks use many small routers and extensive path diversity. Although the specific Clos network shown here is reconfigurably non-blocking instead of strictly non-blocking, we can still minimize congestion with an appropriate routing algorithm (assuming the outputs are not oversubscribed). Unfortunately, Clos networks still require global point-to-point channels and, as with a crossbar, these global channels can be difficult to layout and have significant energy cost.

4. Photonic On-Chip Networks

Silicon photonics is a promising new technology which offers lower power, higher bandwidth density, and shorter latencies than electrical interconnects. Photonics is particularly effective for global interconnects and thus has the potential to enable scalable low-diameter on-chip networks, which should ease manycore parallel programming. In this section, we first introduce the underlying photonic technology before discussing the cost of implementing some of the global photonic crossbars found in the literature. We then introduce our own approach to implementing a photonic Clos network, and compare its cost to photonic crossbars.

4.1. Photonic Technology

Figure 3 illustrates the various components in a typical wavelength-division multiplexed (WDM) photonic link used for on-chip communication. Light from an off-chip two-wavelength (λ_1, λ_2) laser source is carried by an optical fiber and then coupled into an on-chip waveguide. The waveguide carries the light past a series of transmitters, each using a resonant ring modulator to imprint the data on the corresponding wavelength. Modulated light continues through the waveguide to the other side of the chip where each of the two receivers use a tuned resonant ring filter to "drop" the corresponding wavelength from the waveguide into a local photodetector. The photodetector turns absorbed light into current, which is sensed by the electrical receiver. Both 3D and monolithic integration approaches have been proposed in the past few years to implement silicon-photonic on-chip networks.

With 3D integration, a separate specialized die or layer is used for photonic devices. Devices can be implemented in monocrystalline silicon-on-insulator (SoI) dies with

	Modulator and Driver Circuits			Receiver Circuits			
Design	DDE	FE	TTE	DDE	FE	TTE	ELP
Aggressive	20 fJ/bt	5 fJ/bt	16 fJ/bt/heater	20 fJ/bt	5 fJ/bt	16 fJ/bt/heater	3.3 W
Conservative	80 fJ/bt	10 fJ/bt	32 fJ/bt/heater	40 fJ/bt	20 fJ/bt	32 fJ/bt/heater	33 W

Table 2: Aggressive and Conservative Energy and Power Projections for Photonic Devices – fJ/bt = average energy per bit-time, DDE = Data-traffic dependent energy, FE = Fixed energy (clock, leakage), TTE = Thermal tuning energy (20K temperature range), ELP = Electrical laser power budget (30% laser efficiency).

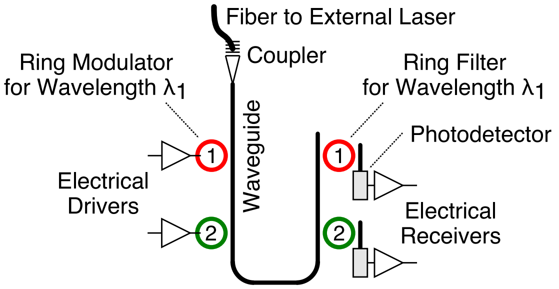


Figure 3: Photonic Components – Two point-to-point photonic links implemented with WDM.

Photonic device	Optical Loss (dB)
Optical Fiber (per cm)	0.5e-5
Coupler	1
Splitter	0.2
Non-linearity (at 30 mW)	1
Modulator Insertion	0 – 1
Waveguide (per cm)	0 – 5
Waveguide crossing	0.05
Filter through	1e-4 – 1e-2
Filter drop	1.5
Photodetector	0.1

Table 3: Optical Loss Ranges per Component

thick layer of buried oxide (BOX) [5], or in a separate layer of silicon nitride (SiN) deposited on top of the metal stack [2]. In this separate die or layer, customized processing steps can be used to optimize device performance. However, this customized processing approach increases the number of processing steps and hence manufacturing cost. In addition, the circuits required to interface the two chips can consume significant area and power.

With monolithic integration, photonic devices are designed using the existing process layers of a standard logic process. The photonic devices can be implemented in polysilicon on top of the shallow-trench isolation in a standard bulk CMOS process [9] or in monocrystalline silicon with advanced thin BOX SoI. Although monolithic integration may require some post-processing, its manufacturing cost can be lower than 3D integration. Monolithic integration decreases the area and energy required to interface electrical and photonic devices, but it requires active area for waveguides and other photonic devices.

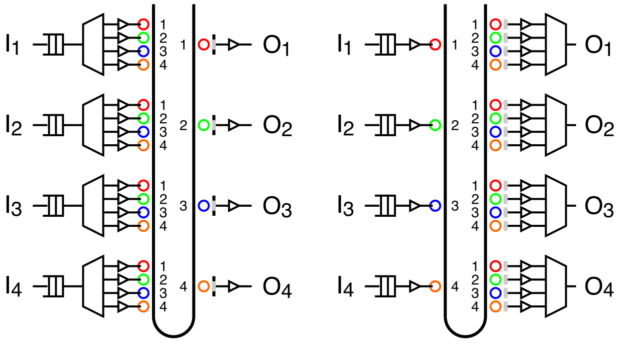
Irrespective of the chosen integration methodology, WDM optical links have many similar optical loss components (see Table 3). Optical loss affects system design, as it sets the required optical laser power and correspondingly the electrical laser power (at a roughly 30% conversion efficiency). Along the optical critical path, some losses such as coupler loss, non-linearity, photodetector loss, and filter drop loss are relatively independent of the network layout, size, and topology. For the scope of this study, we will focus on the loss components which significantly impact the overall power budget as a function of the type, radix, and throughput of the network.

In addition to optical loss, ring filters and modulators have to be thermally tuned to maintain their resonance under on-die temperature variations. Monolithic integration gives the most optimistic ring heating efficiency of all approaches (due to in-plane heaters and air-undercut), estimated at 1 μW per ring per K.

Based on our analysis of various photonic technologies and integration approaches, we make the following assumptions. With double-ring filters and a 4 THz free-spectral range, up to 128 wavelengths modulated at 10 Gb/s can be placed on each waveguide (64 in each direction, interleaved to alleviate filter roll-off requirements and crosstalk). A non-linearity limit of 30 mW at 1 dB loss is assumed for the waveguides. The waveguides are single mode and a pitch of 4 μm minimizes the crosstalk between neighboring waveguides. The ring diameters are ≈10 μm. The latency of a global photonic link is assumed to be 3 cycles (1 cycle in flight and 1 cycle each for E/O and O/E conversion). For monolithic integration we assume a 5 μm separation between the photonic and electrical devices to maintain signal integrity, while for 3D integration the photonic devices are designed on a separate specialized layer. Table 2 shows our assumptions for the photonic link energy and electrical laser power.

4.2. Photonic Global Crossbar Networks

A global crossbar provides non-blocking all-to-all communication between its inputs and outputs in a single stage. Figure 4 shows two approaches for implementing a 4×4 photonic crossbar. Both schemes have multiple single-wavelength photonic channels carried on



(a) DMXbar

(b) CMXbar

Figure 4: Photonic 4x4 Crossbars – Both crossbars have four inputs (I_{1-4}), four outputs (O_{1-4}), and four channels which are wavelength division multiplexed onto the U-shaped waveguide. Number next to each ring indicates resonant wavelength. (a) distributed mux crossbar (DMXbar) with one channel per output, (b) centralized mux crossbar (CMXbar) with one channel per input.

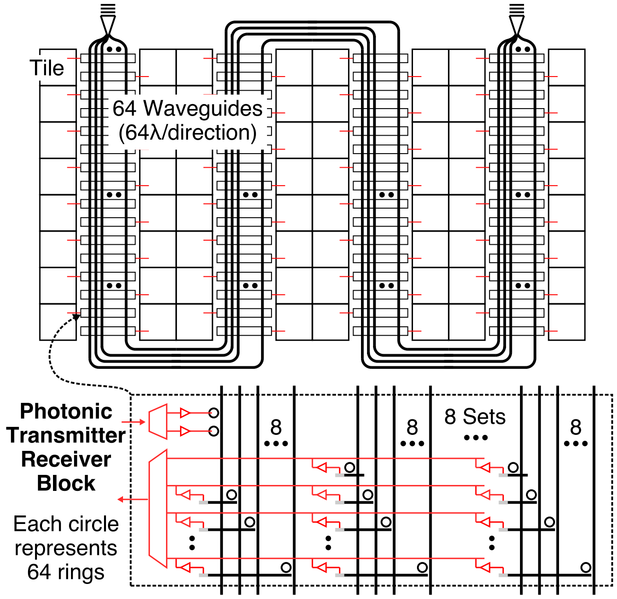


Figure 5: Serpentine Layout for 64x64 CMXbar – Electrical circuitry shown in red. 64 waveguides (8 sets of 8) are either routed between columns of tiles (monolithic integration) or over tiles (3D integration). One 128 b/cycle channel is mapped to each waveguide, with 64 λ going from left to right and 64 λ going from right to left. Each tile modulates a unique channel and every tile can receive from any channel.

a single waveguide using WDM. Crossbars with higher radix and/or greater channel bandwidths will require more wavelengths and more waveguides. Both examples require global arbitration to determine which input can send to which output. Various arbitration schemes are possible including electrical and photonic versions of centralized and distributed arbitration.

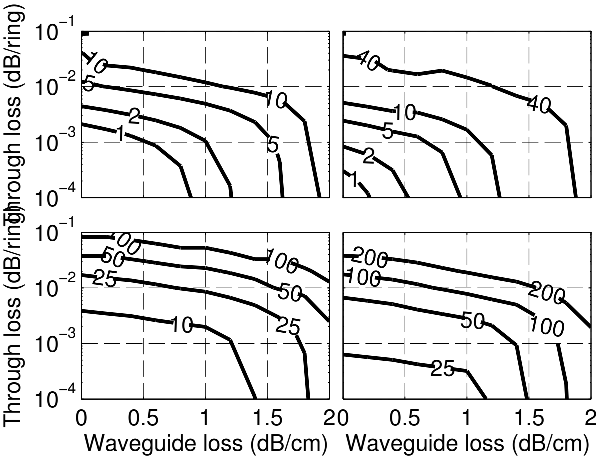
Figure 4a illustrates a distributed mux crossbar (DMXbar) where there is one channel per output and every input can modulate every output channel. As an example, if I_1 wants to send a message to O_3 it first arbitrates and then modulates wavelength λ_3. This light will experience four modulator insertion losses, 13 through losses, and one drop loss. Notice that although a DMXbar only needs one ring filter per output, it requires $O(nr^2)$ modulators where r is the crossbar radix and n is the number of wavelengths per port. For larger radix crossbars with wider channel bitwidths the number of modulators can significantly impact optical power, thermal tuning power, and area. For large distributed-mux crossbars this requires very aggressive photonic modulator device design. Vantrease et al. have proposed a global 64 × 64 photonic crossbar which is similar in spirit to the DMXbar scheme and requires about a million rings [15]. Their work uses a photonic token passing network to implement the required global arbitration.

Figure 4b illustrates an alternative approach called a centralized mux crossbar (CMXbar) where there is one channel per input and every output can listen to every input channel. As an example, if I_3 wants to send a message to O_1 it first arbitrates and then modulates wavelength λ_3. By default all ring filters at the receivers are slightly off-resonance so output O_1 receives the message by tuning in the ring filter for λ_3. This light will experience one modulator insertion loss, 13 through losses, three detuned receiver through losses, and one drop loss. If all ring filters were always tuned in, then wavelength λ_3 would have to be split among all the outputs even though only one output is ever going to actually receive the data. Although useful for broadcast, this would drastically increase the optical power. A CMXbar only needs one modulator per input (and so is less sensitive to modulator insertion loss), but it requires $O(nr^2)$ drop filters. As with the DMXbar, this can impact optical power, thermal tuning power, and area, and it necessitates aggressive reduction in the ring through loss. Additionally, tuning of the appropriate drop filter rings when receiving a message is done using charge injection, and this incurs a fixed overhead cost of 50 μW per tuned ring. Kırman et al. investigated a global bus-based architecture which is similar to the CMXbar scheme [7]. Nodes optically broadcast a request signal to all other nodes, and then a distributed arbitration scheme allows all nodes to agree on which receiver rings to tune in. Psota et al. have also proposed a CMXbar-like scheme which focuses on supporting global broadcast where all receivers are always tuned in [12].

Although Figure 4 shows two of the more common approaches proposed in the literature, there are other schemes which use a significantly different implementation. Zhou et al. describe an approach which replaces the U-shaped waveguide with a matrix of passive ring filters [17]. This approach still requires either multiple modulators per input or multiple ring filters per output, but results in shorter waveguide lengths since all wavelengths do not need to pass by all tiles. Unfortunately, the matrix also increases the number of rings and waveguide crossings. Petracca et al. describe a crossbar implementation which leverages photonic switching elements that switch many wavelengths with a single ring resonator [10]. Their scheme requires an electrical control network to configure these photonic switching elements, and thus is best suited for transmitting very long messages which amortize configuration overhead. In this paper, we focus on the schemes illustrated in Figure 4 and leave a detailed comparison to more complicated crossbars for future work.

The DMXbar and CMXbar schemes can be extended to much larger systems in a variety of ways. A naive extension of the CMXbar scheme in Figure 4b is to layout a global loop around the chip with light always traveling in one direction. Unfortunately this layout has an optical critical path which would traverse the loop twice. Figure 5 shows a more efficient serpentine layout of the CMXbar scheme for our target system of 64 tiles. This crossbar has 128 b/cycle input ports which makes it suitable for a MTBw system (i.e., 128 b/cycle per tile under uniform random traffic). At a 5 GHz clock rate, each channel uses 64 λ (10 Gb/s/λ), and we need a total of 64 waveguides (1 waveguide/channel). An input can send light in either direction on the waveguides, which shortens the optical critical path but requires additional modulators per input.

The total power dissipated in the on-chip photonic network can be divided into two components. The first component consists of power dissipated in the photonic components, i.e., power at the laser source and the power dissipated in thermal tuning. The second part consists of electrical power dissipated in the modulator driver, receiver, and arbitration circuits. Here we quantify the first power component and then in Section 5 we provide a detailed analysis of the second power component.



(a) LTBw (b) HTBw

Figure 6: Laser Optical Power (W) (top row) and Percent Area (bottom row) for 64×64 CMXbar – Systems implemented with serpentine layout on 20×20 mm die.

	Global Crossbar		Clos	
System	Rings	Power	Rings	Power
LTBw	266 k	5.3 W	14 k	0.28 W
HTBw	1,000 k	21.3 W	57 k	1.14 W

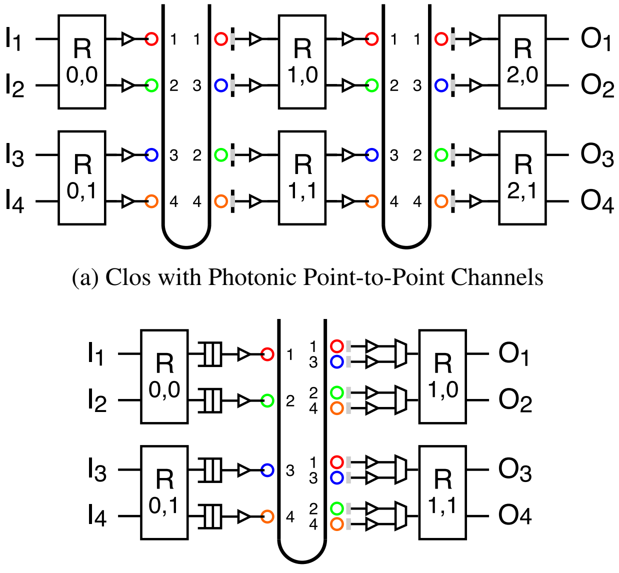
Table 4: Thermal Power – Power required to thermally tune the rings in the network over a temperature range of 20K.

The optical losses experienced in the various optical components and the desired network capacity determine the total optical power needed at the laser source. In the serpentine layout of a CMXbar, the waveguide and ring through loss are the dominant loss components, due to the long waveguides (9.5 cm) and large number of rings (128 modulator rings and 63 × 64 = 4032 filter rings) along each waveguide. Figure 6 shows two contour plots of the optical power required at the laser source for the LTBw and HTBw systems with a photonic CMXbar network. For a given value of waveguide loss and through loss per ring, the number of wavelengths per waveguide is the same for the two systems. However, the higher bandwidth system requires wider global buses which increases the optical power required at the laser source. As a result, the LTBw system can tolerate higher losses per component compared to the HTBw system for the same optical power budget.

Figure 6 shows contour plots of the percent area required for the optical devices for the LTBw and HTBw systems. The non-linearity limit affects the number of wavelengths that can be routed on each waveguide and hence the number of required waveguides, making photonic device area dependent on optical loss. As expected, the HTBw system requires increased photonic area for each loss combination. There is a lower limit on the area overhead which occurs when all of the wavelengths per waveguide are utilized. The minimum area for the LTBw and HTBw systems is 6% and 23%, respectively.

To calculate the required power for thermal tuning, we assume that under typical conditions the rings in the system would experience a temperature range of 20 K. Table 4 shows the power required for thermal tuning in the crossbar. Although each modulator and ring filter uses two cascaded rings, we assume that these two rings can share the same heater. The large number of rings in the crossbar significantly increases both thermal tuning and area overheads.

We can use a similar serpentine layout as the one shown in Figure 5 to implement a DMXbar. There would be one output tile per waveguide and there would be no need to tune or detune the drop filters. We would, however, require a large number of modulators per waveguide



(a) Clos with Photonic Point-to-Point Channels

(b) Clos with Photonic Middle Routers

Figure 7: Photonic 2-ary 3-stage Clos Networks – Both networks have four inputs (I_{1-4}), four outputs (O_{1-4}), and six 2×2 routers ($R_{0-2,0-1}$). (a) four point-to-point photonic channels use WDM on each U-shaped waveguide. (b) the two middle routers ($R_{1,0-1}$) are implemented with photonic 2 × 2 CMXbars on a single U-shaped waveguide. Number next to each ring indicates resonant wavelength.

(63 × 64 = 4032) and modulator insertion loss would most likely dominate the optical power loss. For this topology to be feasible, novel modulators with close to 0 dB insertion loss need to be designed. The area for photonic devices and power dissipated in thermally tuning the rings would be similar to that in the CMXbar implementation.

The large number of rings required for photonic crossbar implementations make monolithic integration impractical from an area perspective, and 3D integration is expensive due to the power cost of thermal tuning (even in the case when all the circuits of the inactive transmitters/receivers can be fully powered down). The actual cost of these crossbar networks will be even higher than indicated in this section since we have not accounted for arbitration overhead. These observations motivate our interest in photonic Clos networks which preserve much of the simplicity of the crossbar programming model, while significantly reducing area and power.

4.3. Photonic Clos Networks

As described in Section 3.2, a Clos network uses multiple stages of small routers to create a larger non-blocking all-to-all network. Figure 7 shows two approaches for implementing a 2-ary 3-stage Clos network. In Figure 7a, all of the Clos routers are implemented electrically and the inter-router channels are implemented with photonics. As an example, if input I_2 wants to communicate with output O_4 then it can use either middle router. If the routing algorithm chooses $R_{1,1}$, then the network will use wavelength λ_2 on the first waveguide to send the message to $R_{1,1}$ and wavelength λ_4 on the second waveguide to send the message to O_4. Figure 7b is logically the same topology, but each middle router is implemented with photonic CMXbar. The channels for both crossbars are multiplexed onto the same waveguide using WDM. Note that we still use electrical buffering and arbitration for these photonic middle routers. Using photonic instead of electrical middle routers removes one stage of EOE conversion and can potentially lower the dynamic power of the middle router crossbars, but at the cost of higher optical and thermal tuning power. Depending on photonic device losses, this tradeoff may be beneficial since for our target system the radix of the Clos routers (8×8) is relatively low. In this paper, we focus on the Clos with photonic point-to-point channels since it should have the lowest optical power, thermal tuning power, and area overhead.

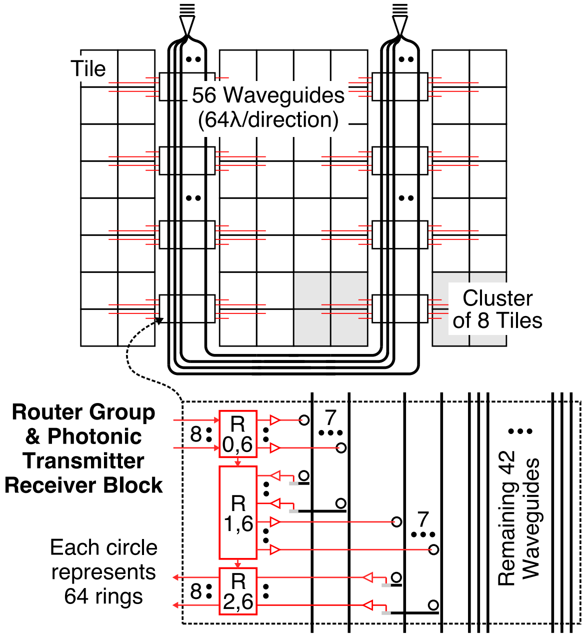
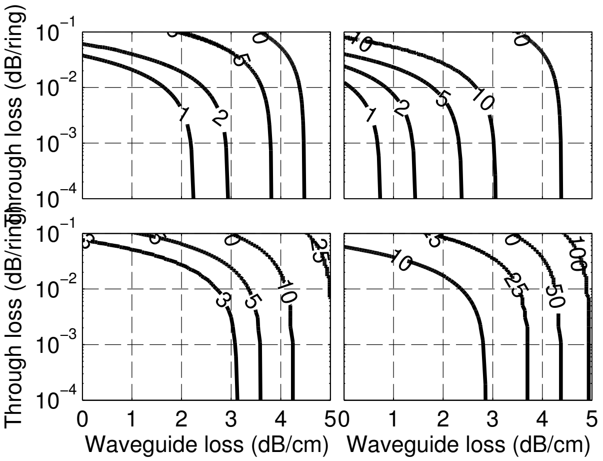


Figure 8: U-Shaped Layout for 8-ary 3-stage Clos – Electrical circuitry shown in red. 56 waveguides (8 sets of 7) are either routed between columns of tiles (monolithic integration) or over tiles (3D integration). Each of the 8 clusters (8 tiles per cluster) has electrical channels to its router group which contains one router per Clos stage. In the inset, the first set of 7 waveguides are used for channels (each 64 λ = 128 b/cycle) connecting to and from every other cluster. The second set of 7 waveguides are used for the second half of the Clos network. The remaining 42 waveguides are used for point-to-point channels between other clusters.

As in the crossbar case, there are multiple ways to extend this smaller Clos network to larger systems. For a fair comparison, we keep the same packaging constraints (i.e., location of vertical couplers) and also try to use



(a) LTBw (b) HTBw

Figure 9: Laser Optical Power (W) (top row) and Percent Area (bottom row) for 8-ary 3-stage Clos – Systems implemented with U-shaped layout on 20×20 mm die.

the light from the laser most efficiently. Figure 8 shows the U-shaped layout of the photonic Clos network in a MTBw system, which corresponds to 64 λ per channel. Each point-to-point photonic channel uses either forward or backward propagating wavelengths depending on the physical location of the source and destination clusters.

In a Clos network, the waveguide and ring through losses contribute significantly to the total optical loss but to a lesser extent than in a crossbar network, due to shorter waveguides and less rings along each waveguide. All the waveguides in the Clos network are roughly 2× shorter and with 20× less rings along each waveguide compared to a crossbar network. Figure 9 shows the optical power contours for the Clos network.

Although the number of optical channels in the Clos network is higher than in the crossbar network, the total number of rings (for same bandwidth) is significantly smaller since optical channels are point-to-point, resulting in significantly smaller tuning (Table 4) and area costs. The area overhead shown in Figure 9 is much smaller than for a crossbar due to shorter waveguides and smaller number of rings and is well suited for monolithic integration with a wider range of device losses. The lower limit on the area overhead is 2% and 8% for LTBw and HTBw, respectively.

Based on this design-space exploration we propose using the photonic Clos network for on-chip communication. Clos networks have lower area and thermal tuning costs and higher tolerance of photonic device losses as compared to global photonic crossbars. In the next section we compare this photonic Clos network with electrical implementations of mesh, cmesh, and Clos networks in terms of throughput, latency, and power.

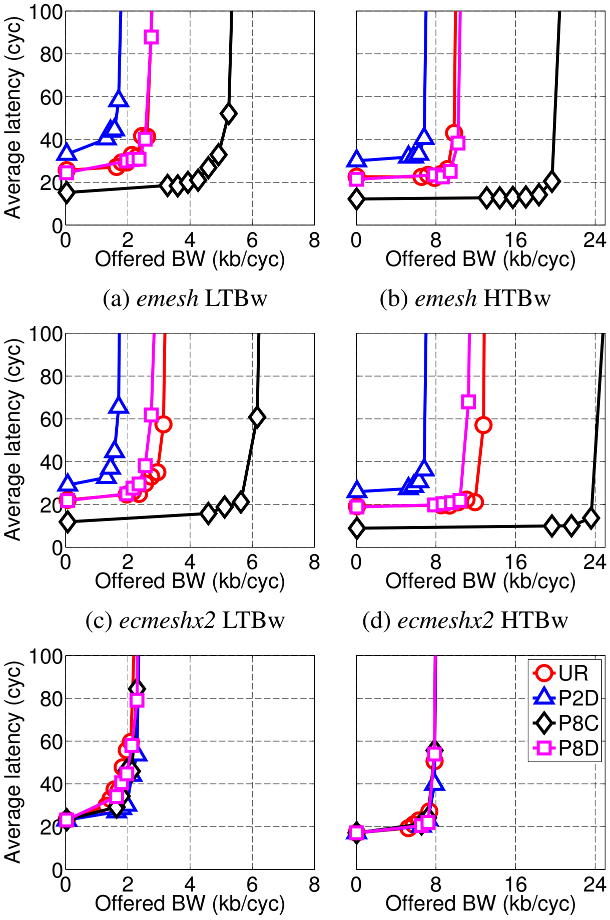
5. Simulation Results

In this section, we use a detailed cycle-accurate microarchitectural simulator to study the performance and power of various electrical and photonic networks for a 64-tile system with 512 b messages. Our model includes pipeline latencies, router contention, flow control, and serialization overheads. Warm-up, measure, and drain phases of several thousand cycles and infinite source queues were used to accurately determine the latency at a given injection rate. Various events (e.g., channel utilization, queue accesses, arbitration) were counted during simulation and then multiplied by energy values derived from first-order gate-level models.

Our baseline includes three electrical networks: a 2D mesh (*emesh*), a mesh with a concentration factor of four (*ecmeshx2*), and an 8-ary 3-stage Clos (*eclos*). Because a single concentrated mesh would have channel bitwidths larger than our message size for some configurations, we implement two parallel cmeshes with narrow channels and randomly interleave messages between them. We also study a photonic implementation of the Clos network (*pclos*) with aggressive (*pclos-a*) and conservative (*pclos-c*) photonic devices (see Table 2). We show results for LTBw and HTBw systems which correspond to ideal throughputs of 64 b/cycle and 256 b/cycle per tile for uniform random traffic. Our mesh networks use dimension-ordered routing, while our Clos networks use a randomized oblivious routing algorithm (i.e., randomly choosing the middle router). All networks use wormhole flow control.

We use synthetic traffic patterns based on a partitioned application model. Each traffic pattern has some number of logical partitions, and tiles randomly communicate only with other tiles that are in the same partition. These logical partitions are then mapped to physical tiles in either a co-located fashion (tiles within a partition are physically grouped together) or in a distributed fashion (tiles in a partition are distributed across the chip). We believe these partitioned traffic patterns capture the varying locality present in manycore programs. Although we studied various partition sizes and mappings, we focus on the following four representative patterns in this paper. A single global partition is identical to the standard uniform random traffic pattern (UR). The P8C pattern has eight partitions each with eight tiles optimally co-located together. The P8D pattern stripes these partitions across the chip. The P2D pattern has 32 partitions each with two tiles, and these two tiles are mapped to diagonally opposite quadrants of the chip.

Figure 10 shows the latency versus offered bandwidth for the LTBw and HTBw systems with different traffic patterns. In both *emesh* and *ecmeshx2*, the P8C traffic pattern requires only local communication and thus has higher performance. The P2D traffic pattern requires global communication which results in lower per-



(a) *emesh* LTBw (b) *emesh* HTBw

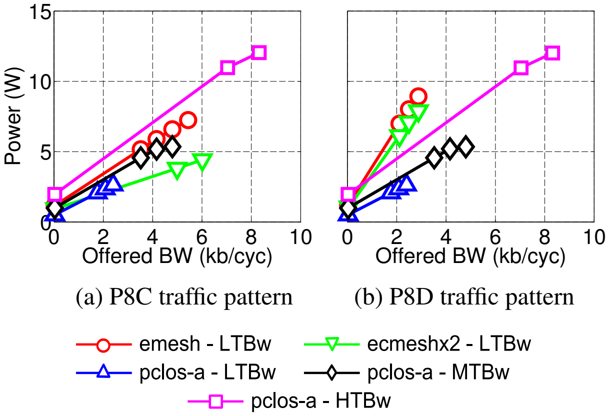
(c) *ecmeshx2* LTBw (d) *ecmeshx2* HTBw

(e) *pclos* LTBw (f) *pclos* HTBw

Figure 10: Latency vs. Offered Bandwidth – LTBw systems have a theoretical throughput of 64 b/cycle per tile for UR; corresponding for HTBw is 256 b/cycle.

formance. On average, *ecmeshx2* saturates at higher bandwidths than *emesh* due to the path diversity provided by the two cmesh networks, and has lower latency due to lower average hop count. Although not shown in Figure 10, the *eclos* network has similar saturation throughput to *pclos* but with higher average latency. Because *pclos* always distributes traffic randomly across its middle routers, it has uniform latency and throughput across all traffic patterns. Note, however, that *pclos* performs better than *emesh* and *emeshx2* on global traffic patterns (e.g., P2D) and worse on local traffic patterns (e.g., P8C). If the *pclos* power consumption is low enough for the LTBw system then we should be able to increase the size to a MTBw or HTBw system. A larger *pclos* network will hopefully have similar performance and energy-efficiency for local traffic patterns as compared to *emesh* and *ecmeshx2* and much better performance and energy-efficiency for global traffic patterns.

Figure 11 shows the power dissipation versus offered bandwidth for various network topologies with the P8C and P8D traffic patterns. In order to match the performance of *ecmeshx2* LTBw system we need to use the *pclos-a* MTBw system which has slightly higher power for the P8C traffic pattern (local communication) and much lower power for the P8D traffic pattern (global communication) assuming we are at medium to high load. Laser power is not included in Figure 11 which may be appropriate for systems primarily limited by the power density of the processor chip, but may not be appropriate for energy-constrained systems or for systems limited by the total power consumption of the motherboard.



(a) P8C traffic pattern (b) P8D traffic pattern

Figure 11: Power Dissipation vs. Offered Bandwidth – 3.3 W laser power not included for the *pclos-a* topology.

Figure 12 shows the power breakdowns for various topologies and traffic patterns, for both LTBw and HTBw design points that can support the desired offered bandwidth with lowest power. Compared to *emesh* and *ecmeshx2*, the *pclos-a* network provides comparable performance and low power dissipation for global traffic patterns, and comparable performance and power dissipation for local traffic patterns. The *pclos-a* network energy-efficiency increases when sized for higher throughputs (higher utilization) due to static laser power component. More importantly, the *pclos-a* network offers a global low-dimensional network with uniform performance which should simplify manycore parallel programming. The energy efficiency of *pclos* network might be further improved by investigating alternative implementations which use photonic middle switch router as shown in Figure 7b.

It is important to note that with conservative optical technology projections, even in relatively simple optical network like *pclos*, the required electrical laser power is much larger than other components, and the photonic network will usually consume higher power than the electrical networks. This strong coupling between overall network performance, topology and underlying photonic components underlines the need for a fully integrated vertical design approach illustrated in this paper.

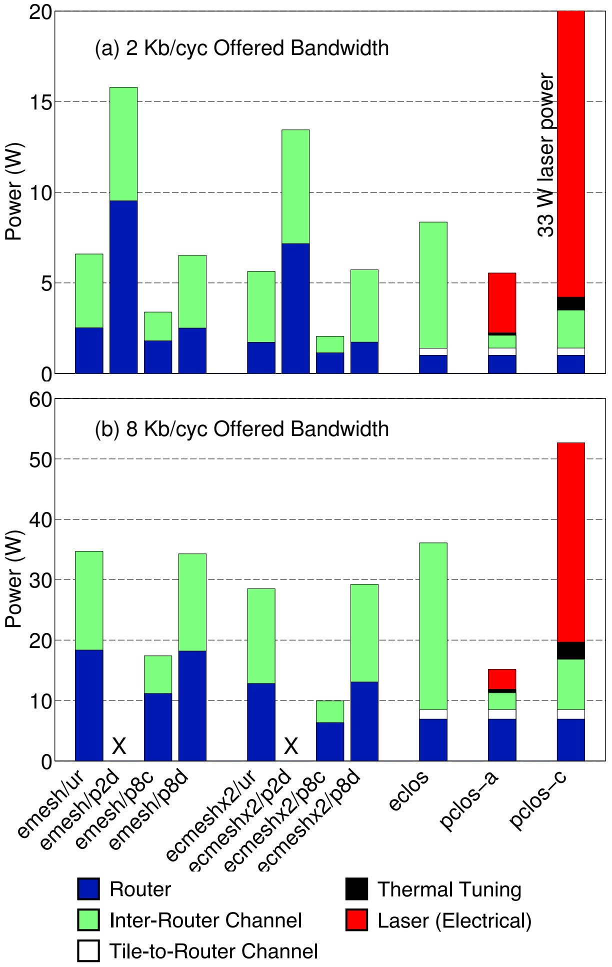


Figure 12: Dynamic Power Breakdown – Power of *eclos* and *pclos* did not vary significantly across traffic patterns. (a) LTBw systems at 2 kb/cycle offered bandwidth (except for *emesh/p2d* and *ecmeshx2/p2d* which saturated before 2 kb/cycle, HTBw system shown instead), (b) HTBw systems at 8 kb/cycle offered bandwidth (except for *emesh/p2d* and *ecmeshx2/p2d* which saturated before 8 kb/cycle).

6. Conclusion

We have proposed and evaluated a silicon-photonic Clos network for global on-chip communication. Since the Clos network uses point-to-point channels instead of the global shared channels found in crossbar networks, our photonic Clos implementations consume significantly less optical power, thermal tuning power, and area overhead, while imposing less aggressive loss requirements on photonic devices. Our simulations show that the resulting photonic Clos networks should provide higher energy-efficiency than electrical implementations of mesh and Clos networks with equivalent throughput. A unique feature of a photonic Clos network is that it provides uniformly low latency and uniformly high bandwidth regardless of traffic pattern, which helps reduce the programming challenge introduced by highly parallel systems.

Acknowledgments

This work was supported in part by Intel Corp. and DARPA awards W911NF-08-1-0134 and W911NF-08-1-0139.

References

[1] J. Balfour and W. J. Dally. Design tradeoffs for tiled CMP on-chip networks. *Int'l Conf. on Supercomputing*, 2006.

[2] T. Barwicz et al. Silicon photonics for compact, energy-efficient interconnects. *Journal of Optical Networking*, 6(1):63–73, 2007.

[3] S. Bell et al. TILE64 processor: A 64-core SoC with mesh interconnect. *Int'l Solid-State Circuits Conf.*, Feb. 2008.

[4] C. Clos. A study of non-blocking switching networks. *Bell System Technical Journal*, 32:406–424, 1953.

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

[6] B. Kim and V. Stojanović. Characterization of equalized and repeated interconnects for NoC applications. *IEEE Design and Test of Computers*, 25(5):430–439, 2008.

[7] N. Kırman et al. Leveraging optical technology in future bus-based chip multiprocessors. *Int'l Symp. on Microarchitecture*, Dec. 2006.

[8] U. Nawathe et al. An 8-core 64-thread 64 b power-efficient SPARC SoC. *Int'l Solid-State Circuits Conf.*, Feb. 2007.

[9] J. Orcutt et al. Demonstration of an electronic photonic integrated circuit in a commercial scaled bulk CMOS process. *Conf. on Lasers and Electro-Optics*, May 2008.

[10] M. Petracca et al. Design exploration of optical interconnection networks for chip multiprocessors. *Symp. on High-Performance Interconnects*, Aug. 2008.

[11] D. Pham et al. The design and implementation of a first-generation CELL processor. *Int'l Solid-State Circuits Conf.*, Feb. 2005.

[12] J. Psota et al. ATAC: On-chip optical networks for multicore processors. *Boston Area Architecture Workshop*, Jan. 2007.

[13] A. Shacham, K. Bergman, and L. Carloni. On the design of a photonic network-on-chip. *Int'l Symp. on Networks-on-Chip*, May 2007.

[14] S. Vangal et al. 80-tile 1.28 TFlops network-on-chip in 65 nm CMOS. *Int'l Solid-State Circuits Conf.*, Feb. 2007.

[15] D. Vantrease et al. Corona: System implications of emerging nanophotonic technology. *Int'l Symp. on Computer Architecture*, June 2008.

[16] W. Zhao and Y. Cao. New generation of predictive technology model for sub-45 nm early design exploration. *Trans. on Electron Devices*, 53(11):2816–2823, Nov. 2006.

[17] L. Zhou et al. Design and evaluation of an arbitration-free passive optical crossbar for on-chip interconnection networks. *Applied Physics A: Materials Science & Processing*, Feb. 2009.