# Silicon-Photonic Network Architectures for Scalable, Power-Efficient Multi-Chip Systems

Pranay Koka
Sun Labs, Oracle
pranay.koka@oracle.com

Michael O. McCracken
Sun Labs, Oracle
michael.mccracken@oracle.com

Herb Schwetman
Sun Labs, Oracle
herb.schwetman@oracle.com

Xuezhe Zheng
Sun Labs, Oracle
xuezhe.zheng@oracle.com

Ron Ho
Sun Labs, Oracle
ron.ho@oracle.com

Ashok V. Krishnamoorthy
Sun Labs, Oracle
ashok.krishnamoorthy@oracle.com

## ABSTRACT

Scaling trends of logic, memories, and interconnect networks lead towards dense many-core chips. Unfortunately, process yields and reticle sizes limit the scalability of large single-chip systems. Multi-chip systems break free of these areal limits, but in turn require enormous chip-to-chip bandwidth. The "macrochip" concept presented here integrates multiple many-core processor chips in a single package with silicon-photonic interconnects. This design enables a multi-chip system to approach the performance of a single large die.

In this paper we propose three silicon-photonic network designs that provide low-power, high-bandwidth inter-die communication: a static wavelength-routed point-to-point network, a "two-phase" arbitrated network, and a limited-connectivity point-to-point network. We also adapt two existing intra-chip silicon-photonic interconnects: a token-ring-based crossbar and a circuit-switched torus.

We simulate a 64-die, 512-core cache-coherent macrochip using all of the above networks with synthetic kernels, and kernels from Splash-2 and PARSEC. We evaluate the networks on performance, optical power and complexity. Despite a narrow data-path width compared to the token-ring or torus, the point-to-point performs 3.3× and 3.9× better respectively. We show that the point-to-point is over 10× more power-efficient than the other networks. We also show that, contrary to electronic network designs, a point-to-point network has the lowest design complexity for an inter-chip silicon-photonic network.

## Categories and Subject Descriptors

B.4.3 [**Hardware**]: Interconnections—*Topology* ; C.1.2 [**Computer Systems Organization**]: Multiprocessors—*Interconnection architectures*

## General Terms

Design, Performance


*ISCA'10,* June 19–23, 2010, Saint-Malo, France.


## Keywords

Interconnection Networks, Nanophotonics

## 1. INTRODUCTION AND MOTIVATION

Today, state-of-the art processors feature multiple cores, such as processors with 4 to 8 cores along with the concomitant cache and memory controllers all on one die [17, 32]. Projections for the next decade indicate tens to hundreds of cores on a die [2,4,39]. As the number of chips and memories grow, the need for efficient on-chip interconnect networks becomes even greater, especially as chip power becomes dominated not by processor cores but by the need to transport data between processors and to memory [23]. Over the years, many electronic technologies [8,18] and on-chip network and router architectures [11, 21, 26] have been proposed to meet the need for increasing on-chip communications. For this communication, on-chip wires can provide suitably high-bandwidth and low-power interconnect [15, 16, 20, 29].

The trend to larger numbers of cores coupled with recent memory and on-chip communications technologies leads to dense, powerful compute blocks–a single many-processor chip. Unfortunately, the scalability of this single processor-chip approach is limited by the low process yields of large chips [31, 38]. One way to overcome this area constraint is to aggregate together several chips in a package, breaking free of the "reticle limit" of a single chip by using many individual smaller chips. Such a strategy requires enormous chip-to-chip bandwidth for these separate chips to perform as a contiguous piece of silicon, as well as the ancillary packaging, power delivery, and heat removal technologies for the aggregated chips. A popular strategy for such integration is to use vertical "3D" stacking of chips, connected using through-silicon-vias [5]. However, limits on delivering power to–and removing heat from–chips placed squarely atop one another means that vertical stacking is best employed for low-power applications such as DRAM integration. By contrast, high-performance and high-power processors are ideally spread out as an array of chips in a larger package, allowing power delivery to the chips' fronts and heat removal from their backs. Interconnecting such an array of chips presents a challenge: the density of off-chip I/O and package routes dramatically lags that of on-chip wires [36], forcing the use of overclocked and high-power serial links. Newer approaches using coupled data communication [12, 19] bypass soldered I/O and package routing and instead employ silicon "bridge" chips between processors, thus carrying all data over dense

on-chip wires [30]. While these systems enable modest arrays of chips, their scalability is limited by the low speed of on-chip wires, especially over distances longer than 10mm.

Silicon photonics is an emerging technology that may help to fill this need for high bandwidth, low power-per-bit channels essential for the deployment of multi-chip systems based on these many-core processors [3, 14, 23, 37, 40, 43]. On-chip optical channels paired seamlessly with inter-chip optical waveguides can provide up to 20 Gb/sec per wavelength of light. The ability to multiplex many wavelengths onto a single waveguide promises very high bandwidth density. Optical links are expected to provide latencies of 0.1 ns/cm at an energy cost of 160 femto-joules/bit (fJ/bit) using coupling structures to silicon waveguides of under 20 micron pitch [24, 25]. At such a low area, energy, and latency cost– especially relative to electrical interconnects–these optical links dramatically reduce the incremental cost of chip-to-chip bandwidth and open up entirely new areas of system exploration.

One such direction might be to widely separate discrete processor chips and interconnect them using fibers, thus using optical links to create physically large but logically dense systems. This would offer simpler packaging, power, and heat requirements yet seemingly provide the bandwidth advantages of wavelength multiplexing. However, chips connect to fibers at a relatively large 250 $\mu$m core pitch, not the 20 $\mu$m pitch of optical proximity couplers, so chip-to-chip bandwidth over fibers would not be much improved over area solder balls connected to package routes. To truly exploit the bandwidth advantages of silicon photonics, a high-performance system should instead employ dense silicon waveguides with fine-pitch connectors and tightly packed processors.

We introduce the macrochip, which is a technology platform for building a large processor node by integrating multiple processor die with a silicon-photonic interconnection network. The network is embedded in a Silicon-on-Insulator (SOI) substrate, and the processor die are connected to the network using optical proximity communication, which together make inter-die and intra-die communication bandwidths nearly equivalent. This approach provides a single-package compute block much larger than a single processor, but requiring neither large chips with low yield nor large numbers of input/output pins, which are expensive in area and power. The macrochip can thus be viewed as a silicon-photonic-based large scale shared memory multi-processor or a "cluster on a chip." In this paper we have attempted to provide all architecturally-relevant information about the Silicon-photonic technology used. Full details about specific devices can be found in [25]. Fundamental to such a system, however, is the silicon photonic network that connects together the individual chips.

Prior work in silicon-photonic networks has introduced network topologies such as an optical crossbar with token-ring arbitration [40] and a torus [37] in the context of single-die systems. In this paper, we propose the architecture and design of three silicon photonic networks for macrochip-like architectures: a statically-routed wavelength division multiplexed (WDM) point-to-point network that requires minimum optical complexity and no arbitration or switching; a limited point-to-point network that requires no more than one electronic router-hop per message; and a shared row data network with two-phase arbitration that provides

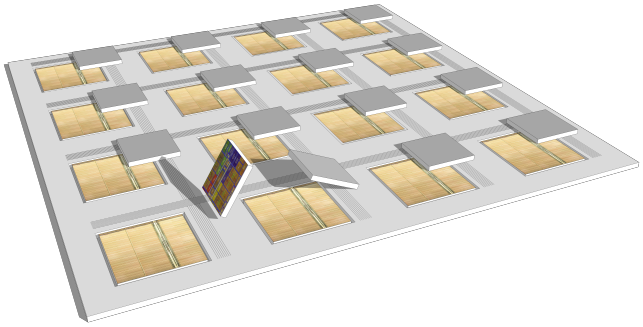


**Figure 1: Physical Layout of a 4×4 Macrochip. Two Face-Down CPU Die Are Shown At An Angle.**

greater die-to-die bandwidth but uses optical switches. We also evaluate the performance, power, and complexity of these networks in a macrochip-like architecture.

Because no prior work has focused on a macrochip-like shared memory multiprocessor, we adapt two promising silicon photonic intra-chip network architectures to the macrochip layout and technology as a point of comparison. We perform a thorough simulation-based performance evaluation of all five network architectures on two of the SPLASH-2 shared-memory benchmarks [42], three of the PARSEC benchmarks [6], and four synthetic benchmarks.

The contributions of this paper include:

- An architectural introduction to the macrochip, a technology platform for a computing system embodying a silicon-photonic intra-die interconnection network with high peak aggregate bandwidth, low latency and low power.
- A classification of optical on-chip networks.
- Three interconnection network topologies designed for the macrochip, two previously unpublished. One network is optimized for all-to-all communications patterns and the other two offer improved link bandwidths.
- A thorough performance, power and complexity evaluation of all network architectures.

The rest of the paper is organized as follows. In section 2, we describe the enabling silicon-photonic technology with regards to the various optical components used including their power, bandwidth and latencies. In section 3, we present a brief description of the macrochip architecture and give some of its advantages. In section 4, we describe the architecture and design of the three proposed network architectures along with the adapted reference architectures. In sections 5 and 6, we describe our evaluation methodology and analyze our results. Finally we survey the background work in related areas in section 7 and conclude in section 8.

## 2. SILICON-PHOTONIC TECHNOLOGY

In this section, we give a brief overview of the silicon-photonic components used in a macrochip. A macrochip consists of a large silicon routing substrate containing optical waveguides. This routing layer has chip-sized openings etched into its surface, and in each opening (called a site) is a processor chip with several cores and caches, and a memory

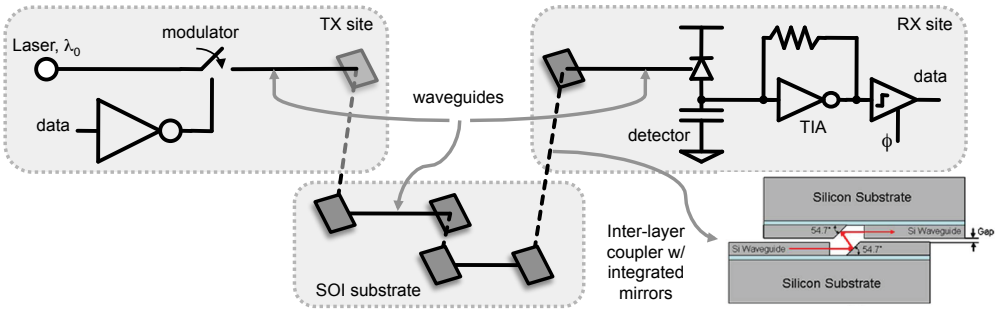



**Figure 2: A Single Photonic Link. Not Shown Are Mux/Demux Devices for WDM.**

| | Energy | Signal Loss |
|---|---|---|
| Modulator | 35 fJ/bit (dynamic) | 4 dB |
| OPxC | *negligible* | 1.2 dB |
| Waveguide | *negligible* | 0.5 dB/cm |
| Drop Filter | *negligible* | 0.1 dB or 1.5 dB |
| Receiver | 65 fJ/bit (dynamic) | N/A |
| Switch | *negligible* | 1 dB |
| Laser | 50 fJ/bit (static) | N/A |

**Table 1: Optical Component Properties**

chip. Site-to-site communication uses the optical waveguides in the routing layer; the silicon routing layer contains nothing but these passive optical waveguides, making its yield reasonable (see [25, 43] for details).

Figure 2 shows a canonical representation of a site-to-site photonic link. Fiber arrays bring off-chip continuous wave laser sources into the macrochip. At each source site, a single-wavelength laser light is modulated by an electronic digital data signal using an electro-optic (EO) modulator. The resulting optical data signal couples to a waveguide on the silicon photonic routing layer through low-loss optical proximity communication (OPxC) [43]. As shown in figure 2, in which two chips placed face-to-face transfer light from waveguide to waveguide through a pair of matching and aligned reflective mirrors [9]. On the silicon-photonic routing layer, the optical data signal travels along waveguides to the destination, where it couples via OPxC up to the destination site. There, a photodetector and electronic amplifier convert the optical data signal into an electronic digital stream.

Wavelength division multiplexing (WDM) in the network increases bandwidth density by reducing the number of routing waveguides and enabling multiple data channels per waveguide. With WDM, several modulated data streams at different wavelengths (from different source lasers) share a single waveguide.

Next, we discuss technology options for each of the major optical components and describe their performance characteristics. The component parameters are based on extensive, on-going device development and reasoned extrapolations to the 2014–2015 time frame. Table 1 presents a summary of the component energy and signal loss characteristics.

Off-chip **laser sources** connect to the macrochip via optical fiber array cables using either edge coupling or surface normal coupling enabled by grating couplers. A commercially available WDM-compatible distributed feedback laser can source a single wavelength at 10 mW of optical power. Optical power sharing can then reduce the total number of laser sources required.

**Modulators** convert an electronic data stream into an optical data stream. One of the most promising candidate technologies is a silicon CMOS modulator consisting of a reverse-biased, carrier-depletion ring resonator [44]. For such rings running at 20 Gb/sec we envision the modulator power to be 0.7 mW, and have a resonator $Q$ of about 12,000. During operation, ring modulators introduce considerable optical loss; we target a total loss of 4 dB. When disabled, ring loss is significantly smaller at 0.1 dB.

A **multiplexer** combines multiple channels on separate wavelengths into a single waveguide. One way to implement a compact multiplexer is to use cascaded silicon CMOS ring resonators [46]. The primary challenge for using such rings is efficiently tuning them to overcome fabrication tolerances and ambient temperature variations of the ring filter; we target 0.1 mW per wavelength tuning power. We target worst-case channel insertion loss of 2.5 dB.

**Optical proximity communication** (OPxC) couples an optical data signal from a waveguide on one chip to a waveguide on another, if the chips are placed face-to-face. The coupling is broadband with a targeted insertion loss of 1.2 dB per coupling. One way to accomplish this is to use mutually aligned waveguide gratings on two chips [28, 41]. Another approach employs a pair of mutually aligned reflecting mirrors [24, 43].

**Waveguides** route optical signals from their source site to their destination site. The macrochip employs two types of waveguides: short local waveguides on a thinned SOI substrate with less than 0.5 dB/cm loss, and global waveguides on a 3 $\mu$m thick SOI routing layer with less than 0.1 dB/cm loss. The low-loss global waveguides have a pitch of 10 $\mu$m. Across the largest envisioned macrochip, the worst case waveguide loss for site-to-site transmission is 6 dB.

A **drop-filter** demultiplexes a single wavelength from a shared multi-wavelength waveguide. A filter has two outputs: one output carries the optical data stream at the selected wavelength, and the other carries the remaining wavelength channels. Silicon CMOS ring resonators are a promising candidate for implementing compact, low-power drop filters. As with multiplexers, the power required to tune the drop-filter is targeted to be 0.1 mW per wavelength. The insertion loss is 0.1 dB for each wavelength that passes through the device and 1.5 dB for the selected drop wavelength.

A **receiver** consists of a waveguide photodetector and

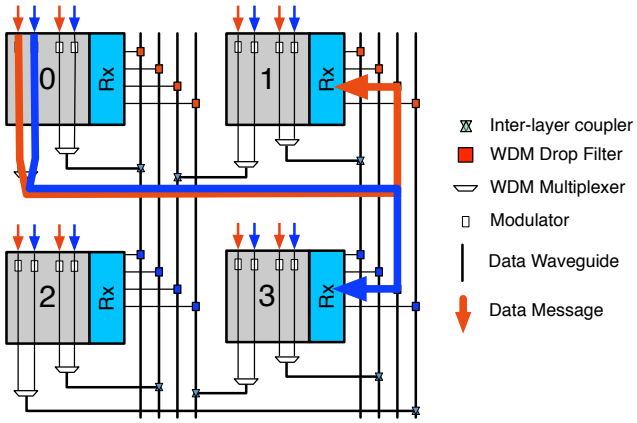



**Figure 3: 2×2 Static-WDM Point to Point Network Routing**

electronic amplifiers, and converts a single-wavelength optical signal into an electronic digital signal [45]. We target operation at 20 Gb/sec with a sensitivity of -21 dBm, and consume 1.3 mW.

A **broadband optical switch** can be implemented using waveguide Mach-Zehnder interferometers [13], although at large area cost. A compact quasi-broadband optical switch can be created using the periodic resonances of a ring resonator of an appropriate size [27]. Such a switch can direct input light to one of two output channels over a 30 nm wavelength range. With aggressive development, we target the maximum insertion loss for the switch to be under 1 dB, with a power consumption of 0.5 mW.

With these envisioned components, the optical link loss for an un-switched link is 17 dB. If a laser launches 0 dBm power at the modulator, a receiver sensitivity of -21 dBm provides 4 dB margin, which should be sufficient for reliable link operation.

Silicon photonic interconnection networks are being actively researched because of at least two important properties. First, they promise high-speed, low-latency data transmission, at about $0.3c$ in SOI waveguides. Second, optics potentially provides low energy data transmission, projected to be significantly less than 1 picojoule/bit [25]. Achieving such energy targets involves many optics and circuits challenges, including high efficiency resonator tuning, efficient WDM lasers, ultra-low power modulator and receiver circuits, precision chip alignment, low cost packaging, seamless fiber-optic off-chip communications, and many more. While these issues are beyond the scope of this paper, they are discussed in many of the references, including [25].

## 3. THE MACROCHIP

The macrochip architecture integrates multiple conventional die, each about 225 $mm^2$ in size, using silicon photonics to achieve performance similar to that of an equivalent 64x225 $mm^2$ single die. This design bypasses die size limits imposed by technology yields and makes possible dramatically more cores on a virtual "chip." The macrochip can be viewed as a large scale shared-memory multiprocessor or a cluster, whose performance is not restricted by the limited pin counts on processor die because all cores are interconnected through dense silicon photonics.

The macrochip achieves multi-die integration through a large SOI substrate supporting an array of physically separate CMOS die, called sites. The substrate contains two layers of silicon optical waveguides; the layers run in orthogonal directions much like on-chip electrical wiring, with via-like connections between the layers built using low-loss OPxC connectors. By using two optical routing layers, orthogonal waveguides avoid physically intersecting and suffering signal crosstalk. The substrate layers are SOI because the silicon waveguides require a buried oxide for light confinement [25], although photonics-enabled bulk silicon may in the future eliminate the need for SOI [33].

The macrochip shown in figure 1 is an $8 \times 8$ array of sites, where each site contains both processor and memory die. The memory can be conventional DRAM or other technology, and occupies up to 225 $mm^2$ in area. It sits face-up in a cutout in the SOI routing substrate. A smaller 125 $mm^2$ processor die sits face-down, partly overlapping the memory and the SOI substrate and spanning the two. The processor and memory die are connected using electrical proximity [12]. Additional memory can be located off the macrochip and accessed via optical fibers. The processor is a multi-core die with memory controllers, a cache hierarchy, and an intra-die interconnect network [10]. The details of the processor die are beyond the scope of this paper. The processor and memory die connect using electrical proximity communication [12]. The processor die also includes optical transmitters, receivers, and waveguides positioned to overlap the SOI routing substrate, and uses OPxC to connect its waveguides to those in the SOI routing substrate [24]. Power is delivered to each site from a top plate and connected using solderless spring contacts that allow chip replaceability for higher system yield [30].

An Oracle Niagara 2 [32] processor scaled for low-power operation in 2015 can operate at 5 GHz, in a 16 nm technology with 64 single-issue cores in 125 $mm^2$. This chip will require 1 W/core or 64 W/site, including the processor, the optical interfaces, memory controller and caches. This means that a 64-site macrochip will dissipate about 4 kW of power. Cooling such a package, while challenging, can be done using liquid and direct-bonded copper cold plates with microchannels to directly shunt cold water to each chip site. Today, vendors sell plates for a few hundred dollars that can cool 3 kW over a 5 cm x 5 cm area [1], so cooling 4 kW over a much larger macrochip area can be done with a similar design.

The inter-site interconnection network is designed to offer 2.56 TB/sec bandwidth into a site and 2.56 TB/sec from a site, by employing 1024 transmitters and 1024 receivers, each running at 20 Gb/sec (2.5 GB/sec). This gives a total peak aggregate bandwidth of 160 TB/sec. The waveguides each carry 16 wavelengths.

The lasers used for the silicon photonic interconnect will be located off-macrochip and brought in using optical fibers to edge connectors on the macrochip. We assume lasers capable of generating eight discrete wavelengths will be available in 2015, and each wavelength can be split to drive eight channels using power sharing. Thus, 1024 lasers will be needed to drive the entire interconnection network. Because a macrochip can support up to 2000 edge fiber connections, the optical bandwidth will be sufficient for inter-site communications, off-macrochip memory accesses, and I/O connections.

The macrochip architecture as described enables building a large-scale shared-memory multiprocessor in a single logical chip made up of 64 die. The remainder of this paper is devoted to describing five different network designs for connecting these die and, for each design, evaluating performance, power and complexity.

# 4. NETWORK ARCHITECTURES

Section 3 described a target system for the 2015 time frame. However, simulating such a large system is currently intractable. Therefore, in the following sections we describe and analyze a scaled-down system where both the compute power and network bandwidth are reduced by a factor of eight. Therefore, each site has 128 transmitters and 128 receivers, the total peak network bandwidth is 20 TB/sec, and there are 8 wavelengths per waveguide and 8 cores per site. These are detailed in tables 4 and 6.

## 4.1 Classification of Optical Network Architectures

Implementing electronic dynamic packet switching is relatively straightforward. Therefore, designers of conventional on-chip networks can choose from a wide range of multi-hop network topologies [10]. However, optical dynamic packet switching is much more difficult. A switched optical network is either circuit switched, using dynamically set optical switches; or requires multiple optical-electrical conversions and routing in the electronic domain. The goal of this paper is to propose and evaluate optical network architectures of various types for the macrochip. Based on the switching or routing architecture used by the network, we classify optical network topologies into the four broad categories below.

*Optical Networks Without Switching or Routing.* The only network that falls under this category is a fully connected optical point-to-point network. A fully connected electronic point-to-point network is difficult to implement due to the quadratic growth in the number of wires and connections. In the optical domain, however, we exploit WDM in silicon waveguides to reduce waveguide area and routing complexity. We show in section 6.4 that a point-to-point network is less complex than other, switch-based architectures. A point-to-point networks has almost no overhead for data transmission, but is limited to low-bandwidth and narrow datapath site-to-site optical channels.

*Circuit Switched Networks.* These architectures use a network of waveguides and optical switches with multiple host access points to interconnect compute nodes. Each compute node sets up a series of optical switches, using an independent optical/electrical path-setup network, to connect the node to the destination. No explicit arbitration among senders is required for data transmission. Depending on the topology and complexity, these networks are either blocking or non-blocking. A non-blocking network implies that a circuit established between any pair of nodes 'A' and 'B' will not block a circuit between any other pair of nodes 'C' and 'D'. We adapt the architecture of the optical circuit-switch torus proposed in [35] to the macrochip technology.

*Arbitration-Based Switched Optical Networks.* Arbitrated networks are fundamentally circuit switched networks but differ in the way the optical circuit is established. All sources contending for a shared data channel arbitrate for data slots prior to data transmission. The arbitration mechanism also sets up the appropriate switches for data transmission. These networks usually require a separate arbitration network for control. One proposed optical crossbar architecture uses optical token ring arbitration to access the network [40]. We adapt this network topology to the macrochip architecture. Based on our power analysis we find the crossbar architecture has high power consumption when adapted to the macrochip. We propose a two-phase arbitration-based network that has lower power requirements at the expense of lower site-to-site bandwidth than the network in [40].

*Optical Networks with Electronic Routing.* These networks use multiple hops between the source and destination nodes; however at each hop, the optical packets are converted to the electronic domain, packet switched to the appropriate output port, and then converted back to the optical domain. We modify the point-to-point network (above) to limit the number of direct connections and use a maximum of one electronic switching hop in order to provide full connectivity in the network. This facilitates higher bandwidth site-to-site optical channels compared to the point-to-point network.

The five networks chosen for the macrochip have remarkably different properties. The point-to-point network has the least overhead but is limited by low-bandwidth narrow-datapath site-to-site channels whereas the switched networks have higher bandwidth wider-data-path data channels but with higher overhead in the form of arbitration, path setup or electronic routing.

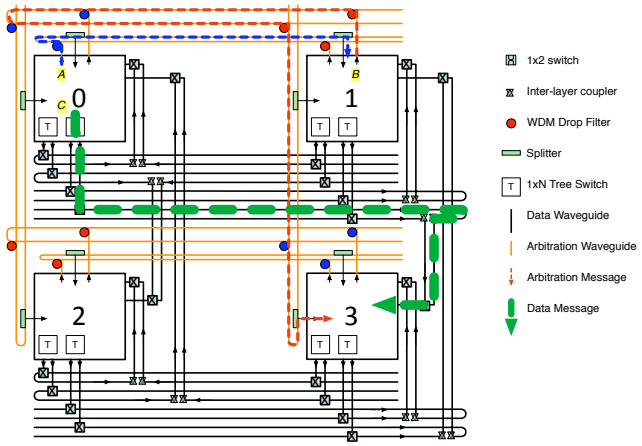
## 4.2 Statically-Routed Point-to-Point Network

In the static WDM-routed point-to-point network, each site has a dedicated optical data path to every other site. Figure 3 shows a $2 \times 2$ point-to-point network. The figure shows a packet transfer from site 0 to sites 1 and 3. Site 0 uses the same horizontal and vertical waveguide for both transfers, but uses a different wavelength for each destination, blue for site 3 and red for site 1.

The network layout consists of horizontal waveguides between the rows of the macrochip and vertical waveguides between the columns of the macrochip. The horizontal and vertical waveguides are laid on the bottom and top layers, respectively, of the SOI routing substrate, and horizontal waveguides connect to vertical waveguides using inter-layer couplers. Each vertical waveguide drops one wavelength at each site in the column. A transmitting site can communicate with any receiving site $S$ by choosing the waveguides leading to the column of site $S$ and the wavelength that is then dropped at site $S$.

In the $8 \times 8$ macrochip configuration, each site sources 16 horizontal waveguides, each carrying 8 wavelengths of light, for a total of 128 wavelengths. At a bit-rate of 20 Gb/sec (2.5 GB/sec) per wavelength, and with 64 sites, the network has a total peak bandwidth of 20 TB/sec. Each point-to-point data channel uses two wavelengths for 5 GB/sec. The network uses 8192 total horizontal and twice as many vertical waveguides, because each vertical channel consists of two waveguides for communicating both up and down. A more detailed component count is presented in table 6.

The statically routed point-to-point network has no switching or arbitration overheads. However, the narrow 2-bit site-




*Site 0 sends to S3. First, S0 sends an arb. request in blue along its row. An arb. slot is assigned to the request. Then S1 sends the switch request in red along the column. S3 sets the input switch. Finally, S0 sends using the data waveguide going to S3.*

**Figure 4: Two-Phase Arbitrated Network**

to-site data-path is a potential performance limiter. This observation motivates the investigation of other types of networks with higher point-to-point bandwidth.

## 4.3 Two-Phase Arbitration-Based Switched Optical Network

*Data Network Topology.* This architecture uses shared data channels between sites to increase the bandwidth per site-to-site logical connection. Figure 4 shows a small $2 \times 2$ version of the macrochip with this network. All sites in a row of the macrochip share an optical data channel to one other site. Thus, an $8 \times 8$ macrochip has 512 shared channels. Sites that share a channel are said to belong to the same arbitration domain. Each site connects to 64 shared horizontal channels, each comprised of two waveguides with 8 wavelengths per waveguide. Each of the waveguides in a row is coupled to a vertical waveguide using an inter-layer coupler, and this vertical waveguide is connected to one destination site. Each shared channel is 40 GB/sec and 16 bits wide. Due to the physical restriction that only one transmitter at a time can feed directly into a waveguide, a shared waveguide is implemented using broadband switches at the feeding points. Each switch hop along the path of an optical signal causes 1 dB loss. To minimize the optical loss through the switches, each waveguide is implemented as two parallel waveguide segments. We refer to the pair of segments that form a logical waveguide as simply waveguides in this section.

In order to minimize the number of transmitters in the network, each site is limited to transmitting to only one site in any column and hence can sustain, at most, 8 simultaneous 40 GB/sec transmissions to different columns. Each site uses a tree of broadband switches per column, shown as "T" in figure 4, to choose a destination in that column. This results in a maximum of 7 switch hops between any source and destination and hence only a 7 dB loss. The use of switch trees to minimize the number of switches on a path can result in contention when a site has multiple packets to send to a single column. An alternate version of this network uses twice the amount of laser power and double the number of switch trees, to reduce the potential for contention. This alternative, called "2-phase Arb ALT" in the remainder of the paper, is evaluated along with the base design in section 6.

*Arbitration Network Topology.* Figure 4 illustrates the topology and operation of the arbitration network. Due to the mesochronous properties of the macrochip [25], we can employ a completely distributed arbitration mechanism to reduce the arbitration overheads. Each node in an arbitration domain makes the same decision for every arbitration request as made by all of the sites in that domain, at the same time.

Data transmission from any site 'A' to any site 'B' requires contention resolution among the sites in the arbitration domain and a control operation to set the appropriate switches for data transmission. The arbitration network consists of a request waveguide for each row for contention resolution, and a notification waveguide for each column to set the appropriate destination switches. Each site can transmit on the request and the notification waveguides using pre-assigned wavelengths. In addition, each site is a column manager of its column in its arbitration domain. The arbitration wavelengths are assigned in cyclic fashion as shown in figure 4, to enable WDM and reduce the number of arbitration waveguides. The request waveguides are snooped by all sites in the corresponding row, and the notification waveguides are snooped by the all sites in the corresponding column. Snooping requires higher input power proportional to the number of sites snooping the waveguide. In this case the arbitration waveguides need to be sourced with $7\times$ more laser power. Since the arbitration network uses a small number of lasers (in comparison to the data network), the increase in power is negligible. The increase in area due to arbitration is also small, because it adds only 16 horizontal and 8 vertical waveguides to the data network.

*Arbitration Mechanism.* Sites arbitrate for access to a destination in arbitration slots. Each arbitration slot is about 0.4 ns, enough to transmit an arbitration request. Multiple arbitration requests are pipelined to improve bandwidth utilization on the optical data channels. The data channels are are also time-slotted. The size of each slot is variable but an integral multiple of a basic slot size. Data transmission for every packet requires the following steps:

Phase 1:

1) All requests during an arbitration interval are posted to the request network.

2) Each site in the arbitration domain sees the request after the propagation time.

3) Each site maintains a round-robin counter for every destination node. The counter specifies the order of assignment of successive slots to the requesters. All sites in the arbitration domain assign the same slot $T_r$ to a requester.

Phase 2:

1) At arbitration slot $T_a$, $T_a < T_r$, the column manager for the destination column sends a switch request on the notification waveguide in the pre-assigned wavelength.

2) Prior to slot $T_r$, all the row sites in the arbitration domain set their respective broadband switches, the destination site sets its input-select switch, and the source site transmits the data over the optical circuit to the destination.

The switch request notification is timed to accommodate

the switch delay. When an optical switch is set, it remains in that position until another switch request notification changes it.

This network provides a wider 16-bit 40 GB/sec data path but incurs arbitration and switch-delay overheads. A detailed component count for the data and arbitration networks is given in table 6.

## 4.4 Token-Ring-Based Optical Crossbar

The Corona architecture is an optical crossbar with token-ring arbitration, using a ring topology with no waveguide crossings [40]. Each site or cluster in the Corona architecture has a dedicated waveguide bundle shared by all sites transmitting data to that site. Access to the shared bundle is arbitrated using a token ring. One token for each destination is propagated on a token bus by that site. A site that needs to transmit data to the destination diverts the token by tuning its receiver on the token channel to the correct wavelength. On completing data transmission, the site releases the token by re-injecting a light pulse into the token bus. We adapted the Corona network to the macrochip by using the bottom substrate to route both the token and data waveguides and the modulators.

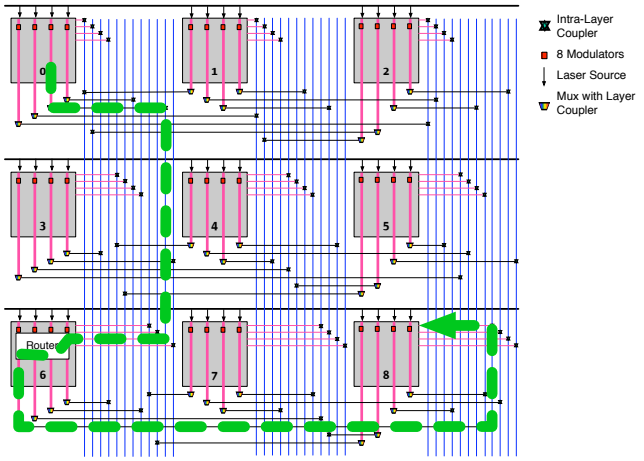
The Corona network architecture uses 64-way WDM and hence each wavelength of light passes by 4,096 modulator rings. Each modulator when tuned off-resonance causes a 0.1 dB loss to the signal causing a 409.6 dB loss along the path. Even a WDM factor of 8 as used by the macrochip, causes a 51.2 dB loss. In order to limit the power loss due to the rings to 12.8 dB, we reduced the WDM factor to 2 and increased the number of waveguides by a factor of 4.

The macrochip dimensions are 10× the dimensions of the chip proposed for Corona. We scaled the 8 cycle round-trip token latency specified in [40] based on dimension. In our adaptation, it takes 80 cycles for a token to complete a round-trip.

## 4.5 Circuit Switched Network

A circuit switched network requires a path setup procedure to set the appropriate optical switches along the path from the source to the destination. The circuit-switched-style network in [35] uses an electronic torus overlaying an optical torus. The optical torus uses a $4 \times 4$ optical switch at each switching point. This optical torus network enables non-blocking operation. The electronic network is a low-bandwidth packet-switched network and is used to set up an optical path from the source to the destination. Each switching point in the electronic network is attached to the $4 \times 4$ optical switch that it controls. To establish an optical circuit, a source node initiates a path-setup packet on the electronic network from its gateway; this setup packet is routed through multiple switch points to the destination. At each switch point the router sets the corresponding optical switch and routes the packet towards the destination site. Similarly, a post-communication path-breakdown procedure is followed to tear down the optical circuit.

We adapted the network topology and architecture in [35] for the macrochip. Adding an inter-site electronic network for path setup on the macrochip would require an active substrate with long running wires/lines, complicating the design of the macrochip. Therefore, in our adaptation, we used a low bandwidth optical network for path setup instead of an electronic network.




*S0 sends to S8, forwarded through S6. S0 transmits in all wavelengths in the marked channel. S6 receives, converts, routes electronically, then re-sends the packet as an optical signal to S8.*

**Figure 5: 3×3 Limited Point to Point Network Topology**

We perform the path setup, acknowledgment and path tear-down using the optical control network. We added the the additional routers required for non-blocking operation to the macrochip sites. The optical torus adapted to the macrochip is similar to that in [35] and is completely routed in the lower substrate to reduce the coupler loses. Each macrochip site sources 16 waveguides with 8 wavelengths per waveguide. This requires 64 waveguide loops between each row of sites in the macrochip, resulting in 50% fewer waveguides compared to the WDM point-to-point network.

This network however has two major drawbacks. First, it requires a large number of waveguide crossings. Waveguide crossings induce significant crosstalk into a waveguide, especially when one waveguide is crossed at multiple points [7]. Because we do not know the crosstalk loss assumptions used in [35], we assume negligible crosstalk at waveguide crossings for the macrochip adaptation of this network. Second, there are similar questions about the power loss due to broadband switches. As shown in Table 1, we project that a broadband $1 \times 2$ switch will cause a 1 dB power loss. For the adaptation of the circuit-switched network, we use a more aggressive power loss assumption of 0.5 dB loss per $4 \times 4$ optical switch. The worst case path in the network requires 31 optical switch hops causing approximately 15 dB loss to the input signal, requiring an approximate 30x increase in the laser power. More detailed component counts and power estimates are shown in tables 6 and 5.

## 4.6 Limited Point-to-Point Network with Electronic Routing

Figure 5 shows a $3 \times 3$ version of the limited point-to-point network. The topology of this network is similar to that of the point-to-point network. Each site has direct optical connectivity using a separate waveguide to every site in its row and its column, called row-peers and column-peers respectively. This provides a 20 GB/sec direct optical channel to each of its row and column peers and a total of 20 TB/sec peak network bandwidth. In this arrangement, there are no direct optical channels to the other sites - the sites that

are not in the row or column of the source site. Electronic routing is used to extend the connectivity of a site to the remaining sites. This routing is implemented by adding two $7 \times 7$ electronic routers on each site: one for forwarding packets from row peers to column peers and one for forwarding from column peers to row peers. Communication between non-peer sites requires that a data packet be first forwarded to a site that is a peer to both the source and the destination. At the forwarding site, the data packet is converted to the electronic domain, forwarded to the appropriate output port, and then converted back to the optical domain to reach the destination. By the use of this configuration, each transmission requires a maximum of one intermediate O-E/E-O conversion. Table 6 gives the component count for this network architecture. The routers used in the sites are assumed to have a latency of one cycle.

# 5. SIMULATION METHODOLOGY

We performed a detailed evaluation of the five network architectures for the macrochip using two kinds of workloads: application kernels and synthetic benchmarks. We used five shared memory application benchmarks, two from the SPLASH-2 suite and three from the PARSEC suite. The benchmarks and the data sets used are listed in Table 2. The SPLASH-2 benchmarks were compiled with the Sun Studio 12 compiler suite with `-fast` optimization and the PARSEC benchmarks were compiled using g++ version 3.4.3 with `-O3` optimization for the UltraSparc T2+ processor.

The four synthetic benchmarks were chosen to represent a range of traffic patterns. In the butterfly and transpose pattern, each site only sends to one unique destination, whereas in the nearest-neighbor pattern, each site communicates with four neighbors. Table 3 lists and describes the synthetic benchmarks. These benchmarks were driven by two types of coherence mixes: Less Sharing (LS) and More Sharing (MS). In the LS mix, 90% of coherence requests have no sharers for the cache block, while in the MS mix, 40% of requests have three sharers. All of the synthetic benchmarks are driven at a rate equivalent to an L2 cache miss rate of 4% per instruction.

Table 4 shows the simulated macrochip configuration. We reduced the number of cores per site from 64, in the configuration proposed in section 3, to 8 to make the simulation more tractable. Accordingly we reduced the total network bandwidth by 8 times, to a peak of 20 TB/sec for all the networks. We also used a 256 KB cache, shared by all cores on the site, to suit the data set sizes of the applications. The optical-fiber-connected main memory is not modeled in detail. We leave the study of effect of main memory technologies on performance to future work.

Our simulation infrastructure consists of two parts. The macrochip CPU simulator is an instruction-trace driven multiprocessor core/cache simulator that models an MOESI coherence protocol. The CPU simulator generates L2 miss traffic along with detailed coherence information for all 512 cores. The network simulator models all five network architectures and is driven using the coherence traffic generated by the CPU simulator. The network model simulates all necessary network messages required by the coherence protocol to satisfy a coherence request. We model finite MSHRs, network I/O buffers and virtual channels for cache coherence operations. To keep simulation time manageable, we do not model the intricate details of the cache coherency protocol.

| Benchmark | Size | Suite |
|---|---|---|
| Radix | 32 M integers | Splash-2 |
| Barnes | 16 K particles | Splash-2 |
| Blackscholes | simlarge | PARSEC |
| Fluidanimate (forces) | simlarge | PARSEC |
| Fluidanimate (densities) | simlarge | PARSEC |
| Swaptions | simlarge | PARSEC |

**Table 2: Benchmarks Used**

| Pattern | Destination ID |
|---|---|
| Uniform | Random for every packet |
| Transpose | First half of the bits in source site-id are swapped to second half |
| Butterfly | Swap LSB and MSB of source site-id |
| Neighbor | Source (x,y) randomly selects from (x,y-1), (x,y+1), (x-1,y), (x+1,y) |

**Table 3: Synthetic Patterns**

# 6. EVALUATION

## 6.1 Latency and Throughput Analysis

We performed tests with 64-byte raw data packets to determine the maximum throughput of each network. In these tests, we compared the five networks using the synthetic patterns listed in table 3. The input driver for these simulations probabilistically generates data packets in a specific communication pattern. Each data packet is 64 bytes, to represent a cache line transfer. We measured the latency per packet, defined as the time elapsed from when the packet was generated to when the packet was received by the destination site. The bandwidth per site and the total peak bandwidth used for each network are as shown in table 4. Figure 6 shows the latency response for all the networks on the synthetic patterns. The latency per packet increases with load. The vertical asymptote of the latency response curve shows the maximum sustainable bandwidth for that network.

The point-to-point network performs best on the uniform random pattern, shown in figure 6, because it has no arbitration or path setup overheads. The sustained bandwidth scales well up to 95% of peak. The token-ring network is limited by token arbitration overheads, and so scales only to 40% of peak. The limited point-to-point network sustains up to 47% of the peak bandwidth. This is because 75% of the traffic is forwarded through another site. This causes each packet to use two point-to-point links, reducing the effective bandwidth by about 50%. The circuit-switched network has high path setup overhead at this message size and has the lowest sustained bandwidth, only 2.5% of peak. The original two-phase arbitration network wastes bandwidth due to

| | |
|---|---|
| Number of sites | 64 |
| Shared L2 Cache per site | 256 KB |
| Bandwidth per site | 320 GB/sec |
| Total peak bandwidth | 20 TB/sec |
| Cores per site | 8 |
| Threads per core | 1 |
| FPU per core | 1 |

**Table 4: Simulated Macrochip Configuration**

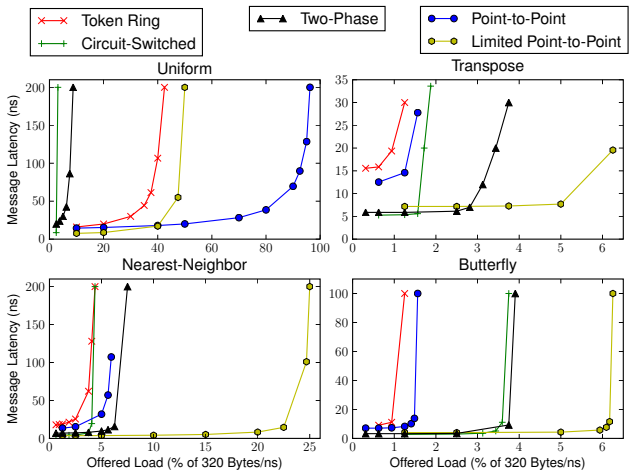



**Figure 6: Latency vs. Offered Load for Four Message Patterns**

switch tree contention, resulting in a sustained bandwidth of 7.5% of peak. As discussed in section 4.3, this contention can be resolved by either increasing the number of transmitters or by using a more logic-intensive arbitration protocol that provides an efficient assignment of data slots. One such alternate configuration is evaluated further in section 6.2.

In the butterfly and transpose traffic patterns, each site communicates with only one other site. These patterns limit the throughput of the point-to-point network to 5 GB/sec because they use only one of the available 64 links. The token-ring network throughput on these patterns reaches a maximum below 1% of peak. It takes only one cycle to transmit a 64-byte packet, but requires 80 cycles to reacquire the token. Because the communication pattern is one-to-one, the token latency reduces the bandwidth utilization.

In the nearest-neighbor pattern, each site communicates with four other sites at a time. This pattern maps well to the row/column connectivity of the limited point-to-point network. No packet requires any intermediate router hops. Each site uses four of its 20 GB/sec point-to-point links with no contention or overheads and the network achieves a sustained bandwidth of up to 25% of the peak.

## 6.2 Benchmark Performance Analysis

Figure 7 shows the speedup of each network relative to the circuit-switched network, which had the lowest performance. The five columns on the right show results for the synthetic benchmarks shown in table 3, and the six on the left show results for the application benchmarks in table 2. Figure 8 shows corresponding data for average latency per coherence operation.

In figure 7 we see that the networks that require arbitration perform poorly on the transpose and butterfly patterns. In these patterns, each site only sends to one other site. As discussed above, the token-ring network's token latency results in low bandwidth utilization. In comparison, the point-to-point network has lower data-path width, but avoids this setup overhead, providing a higher overall speedup and a reduction in latency. There is a relatively small variation in speedup on the butterfly pattern because 50% of the communication is intra-node, and we have modeled intra-node traffic as a single cycle loop-back link.

| Network Type | Power Loss Factor | Laser Power (W) |
|---|---|---|
| Token-Ring | 19× | 155 |
| Point-to-Point | 1× | 8 |
| Circuit-Switched | 30× | 245 |
| Limited Pt.-to-Pt. | 1× | 8 |
| Two-Phase: | | |
| Data | 5× | 41 |
| Data (ALT) | 4× | 65.5 |
| Arbitration | 8× | 1 |

**Table 5: Network Optical Power**

The "MS" sharing mix consists of a large number of invalidate and acknowledgment packets which are small in size, and so the arbitration overhead dominates performance. Due to this, the point-to-point network performance on this mix, regardless of message pattern, is at least 4.5× better than the arbitrated networks. Therefore, the figures show results for only one pattern with the "MS" mix.

The two-phase arbitration network has lower overhead than the token-ring and circuit-switched networks, and thus provides a speedup of at least 1.6× compared to these networks. For the reasons discussed in section 6.1, the limited point-to-point network performs better than any other network on the nearest-neighbor pattern and has a speedup of 5× compared to the circuit-switched network. The all-to-all pattern causes the maximum network load of all the synthetic benchmarks and causes more contention for the input switch tree in the two-phase arbitration network. The "Two-phase (ALT)" alternate design discussed above improves performance by 1.4× by reducing this contention.

The left six bars in figure 7 show results for the application benchmark kernels normalized to the circuit-switched network. The point-to-point network consistently outperforms the other networks on the application benchmarks, with a maximum speedup of 8.3× over the circuit-switched network and 3× over the token-ring network on the swaptions benchmark. This follows from the latency per coherence operation shown in figure 8. The point-to-point network has a maximum latency of only 54 nanoseconds for the application benchmarks and 100 nanoseconds on the synthetic benchmarks.

Despite relatively low data-path width per link, the point-to-point network is a better choice for the small, latency-sensitive messages in cache coherence traffic. This is due to the absence of arbitration or path setup overhead.

The token-ring and circuit-switched torus networks have been shown to provide good performance when used as an intra-chip network, but when scaled to the dimensions of the macrochip, token propagation and path-setup latency hurt performance on the class of applications we evaluated.

The Barnes benchmark shows relatively low speedups. This benchmark does not stress any of the networks, due to a relatively low L2 cache miss rate.

## 6.3 Power Analysis

In this section, we discuss power estimates for each network. The static power consumed by the networks, shown in table 5, is calculated using the component counts shown

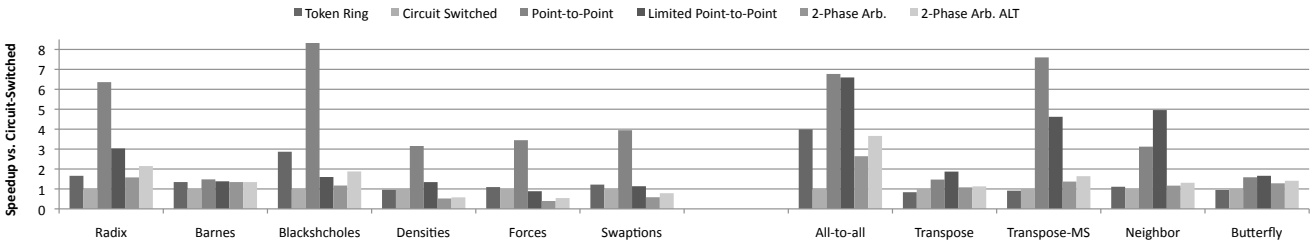



**Figure 7: Speedup For Benchmarks and Synthetic Message Patterns, Normalized to Circuit-Switched Network**

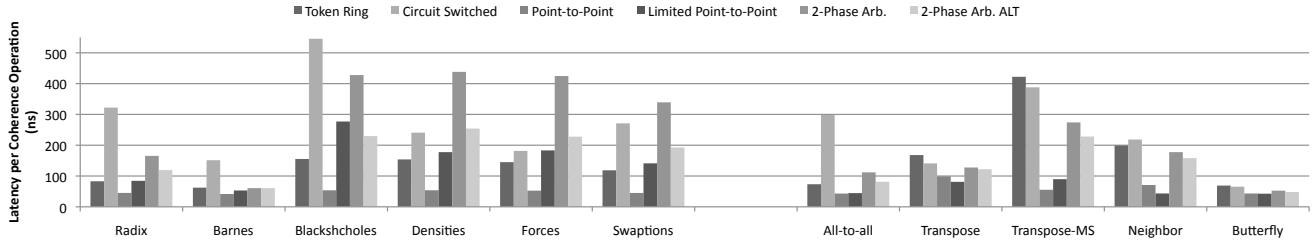



**Figure 8: Latency per Coherence Operation**

in table 6 and the per-component static energy from table 1. The dynamic power is calculated from the transmitter and receiver energies shown in table 1.

When the optical signal passes through switches, non-resonant modulators, or is split into two waveguides, it suffers signal loss. The total loss is listed as the power loss factor in table 5. To compensate for this signal loss, the laser power must be increased by the loss factor. The basic laser power is assumed to be 1 mW per wavelength. For example, in the token-ring network, which has two wavelengths per waveguide, each wavelength passes through 128 modulators, which have an off-resonance coupling loss of 0.1 dB. This causes a total signal loss of 12.8 dB, or a 19× power loss factor.

The power estimates for the limited point-to-point network include dynamic router power. The energy required by the routers to switch a single byte was conservatively assumed to be 60 picojoules [34]. Figure 9 shows the energy consumed by the routers for each workload as a percentage of total energy. The energy consumed by routers was a maximum of 17% for the synthetic benchmarks, and a maximum of 10.4% for the application benchmarks.

Figure 10 shows the energy-delay product (EDP) for each network on each workload. This graph is normalized to the point-to-point network, which has the lowest power requirements. On all but one of the application benchmarks, the point-to-point network has more than 100× lower EDP than the arbitrated or circuit-switched networks. The point-to-point network also has up to 26× lower EDP than the limited point-to-point network. The alternate configuration of the two-phase arbitration network is an improvement on four of the six application benchmarks. On the Blackscholes kernel, it has an EDP 62% of the base two-phase configuration, but on the Barnes kernel, it has an EDP of 160% of the base. The point-to-point network not only performs well, but has superior EDP when compared to the other networks.

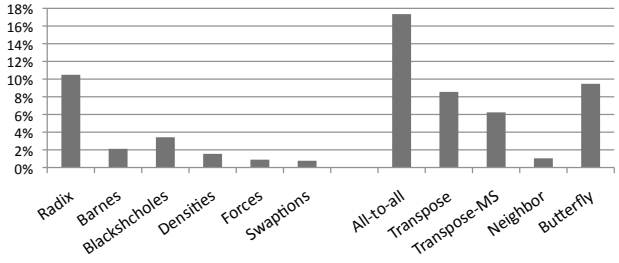



**Figure 9: Energy Used By Routers in Limited Point-to-Point Network as a Percentage of Total**

## 6.4 Complexity Analysis

An important factor to be considered in evaluating silicon-photonic network architectures is scalability in terms of the number and types of the individual components. In this section we make a qualitative assessment of network complexity based on those factors.

Table 6 shows total component counts for each of the six networks. The number of physical waveguides used by the token ring network is only 8192. However, since every waveguide is routed along every row, it adds to the total area for waveguides and hence is shown as 32 K waveguides in Table 6. The waveguide counts for the point-to-point, limited point-to-point and the two phase arbitration networks include both the vertical and horizontal waveguides.

As the number of wavelengths per waveguide increases with improvements in technology, the peak bandwidth for a point-to-point network can increase without increasing the number of waveguides. This is contrary to the case of electronic point-to-point networks where scalability is limited by the quadratic increase in the number of wires. The other networks analyzed in this paper require additional compo-

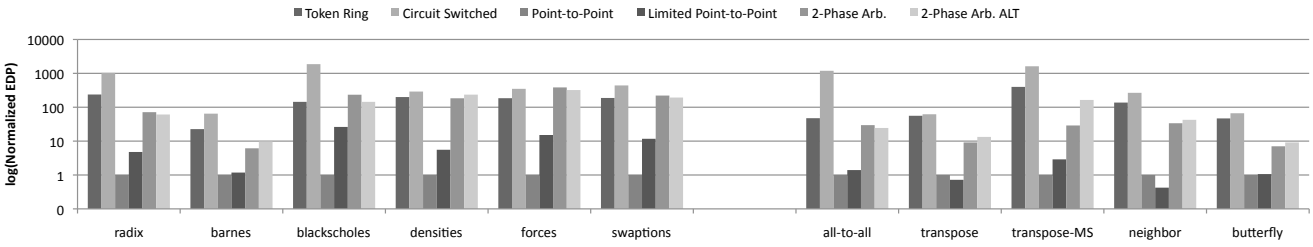



**Figure 10: Log Plot of Energy-Delay Product, Normalized to the Point-to-Point Network**

| Network Type | Tx | Rx | Wgs | Switches |
|---|---|---|---|---|
| Token-Ring | 512K | 8192 | 32K | 0 |
| Point-to-Point | 8192 | 8192 | 3072 | 0 |
| Circuit-Switched | 8192 | 8192 | 2048 | 1024[a] |
| Limited Point-to-Point | 8192 | 8192 | 3072 | 128[b] |
| Two-Phase: | | | | |
| Data | 8192 | 8192 | 4096 | 16K |
| Data (ALT) | 16384 | 8192 | 4096 | 15K |
| Arbitration | 128 | 1024 | 24 | 0 |

**Table 6: Total Optical Component Counts**

[a] $4 \times 4$ switches

[b] $7 \times 7$ electronic routers

nents, such as switches and arbitration/control networks, which increases contention and limits the scalability of these networks.

# 7. RELATED WORK

In recent years, a variety of architectures have been proposed using Silicon-photonic technology to build on-chip networks for single-die multicore systems.

The Corona architecture uses a high-speed optical crossbar with token-ring arbitration, interconnecting 64 cores on a single die [40]. Because the links on a single die are short, the token propagation latency is very low. In a multi-die system with link distances an order of magnitude larger, this propagation latency becomes a significant performance issue.

The non-blocking torus proposed in [35, 37] uses a packet-switched electronic network to setup an end-to-end optical circuit in the companion optical network. This is also an on-chip network for a single-die multicore chip. The path-setup latency in such a network causes significant delays for small transfers such as cache lines.

The Firefly network is a hierarchical network for a single die that uses an electronic network for messages local to a cluster of cores and an optical crossbar between clusters [34]. The optical crossbar is similar to that used in the Corona architecture, and will suffer the same token latency problems when scaled to a large multi-die system.

A detailed investigation of a silicon-photonic cache-coherent bus for a single multicore die is presented in [22]. The authors demonstrate that the use of silicon-photonics can offer improvements in performance, power and area over a similar all-electric bus for the shared-memory workloads they analyzed.

In contrast to the above papers, our work focuses on design and evaluation of silicon-photonic networks for large, multi-die systems. We have proposed three optical networks that have low power requirements for a multi-die system. When evaluated on a large multi-die system, the performance and power characteristics of the previously proposed intra-die designs change considerably. We show that a low-complexity point-to-point network performs better than other solutions in the context of the evaluated shared-memory workloads on a multi-die system.

# 8. CONCLUSIONS

In this paper, we describe the macrochip technology platform, which uses a silicon-photonic interconnection network to enable building a large, high-performance single logical chip with very high core counts.

To achieve high-bandwidth, low-power communication on the macrochip, we proposed three silicon-photonic network designs and adapted two promising designs from previous work on intra-chip silicon-photonic interconnects. The networks covered in this paper vary widely in communication overheads, power consumption, complexity and data-path width.

We simulated the performance of all the networks on five synthetic benchmarks and five application kernels, two from the SPLASH-2 benchmark suite and three from the PARSEC benchmark suite. We also performed a detailed power and complexity estimation for each of the networks. Based on our evaluation, we find that the static WDM point-to-point network, despite having 1/64th the per-link data-path width of some of the other networks, performs between 3 to 8 times better than other networks that have wider data-paths but larger communication overheads. Due to its simplicity, the energy-delay product of the point-to-point network is between 10 and 100 times lower than the wider data-path networks on some of the benchmarks. We have also shown that, contrary to the case in electronic networks, a silicon-photonic point-to-point network has the lowest design complexity and a high degree of scalability. Future work will evaluate network architectures for message passing workloads and the performance impacts of different memory technologies and site architectures on the macrochip.

## Acknowledgment

We would like to thank the ISCA reviewers, in particular Al Davis, for comments that improved the quality of the final paper.

This material is based upon work supported, in part, by DARPA under Agreement No. HR0011-08-09-0001. The authors thank Dr. Jag Shah of DARPA MTO for his inspi-

ration and support of this program. The views, opinions, and/or findings contained in this paper are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense. Approved for public release, distribution unlimited.

# 9. REFERENCES

[1] High performance coolers. `http://www.electrovac.com/sprache2/n160352/i229935.html`.

[2] K. Asanovic, R. Bodik, et al. The landscape of parallel computing research: A view from Berkeley. Technical Report UCB/EECS-2006-183, EECS, UC Berkeley, 2006.

[3] R. Beausoleil, J. Ahn, et al. A nanophotonic interconnect for high-performance many-core computation. In *HOTI*, Aug. 2008.

[4] S. Bell, B. Edwards, et al. Tile64 - processor: A 64-core SoC with mesh interconnect. In *ISSCC*, 2008.

[5] K. Bernstein, P. Andry, et al. Interconnects in the third dimension: design challenges for 3D ICs. In *DAC*, 2007.

[6] C. Bienia, S. Kumar, et al. The PARSEC benchmark suite: Characterization and architectural implications. In *PACT*, 2008.

[7] W. Bogaerts, P. Dumon, et al. Low-loss, low-cross-talk crossings for silicon-on-insulator nanophotonic waveguides. *Opt. Lett.*, 32(19), 2007.

[8] M. Chang, J. Cong, et al. CMP network-on-chip overlaid with multi-band RF-interconnect. In *HPCA*, 2008.

[9] J. Cunningham, X. Zheng, et al. Optical proximity communication in packaged SiPhotonics. In *IEEE Group IV Photonics*, 2008.

[10] W. Dally, B. Towles, et al. *Principles and Practices of Interconnection Networks*. Morgan Kaufman, 2004.

[11] R. Das, S. Eachempati, et al. Design and evaluation of a hierarchical on-chip interconnect for next-generation cmps. In *HPCA*, Feb. 2009.

[12] R. Drost, R. Hopkins, et al. Proximity communication. *JSSC*, 2004.

[13] R. Espinola, M. Tsai, et al. Fast and low-power thermooptic switch on thin silicon-on-insulator. *Photonics Technology Letters, IEEE*, 2003.

[14] R. Ho, J. Lexau, et al. Circuits for silicon photonics on a "macrochip". In *ASSCC*, Nov. 2009.

[15] R. Ho, K. Mai, et al. The future of wires. *Proceedings of the IEEE*, 89(4), Apr 2001.

[16] R. Ho, I. Ono, et al. High-speed and low-energy capacitively-driven on-chip wires. In *ISSCC*, Feb. 2007.

[17] Intel. First tick, now tock: Next generation Intel microarchitecture. techreport, `http://intel.com/`, 2009.

[18] A. Jose, K. Shepard, et al. Distributed loss-compensation techniques for energy-efficient low-latency on-chip communication. *JSSC*, 2007.

[19] K. Kanda, D. Antono, et al. 1.27Gb/s/pin 3mW/pin wireless superconnect (WSC) interface scheme. In *ISSCC*, 2003.

[20] B. Kim, V. Stojanovic, et al. A 4Gb/s/ch 356 fJ/b 10mm equalized on-chip interconnect with nonlinear charge-injecting transmit filter and transimpedance receiver in 90nm CMOS. In *ISSCC*, 2009.

[21] J. Kim, W. Dally, et al. Microarchitecture of a high-radix router. In *ISCA*, 2005.

[22] N. Kırman, M. Kırman, et al. Leveraging optical technology in future bus-based chip multiprocessors. In *MICRO*, 2006.

[23] P. Kogge, K. Bergman, et al. Exascale computing study: Technology challenges in achieving exascale systems, 2008.

[24] A. V. Krishnamoorthy, J. E. Cunningham, et al. Optical proximity communication with passively aligned silicon photonic chips. *J. Quant. Elec.*, 45(4), 2009.

[25] A. V. Krishnamoorthy, R. Ho, et al. Computer systems based on silicon photonic interconnects. *Proc. IEEE*, 97(7), 2009.

[26] R. Kumar, V. Zyuban, et al. Interconnections in multi-core architectures: understanding mechanisms, overheads and scaling. In *ISCA*, 2005.

[27] B. Lee, A. Biberman, et al. All-optical comb switch for multiwavelength message routing in silicon photonic networks. *IEEE Photon. Technol. Lett*, 20(10):767–769, 2008.

[28] G. Masanovic, V. Passaro, et al. Coupling to nanophotonic waveguides using a dual grating-assisted directional coupler. *IEEE Proc. Optoelectronics*, 152(1), 2005.

[29] E. Mensink, D. Schinkel, et al. A 0.28pJ/b 2Gb/s/ch transceiver in 90nm CMOS for 10mm on-chip interconnects. In *ISSCC*, 2007.

[30] J. Mitchell, J. Cunningham, et al. Integrating novel packaging technologies for large scale computer systems. In *InterPACK*. ASME, July 2009.

[31] B. Murphy. Cost-size optima of monolithic integrated circuits. *Proc. IEEE*, 52(12), 1964.

[32] U. Nawathe, M. Hassan, et al. An 8-core 64-thread 64b power-efficient SPARC SoC. In *ISSCC*, Feb. 2007.

[33] J. Orcutt, A. Khilo, et al. Demonstration of an electronic photonic integrated circuit in a commercial scaled bulk cmos process. In *Conf. on Lasers and Electro-Optics*, 2008.

[34] Y. Pan, P. Kumar, et al. Firefly: illuminating future network-on-chip with nanophotonics. *ISCA*, 2009.

[35] M. Petracca, B. G. Lee, et al. Design exploration of optical interconnection networks for chip multiprocessors. In *HOTI*, 2008.

[36] Semiconductor Industries Association. International technology roadmap for semiconductors. webpage, 2008. `http://www.itrs.net/Links/2008ITRS/Home2008.htm`.

[37] A. Shacham, B. Lee, et al. Photonic NoC for DMA communications in chip multiprocessors. In *HOTI*, 2007.

[38] C. Stapper. On murphy's yield integral. *IEEE Trans. Semiconductor Manufacturing*, 4(4), 1991.

[39] S. Vangal, J. Howard, et al. An 80-tile sub-100-W TeraFLOPS processor in 65-nm CMOS. *JSSC*, 2008.

[40] D. Vantrease, R. Schreiber, et al. Corona: System implications of emerging nanophotonic technology. In *ISCA*, June 2008.

[41] L. Vivien, G. Maire, et al. A high efficiency silicon nitride grating coupler. In *IEEE Group IV Photonics*, pages 1–3, Sept. 2007.

[42] S. Woo, M. Ohara, et al. The SPLASH-2 programs: Characterization and methodological considerations. In *ISCA*, 1995.

[43] X. Zheng, P. Koka, et al. Silicon photonic WDM point-to-point network for multi-chip processor interconnects. In *IEEE Group IV Photonics*, 2008.

[44] X. Zheng, J. Lexau, et al. Ultra-low energy all-CMOS modulator integrated with driver. *Optics Express*, 18(3):3059–3070, 2010.

[45] X. Zheng, F. Liu, et al. An sub-picojoule-per-bit CMOS photonic receiver for densely integrated systems. *Optics Express*, 18(1):204–211, 2010.

[46] X. Zheng, I. Shubin, et al. A tunable $1 \times 4$ silicon CMOS photonic wavelength multiplexer/demultiplexer for dense optical interconnects. *Optics Express*, 18(5):5151–5160, 2010.