Leveraging Optical Technology in Future Bus-based Chip Multiprocessors

Nevin Kırman Meyrem Kırman Rajeev K. Dokania José F. Martínez
Alyssa B. Apsel Matthew A. Watkins David H. Albonesi

Computer Systems Laboratory
Cornell University
Ithaca, NY 14853 USA

http://csl.cornell.edu/

ABSTRACT

Although silicon optical technology is still in its formative stages, and the more near-term application is chip-to-chip communication, rapid advances have been made in the development of on-chip optical interconnects. In this paper, we investigate the integration of CMOS-compatible optical technology to on-chip cache-coherent buses in future CMPs.

While not exhaustive, our investigation yields a hierarchical opto-electrical system that exploits the advantages of optical technology while abiding by projected limitations. Our evaluation shows that, for the applications considered, compared to an aggressive all-electrical bus of similar power and area, significant performance improvements can be achieved using an opto-electrical bus. This performance improvement is largely dependent on the application's bandwidth demand and on the number of implemented wavelengths per optical waveguide. We also present a number of critical areas for future work that we discover in the course of our research.

1 INTRODUCTION

Current research and technology trends indicate that future chip multiprocessors (CMPs) may comprise tens or even hundreds of processing elements. An important hurdle towards attaining this scale, however, is the need to feed data to such large numbers of on-chip cores. This can only be achieved if architecture and technology developments provide sufficient chip-to-chip and on-chip communication performance to these future generations of CMPs.

Optical technology [17, 33, 60] and 3D integration [46, 49] are two potential solutions to current and projected limitations in *chip-to-chip* communication performance. Still, *on-chip* communication faces considerable technological and architectural challenges of its own. On the one hand, global on-chip interconnects do not scale well with technology [23, 25]. Although delay-optimized repeater insertion [2, 23, 47] and proper wire sizing [22] can keep the delay nearly constant, this comes at the expense of power [23, 27] and active area, as well as a reduction in wire count (and thus bandwidth). Techniques for optimizing the power-delay product have been developed [3, 27], but unfortunately their most obvious shortcoming is that neither power nor latency are optimal. This, combined with various other technological issues such as manufacturability, conductivity, crosstalk, etc., constitute important roadblocks for future electrical interconnects [25]. On the other hand, to date there is no clear consensus on the architecture of on-chip core interconnects. Although well-understood solutions exist for off-chip interconnects [12, 18, 31, 32, 53], the on-chip power, area, connectivity, and latency constraints make it challenging to port those solutions to the context of CMPs.

Whereas ten years ago the electrical-optical translation costs and CMOS incompatibility were viewed as insurmountable barriers for the use of optics in on-chip communication, today the outlook is dramatically more optimistic. Due to rapid progress in the past five years in CMOS-compatible detectors [58], modulators [1], and even light sources [52], the latest ITRS entertains on-chip optical interconnects as a potential replacement for global wires by 2013 [25]. In global signaling applications, optical interconnects have the potential to fare favorably against their electrical counterparts due to their high speed, high bandwidth, low on-chip power, good electrical isolation, low electromagnetic interference, and other benefits [38]. Although several efforts have attempted to identify under what conditions optics will be favorable over on-chip electrical signaling [13, 15, 21, 28, 29, 40, 42, 43], these studies have been limited in scope to clock distribution networks and comparisons of point-to-point signaling. Although the technology is admittedly still in its formative stages, there is now enough understanding and data regarding on-chip, CMOS-compatible, optical components to consider the broader architectural trade-offs in designing an on-chip optical network for future high performance microprocessors.

In this paper, we investigate the potential of optical technology as a low-latency, high-bandwidth shared bus supporting snoopy cache coherence in future CMPs. We discuss possible optical bus organizations in terms of power, scalability, architectural advantages, and other implementation issues, as well as the implications on the coherence protocol. Through a carefully projected case study for a 32nm CMP, we conduct the first evaluation of on-chip optical buses for this application. This initial step yields insights into the advantages and current limitations of the technology to catalyze future interdisciplinary work.

The rest of the paper is organized as follows: In Section 2, we review the components of an on-chip optical transmission system. We discuss the microarchitecture of our CMP at 32nm technology and the design of our optical-based shared bus in Section 3. Evaluation results are presented in Section 4. In Section 5 we discuss some of the main technical challenges going forward for optical interconnects in CMP applications. We present related work in Section 6, and our concluding remarks in Section 7.

2 OPTICAL TECHNOLOGY OVERVIEW

In this work we consider on-chip modulator-based optical transmission (Figure 1), which comprises three major components: a transmitter, a waveguide, and a receiver. We briefly describe each component, and discuss technology trends in order to estimate the specifications of future designs. We propose one such design later in Section 3.

2.1 Transmitter

Optical transmission requires a laser source, a modulator, and a modulator driver (electrical) circuit. The laser source provides light to the modulator, which transduces electrical

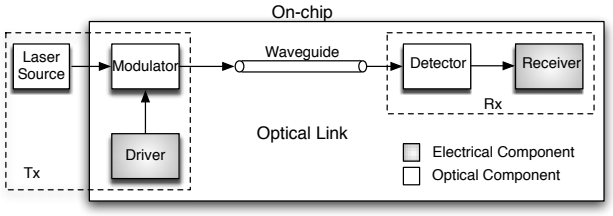


Figure 1: Simplified diagram showing the main components involved in on-chip optical transmission. Tx and Rx stand for transmitter and receiver, respectively.

information (supplied by the modulator driver) into a modulated optical signal.

While both off- and on-chip laser sources are feasible, in this work we opt for an off-chip laser source because of its greater on-chip power, area, and cost savings. As the light enters the chip, optical splitters and waveguides (not shown in Figure 1) route it to the different modulators used for actual data transmission. These distribution paths are a source of signal losses (Table 4).

The modulator translates the modulator driver's electrical information into a modulated optical signal. High-speed electro-optic modulators are designed such that injection of an electrical signal changes the refractive index or the absorption coefficient of an optical path. Among different types of proposed modulators [4, 34, 35, 48], the most recent optical resonator-based implementations are preferable for integrated circuit design, due to their low operating voltage and compact size [4]. We assume this type of modulator in our work.

Modulators are the optical equivalent of electrical switches (or transistors acting as such). Their performance in part is dependent on the on-to-off light intensity ratio, called the *extinction ratio*, which is dependent upon the strength of the electrical input signal. Higher extinction ratio is better for proper signal detection. A poor one may cause transmission errors in the channel. This ratio also puts constraints on the number of transmitters that can time-share the same wavelength on the same channel. An extinction ratio greater than 10dB has been recently reported with high input signal swing [1].

Modulator size is another important criterion for integrated applications. There has been significant recent activity towards realizing compact-sized modulators. Already 10μm ring-modulators (circularly shaped) have been proposed [1], and their size is likely to be reduced with each successive generation, albeit bounded by lithographic process and bending curvature limitations.

The modulator driver consists of a series of inverter stages driving the modulator's capacitive load. A smaller capacitance will improve the power and latency specifications of the overall transmitter, thereby requiring fewer stages. We assume a modulator capacitance of 50fF, even though it is expected to get smaller with technology improvements.

2.2 Waveguide

Waveguides are the paths through which light is routed. The refractive index of the waveguide material has a significant impact on optical interconnect bandwidth, latency, and area. For on-chip applications, silicon (Si) and polymer are the most promising materials. Some of the most relevant features of silicon and polymer waveguides are given in Table 1.

The smaller refractive index of polymer waveguides results in higher propagation speed. On the other hand, polymer waveguides require a larger pitch than Si, which reduces bandwidth density (the number of bits that can be transmitted per unit surface area).

For integrated applications, an additional disadvantage of polymer waveguides is the lack of a compact modulator. Although modulators exist for both silicon [1] and polymer waveguides [45], polymer-based modulators are bulky, and require high voltage drive for high frequency operation. These drawbacks limit their applicability to on-chip optical links.

Waveguide	**Si**	**Polymer**
Refractive index	3.5	1.5
Width (μm)	0.5	5
Separation (μm)	5	20
Pitch (μm)	5.5	25
Time of flight (ps/mm)	10.45	4.93
Loss (dB/cm)	1.3	1

Table 1: General characteristics of silicon and polymer on-chip waveguides.

Polymer waveguides are feasible in a transmission system based on VCSELs (Vertical Cavity Surface Emitting Laser) [52], where the modulator is not required. However, a VCSEL-based solution tends to increase on-chip power with the added complexity of on-chip/flip-bonded laser sources. Also, the light is emitted vertically and must be transferred to the horizontal chip surface, which requires integrated mirrors and sophisticated lithographic technologies. For these various reasons, we choose to study systems using silicon waveguides, although we understand that with technological advances feasible options might become available with polymer waveguides.

2.3 Receiver

An optical receiver performs the optical-to-electrical conversion of the light signal. It comprises a photodetector and a trans-impedance amplifier (TIA) stage. In wave division multiplexing (WDM) applications, which involve simultaneous transmission at different wavelengths per waveguide, the receiver also requires a wave-selective filter for each received wavelength.

The photodetector that is most often proposed is a P-I-N diode [59]. The photodetector's *quantum efficiency* is an important figure of merit for the system. A high quantum efficiency means lower losses when converting optical information into electrical form. Detector size is also an important criteria for both compactness and next stage capacitance. Typically, the detector has large base capacitance and pose a design challenge for high-speed gain stages following it. For our analysis we have assumed 100fF detector capacitance [40], which is achievable even with current technologies.

The TIA stage converts photodetector current to a voltage which is thresholded by subsequent stages to digital levels [44]. To achieve high-gain and high-speed detection, an analog supply voltage higher than the digital supply voltage may be required, thereby requiring higher power. We assume a TIA supply voltage that is 20% higher than the nominal supply for our power calculations in the next section.

3 OPTO-ELECTRICAL BUS ARCHITECTURE

In this section, we explore the opportunities and challenges of building an optical bus for a particular application and technology node. Working bottom-up, we first determine a reasonable CMP organization (in terms of cores, memory hierarchy, operating frequency, etc.), using available data from ITRS and other sources. Then, we address the specifics of designing a cache-coherent network with integrated optical system components (Section 3.2).

We target a 32nm process technology, and assume a 400mm^2 die area. Assuming 10mm^2 per core+L1 at 65nm (for a stripped version of an out-of-order Power4-like core [30]), and extrapolating to 32nm, we find that 64 cores fit comfortably on the die (occupying 40% of the die area), with enough additional space to allocate L2 caches (20%), interconnect (15%), controllers for off-chip L3 cache and memory,

and other system components (25%). The area breakdown closely follows the one in [19].

We opt for sixteen L2 caches, each shared among four cores, as a compromise between the demonstrated benefits of cache sharing [18, 24, 53] and the area/power overhead [30]. Using CACTI4.1 [51], we find sixteen 2MB L2 caches to fit in the allocated area.

We reasonably assume that the use of chip-to-chip optical technology will precede its on-chip application [6], and set off-chip pin bandwidth to 256GB/s and 128GB/s to L3 and memory, respectively. The aggregate pin bandwidth is therefore 384GB/s (3Tbit/s), which is well within current industry projections for our proposal's time frame [17, 33].

We estimate that core frequency will remain approximately constant in subsequent technologies, in agreement with [9]. (For a quantitative analysis, see Appendix A.) Thus, if we reasonably assume a 4GHz core frequency at 65nm, we can set core frequency in our 32nm CMP also to 4GHz.

3.1 Optical Medium

Optical waveguides do not lend themselves gracefully to H-tree or highly angled structures that may be more common in electrical topologies, for turns and waveguide crossings may result in significant signal degradation. This is aggravated when attempting to lay out multiple waveguides for multi-bit transmission, which is the case in a typical bus. Instead, we propose to build upon a simple loop-like structure, which is much better suited to the structural characteristics of optical waveguides. In the rest of this section, we discuss the design implications of this structural choice.

The proposed loop-shaped bus comprises optical waveguides (residing on a dedicated Si layer) that encircle a large portion of the chip (Figure 2). Multiple nodes connected to the bus, each of them responsible for issuing transactions on behalf of a processor or a set of processors, are equipped with necessary transmitters and receivers to interface with the optical medium, as explained earlier (Section 2).

We assume a bus comprising a total of b address, data, and snoop response bits (and thus waveguides). We further presume the availability of w wavelengths per waveguide through wave division multiplexing (WDM) [14, 29], which we use to realize a w-way multibus.

We explore two typical ways to multiplex this multibus organization: by address and by node. In multiplex by address, where wavelengths are assigned to different address spaces, any node can drive any of the w wavelengths, and thus requires arbitration. On the other hand, multiplexing by node gives each of the n nodes exclusive access to $\frac{w}{n}$ wavelengths (with w an integer multiple of n), which we will see has numerous advantages; however, the downside is that the number of nodes directly connected to the bus is then limited to w at best. (Other options are possible, for example leveraging WDM to decrease the number of physical waveguides by w. For the sake of simplicity, we leave this and other options for future work.)

An important consideration for both organizations is to prevent the light from circulating around the loop for more than one complete cycle, or older messages can cause undesirable interference. This can be easily handled in multiplex-by-node organizations by placing attenuator immediately before each modulator, to act as "sink" for the corresponding wavelength once the signal goes full circle. Alternatively, both multiplex-by-address and multiplex-by-node organizations may use an attenuator to "open" the loop, as long as modulators transmit in both directions simultaneously.

One power advantage of the multiplex-by-node organization is that it only requires $nb\frac{w}{n}$ transmitters ($\propto n$ if $w = n$), vs. nbw transmitters ($\propto n^2$ if $w = n$) in the multiplex-by-address organization. Since the optical power in a continuous laser source based system is dependent upon the number of modulators (Section 3.2.2), this difference may result in substantial optical power advantage for the multiplex-by-node organization.

Another power advantage of multiplex-by-node over multiplex-by-address is the possibility to optimize light power through individual coupling-ratio tuning at detectors at design time. This is because in multiplex-by-node organizations, the relative position of each detector with respect to the (sole) transmitter is known for every wavelength, and thus coupling at each detector can be designed to absorb just the right fraction of light power as to allow for efficient delivery to all detectors involved. In multiplex-by-address organizations, coupling at all detectors must be identical, since the signal may come from any one of the transmitters on the same wavelength, and thus the relative order in which they tap onto the signal is not known at design time. It can be shown mathematically that this results in wasted light power.

A third source of power waste in the multiplex-by-address organization comes from the fact that modulators do leak some light into the waveguide even in the "off" position. The more modulators coupled to a particular wavelength, the more aggregate light power leaks into the waveguide. In order for detectors to identify "on" and "off" states correctly, a proportional current bias must be applied to the receivers, which may result in a significant power waste.

For all the above reasons, in this paper we opt for the more practical multiplex-by-node organization.

3.2 Bus Design

We propose an opto-electrical hierarchical bus, where the optical loop constitutes the top level of the hierarchy, and nodes deliver information to processors via electrical sublevels. Figure 2 depicts a possible four-node organization for our 64-processor CMP, where each node is shared among four electrically interconnected L2 caches.

Our bus comprises an address/command bus, a data bus, and a snoop response bus. We allocate 64 bits to address/command (including ECC and tag bits), 72 bits to data (including 8-bit ECC and assuming that tags are provided at the header), and 8 bits per snoop response. Therefore, the number of waveguides is 136 for address/command plus data buses, and $8n$ to support snoop responses (each node provides w snoop responses using $\frac{w}{n}$ different wavelengths, for a total of $\frac{8w}{\frac{w}{n}} = 8n$ waveguides).

3.2.1 Protocol

Before delving into the details of a design space exploration, we give a high-level description of the bus protocol. The specifics of the cache coherence protocol are not relevant here; we focus on the handling of coherence requests by the split-transaction, fully pipelined hierarchical bus.

L2 cache accesses by processors may result in coherence requests, which travel down the electrical sublevel to the corresponding node where they are enqueued. Node switches arbitrate among the incoming coherence requests, and broadcast the winner(s) on the optical address bus.

Every node snoops in the requests put on the optical address bus by every other node. (Recall that each node transmits through different wavelengths.) Then, nodes arbitrate among concurrent requests, using the same finite state machine so that they all reach the same outcome independently. (This requires factoring in requests even at their originating switch.) Next, the selected requests are delivered to all caches simultaneously, and the rest are retried later. Caches compose individual snoop responses, which are relayed back down to the optical snoop response bus, which again all nodes read and process concurrently. Finally, the appropriate decision is made and the final snoop result is propagated up to the caches where the appropriate action is taken. Eventually, if indicated, data is generally sent down to the optical data bus (after winning arbitration over possibly competing responses from other caches in the same node), which the original requesting node collects and sends up to the request-

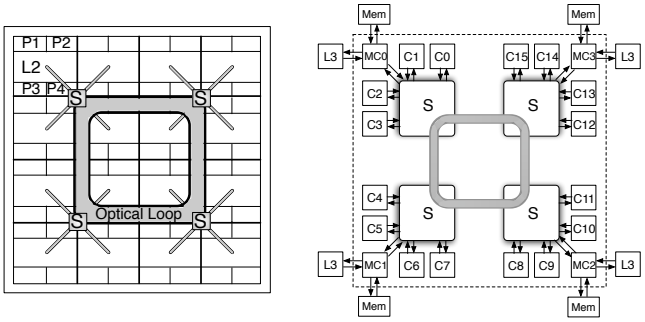


Figure 2: Simplified CMP floorplan diagram (left) and high-level system organization (right), showing the optical loop and the rest of the hierarchical bus. In the figures, S, MC(0-3), and C(0-15) stand for switch (separate switches for address/snoop and data buses), memory controller, and (L2) cache, respectively.

Delays of Optical Components [14]			
Technology	**45nm**	**32nm**	**22nm**
Modulator driver (ps)	25.8	16.3	9.5
Modulator (ps)	30.4	20	14.3
Detector (ps)	0.6	0.5	0.4
Amplifier (ps)	10.4	6.9	4.0
Si waveguide delay (ps/mm)	10.45	10.45	10.45

Table 2: Delays of various optical components at different technology nodes [14].

ing L2 cache.

3.2.2 Topology

Different literature sources offer varying projections on the number of available wavelengths per on-chip waveguide. Chen et al. [14] project that the number of wavelengths per waveguide will increase by four with each technology generation, reaching thirteen wavelengths at 32nm, while Kobrinsky et al. [29] assume an increase of one wavelength every other generation, resulting in three-four wavelengths at 32nm. Accordingly, we explore a range of four to twelve available wavelengths per waveguide.

We investigate several possible bus topologies, deriving for each of them area and power. Table 3 lists such topologies. In the table, H-nxkAkD (H for Hierarchical) designates a topology with n nodes on the optical bus and k address (data) wavelengths per node, totaling to nk wavelengths per waveguide in the address (data) bus. Beyond the optical loop, appropriately-sized electrical switches connect each of the sixteen quad-processor nodes to the network (hence the name Hierarchical). We sweep through all possible configurations given the WDM projections stated earlier on in this section: $k \in \{1, 2, 3\}$ for $n = 4$, and $k = 1$ for $n = 8$. For the sake of completeness, we also investigate a F-16x1A1D (F for Flat) topology, which requires no electrical routers (hence the name Flat), but that is unrealizable under WDM projections.

In the case of four nodes and $k > 1$, we also investigate topologies with a more limited support for new address transactions per cycle, H-nx1AkD, as we empirically observe in the course of our evaluation (Section 4) that this is enough to satisfy the applications' bandwidth demand on the address bus in the simulated system under consideration. This should generally result in area and power savings. Similarly, for the sake of area and power savings in the case of eight and sixteen nodes, we explore reducing the electrical snoop bandwidth to four (matching the bandwidth of the H-4x1AkD topologies). This is indicated by appending (4S) to the topology encoding.

Frequency Estimation

We estimate the operating frequency of the bus by calculating the time needed for the light to travel from any node to the farthest node on the (unidirectional) optical loop, so that information can be transmitted to all nodes in one bus cycle. With the loop bus centered on the die (Figure 2), and through simple geometric calculations, we estimate its total length to be 36mm, 45mm, and 45mm for 4-, 8-, and 16-node topologies, respectively. If we assume for simplicity that all neighboring nodes are equidistant, then the distance between any two nodes that are farthest apart is 27mm, 39.4mm, 42.2mm, respectively. Using the waveguide and optical-component delays provided in Table 2, and accounting for 4 FO4 latching delay (estimated using ITRS data), we obtain the maximum operating frequencies: 2.9GHz, 2.1GHz, and 2GHz, respectively. This implies that all three buses can run safely at 2GHz–exactly half the cores' frequency. (For simplicity, we assume that the electrical routers in the Hierarchical topologies can operate at this frequency regardless of their size.)

Area Estimation

We estimate the required areas on the active, optical, and metal layers for each organization (Table 3). All address, snoop, and data buses are considered in the area calculations.

In the active area, we account for electrical switches in each node, as well as transmitters and receivers on the optical bus. For simplicity, however, we do not include the area occupied by the repeaters in the electrical wiring, although we do include their contribution to power consumption later in this section. We use Orion [54] to estimate the area of input and output buffers, as well as the crossbar areas inside the switches. We assume four-entry input buffers to receive requests/addresses from each L2 cache, and single-entry input buffers for snoop request/response networks. In the data network, we allocate sixteen-entry buffers to collect data from each L2 cache, but compensate input buffer size at the optical end with optical width as follows: sixteen-, eight-, or four-entry input buffers in four (4x1D), eight (4x2D and 8x1D), or wider (4x3D and 16x1D) optical bus topologies, respectively. Output buffers are single-entry in all cases. We carefully specify the number of input and output ports considering the components connected to each switch (Figure 2), which in turn determines the number of input and output buffers, as well as the size of the crossbar in each case.

We estimate the active area taken up by transmitters and receivers required for the optical buses by conservatively assuming that modulator driver and TIA each occupy $50\mu m^2$, although standard scaling rules predict smaller areas for these components [40]. We assume $80\mu m^2$ modulators (10μm-diameter ring), $10\mu m \times 10\mu m$ detectors [14], and $80\mu m^2$ wave-selective filter areas (10μm-diameter ring resonator). Modulators and detectors consume area in both the active and optical layers; modulator drivers and TIAs are on the active layer, and filters are on the optical layer.

For the multiplex-by-node optical buses, the number of transmitters in each node is $\mathrm{tx_{node}} = b_a a + b_d d + b_s s$, where b_a, b_d, and b_s are the number of address, data, and snoop-response bits, respectively, and a, d, and s are address, data, and snoop bandwidth per node, respectively. Since each node has to be able to receive all the transmitted information by other nodes, the total number of receivers is $(n-1)\mathrm{tx}$, where n is the number of nodes on the optical bus, and $\mathrm{tx} = n \cdot \mathrm{tx_{node}}$ is the total number of transmitters on the bus. Therefore, while the number of transmitters is O(n), the number of receivers is $O(n^2)$.

The area occupied in the optical layer is calculated as the sum of waveguide, modulator, detector, and wave-selective filter areas. We assume the component areas specified above, and Si waveguide pitch as provided in Table 1.

The resulting active area is relatively modest, and the required optical layer easily fits within $400mm^2$ (Table 3).

Finally, we estimate the metal wiring area required for

Optical Topology	Snoop Requests /Bus clk	Area (mm^2)				Power (W)					
		Active Si Layer		Metal	Optical	Electrical Level		Optical Level		Total On-chip	
		Switch	Tx/Rx	Layer	Layer	Switch	Wiring	Tx/Rx	Optical	(α=1)	(α=0.5)
H-4x1A1D	4	1.71	0.39	15.21	33.68	1.75	12.82	0.60	0.79	15.56	9.04
H-4x2A2D	8	2.72	0.78	24.42	34.10	3.03	20.59	1.19	1.58	25.60	15.13
H-4x3A3D	12	4.00	1.17	33.64	34.51	4.64	28.36	1.79	2.37	35.98	21.49
H-4x1A2D	4	1.93	0.56	15.21	33.86	2.06	12.82	0.85	1.13	16.30	9.73
H-4x1A3D	4	2.13	0.72	15.21	34.04	2.37	12.82	1.11	1.47	17.03	10.41
H-8x1A1D	8	4.05	1.89	12.21	51.64	4.50	10.30	3.07	6.35	21.05	15.44
H-8x1A1D(4S)	4	3.08	1.59	7.6	51.3	3.25	6.41	2.58	5.33	14.91	11.34
F-16x1A1D	16	14.38	10.05	N/A	77.08	16.70	N/A	16.78	39.06	53.01	50.90
F-16x1A1D(4S)	4	6.77	6.4	N/A	72.81	7.42	N/A	10.68	24.86	30.53	29.53

Table 3: Area and power characterization of different optical bus topologies. Tx/Rx stands for transmitter/receiver; α is switching activity factor. Total on-chip power is the sum of switch, wiring, Tx/Rx, and half the optical power components (due to a 3dB coupling loss (Section 3.2.2), only half of the optical power is actually consumed on chip). All dynamic power components in switching, wiring, and Tx/Rx columns assume α=1. For α=0.5, only the total sum is provided.

the electrical sub-interconnects in hierarchical organizations. We assume a global wire pitch of 400nm and wire length of 4.5mm and 2.25mm (estimated according to the floorplan in Figure 2) for four- and eight-node configurations, respectively. From each cache to its node, the links include single address and data paths, and as many snoop-response paths as needed in each topology (number of snoop requests per cycle in Table 3). From each node to a cache, the links include single data path and as many snoop-request and snoop-result paths as indicated in the table.

Power Estimation

We categorize the power consumption of the interconnect system into two: the power consumed in the electrical sublevels (switches and wiring), and the power consumed in the optical bus. Table 3 shows a detailed breakdown of power consumption in all topologies under consideration. We report power calculations for each component assuming full switching activity ($\alpha = 1$), but report total power consumption at full, as well as 50% activity ($\alpha = 0.5$).

We estimate the static and dynamic power consumed by the switches in the nodes again using Orion [54] following the structural assumptions outlined in Section 3.2.2.

The static and dynamic power consumption of the wires is estimated following the methodology in [22, 23] for power-delay optimized repeater insertion and wire sizing.[1] We estimate according to ITRS [25] projections that a minimum-sized repeater has approximately 1μW of leakage power consumption.

There are two main power components due to the optical loop: electrical and optical power. Electrical power is the *on-chip* power consumed by the modulator drivers in transmitters (117μW per driver), and TIAs in receivers (257μW per TIA). For calculating the modulator driver and TIA power we used ITRS device projections [25] and standard circuit procedures.

Optical power is the *off-chip* power required by the modulator to modulate and transmit the information optically from one node to the others. In our analysis, we first calculate the minimum optical threshold power required for a detector to detect a signal correctly, which is based on the voltage swing requirement and signal-to-noise ratio of the receiver as suggested by Connor et al. [40]. In our case, the minimum detector current requirement comes to 30μA. Because only one node transmits with a specific wavelength, and the relative distance between a transmitter and a receiver is known at design time, it is possible to design the detectors to tap only the minimum amount of power adequate for signal detection, resulting in minimum overall optical power. Beginning with the minimum power required at the farthest receiver in the optical loop, we calculate the input power required at the transmitter's modulator by visiting nodes in reverse order up to the transmitter, and accumulating at each step the power losses incurred (Table 4). Each modulator requires this amount of optical power, since we assume a continuous wave laser source which will be always on, irrespective of whether data is being transmitted.

[1]We estimate 26ps/mm repeatered wire delay.

	Losses
On-chip coupling loss (dB) [40]	3
Si waveguide loss (dB/cm) [40]	1.3
Splitter loss (dB) [40]	0.2
Modulator insertion loss (dB) [1]	1
Interlayer coupling loss (dB)	1
Bending loss (dB) [40]	0.5
Quantum efficiency [40]	0.8

Table 4: Major power losses incurred by an on-chip optical transmission system.

We formulate the minimum power per modulator in Equation 1. In the equation, P_{th} is the minimum power that is required for a detector to detect the optical signal, P_{loss} is the waveguide loss per unit length, L is the length of the bus, and N is the number of nodes on the bus. The first term in the equation accounts for the power required for all detectors, the last term accounts for the waveguide loss, and K accounts for the other losses in the path, such as bending losses, etc.

$$P_{\text{modulator}} = (N-1)P_{\text{th}}K \cdot 10^{\frac{P_{\text{loss}}L(N-1)}{10N}} \quad (1)$$

Using these analytical models, and accounting for the remaining losses in the optical system such as on-chip coupling, splitters, etc., we report the minimum required total optical power for each configuration (Table 3). Note, however, that only half of this optical power contributes to the total on-chip power consumption (Table 3), as the other half is lost during the coupling of light into the chip (3dB coupling loss).

Discussion

We observe that the most preferable topologies in terms of area and power are H-4x1A{1,2,3}D and H-8x1A1D(4S), although we empirically observe that H-4x1A1D has too low data bandwidth (Section 4.3). All other configurations have excessive power and area expenses in comparison, due to a variety of factors: higher snoop bandwidth, greater number of receivers and transceivers, larger switch crossbars and arbitration logic, etc. Another observation is that, in the four-node configuration, the power consumption of the optical components is relatively low compared to the electrical subnetwork.

Among the preferred organizations, we opt for H-4x1A{2,3}D for our evaluation, mainly because (1) they require lower laser power, and (2) they are more flexible, since they can dynamically allocate the wavelengths for requests from every four L2 caches, while in the eight-node configuration the wavelengths are highly partitioned among nodes,

leaving little room for flexibility.

4 EVALUATION

We now provide a first look at the performance impact of incorporating on-chip optical technology for bus-based CMPs. We first present the experimental setup, including the electrical baseline that we model; then, we describe the simulated applications, followed by our results.

4.1 Experimental Setup

We conduct our evaluation using a cycle-accurate execution-driven simulator based on SESC [41]. Latencies and occupancies of all structures are modeled in detail. The simulator models a 64-core chip multiprocessor featuring dynamic superscalar cores and a snoopy-coherent memory subsystem. Each core is 4-way out-of-order and runs at 4GHz. We summarize the core parameters in Table 5. Each core has access to a private, write-through L1 data cache. An eight-way banked, write-back L2 cache is shared every four cores through a crossbar. All sixteen L2 caches are connected through a snoopy, fully pipelined bus (the object of our study). The coherence protocol is MESI-based and permits cache-to-cache clean block transfers. A banked, shared L3 resides off chip, but with tags on chip. L3 is accessed in parallel with main memory, and it is exclusive of L2 caches. We model four on-chip L3/memory controllers, each connecting to one fourth of L3 and memory via 64GB/s and 32GB/s links, respectively.

Recall that, to fit 64 cores on a 400mm^2 die, we estimated to be able to accommodate no more than 16×2MB L2 caches (Section 3). This results in a L2 effective capacity of 512KB per core, which is somewhat small, and likely to result in increased miss rates (and thus traffic). We believe this core vs. cache area trade-off is likely to become an important factor in future many-core chips. To gain some insight into the effect of this area trade-off in performance, we also conduct simulations assuming area-unconstrained 16×8MB L2 caches (2MB per core).

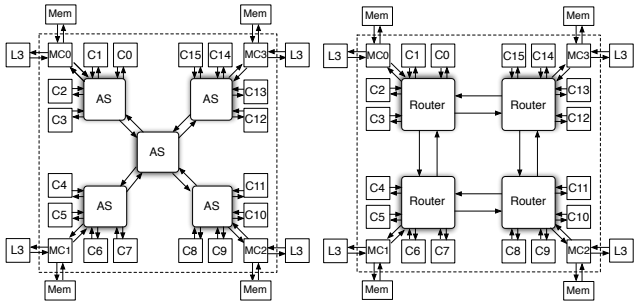
Nevertheless, following common practice for SPLASH-2 applications (Section 4.2), we use reduced cache sizes to compensate for the applications' reduced working sets [57] as follows: 64×8KB L1, 16×128/256KB L2, 1×16MB L3. Still, we use CACTI 4.1 [51] to obtain and use the latencies of the on-chip caches' full-size equivalents: 64×16KB L1, 16×2/8MB L2. As a sanity check, the last two columns of Table 7 list the global L2 miss rates obtained with the two configurations, as obtained during a bandwidth characterization experiment, which we describe later. Table 5 summarizes the memory subsystem parameters.

4.1.1 Electrical Bus

To conduct a meaningful evaluation of the impact of incorporating optical technology to bus-based interconnects, we establish a competitive, state-of-the-art electrical baseline with similar power and active/metal area characteristics as the competing opto-electrical buses. We discuss the address network first, followed by the data network.

An address bus can be implemented in a variety of ways, including a hierarchical tree organization (e.g., a single snooping coherence domain in the Sun Fireplane System Interconnect [12] implemented as two-level tree structure), and unidirectional [5, 26, 50] and bidirectional [7] ring-based interconnects.

We empirically found the tree topology to yield low latency and competitive bandwidth relative to other alternatives for our configuration, and therefore choose it as our baseline. We model it after existing proposals [12, 20]. In the modeled tree organization (Figure 3a) four L2 caches and a memory controller (which in turn manages one-fourth of the off-chip L3 and memory) connect to an address switch (AS), and four such address switches connect to a top-level address switch,



a) Address/Snoop Network b) Data Network

Figure 3: Modeled electrical baseline address and data networks. AS, MC(0-3), and C(0-15) stand for address switch, memory controller, and (L2) cache, respectively. Figures are not to scale.

all through point-to-point links. Requests issued by L2 caches are arbitrated in the switches at each level of the tree, until they reach the top level and are selected. From that point on, broadcasting a snoop request down to all caches, combining snoop responses up at the top-level switch, and again broadcasting the final snoop result down to the caches, takes a fixed amount of cycles. We implement a multibus by selecting multiple snoop requests at the top-level address switch and employing as many snoop request/response buses as needed.

We assume an H-tree layout with 4.5mm first-level (from the L2 caches) and 9mm second-level wire links. By using power-delay optimized repeatered-wires, we can accommodate a 2GHz bus clock frequency–half the cores' speed. Under no contention, the address phase of a request spends a total of 13 bus cycles on the bus: 4 cycles for request arbitration, 3 bus cycles for snoop request, and 6 bus cycles for snoop-response combining and result broadcasting (excluding time spent in the caches).

The data network (Figure 3b) consists of a four-node bidirectional ring. As in the case of the address switches, each data router serves requests from/to four local caches and a memory controller connected to it through point-to-point links. Routing is deterministic and balanced. Transfers within a node use a 16GB/s bypass path within the local router. Bandwidth at each ring link is 16GB/s in each direction, as is the read and write bandwidth of each L2 cache. Bandwidth from (to) the memory controller is 48GB/s (32GB/s). In the absence of contention, it takes 14 bus cycles to transfer a cache line on the data network to a cache in the farthest node.

Finally, we do not simulate I/O, and therefore we do not include it in the system we model.

To obtain area and power characteristics of the electrical bus (Table 6), we follow the estimation methodology described in Section 3.2.2 for the relevant electrical components. When compared to H-4x1A{1,2,3}D buses, an electrical bus with support for an equal number of snoop requests per bus cycle (four) exhibits comparable power consumption and active device area, but a 50% increase in metal area overhead. On the other hand, an electrical baseline with support for half as many snoop requests per bus cycle adds up to similar area and power characteristics as the opto-electrical counterparts. Thus, for our comparison, we choose the latter configuration as our baseline.

4.1.2 Opto-electrical Bus

We model the opto-electrical buses H-4x1A{1,2,3}D as described in Section 3.2. The uncontended latencies in these optical buses are 10 bus cycles for arbitration plus snoop request/response phases, and 12 bus cycles for a cache line data to be transferred on the bus across bus nodes.

Processor Core	
Frequency	4GHz
Fetch/issue/commit width	4/4/6
Inst. window [(Int+Mem)/FP]	56/48
ROB entries	128
Int/FP registers	96/96
Int ALUs/Branch units	4/2
Int Mul/Div units	1/1
FP ALUs	3
FP Mul/Div units	2/2
Ld/St units	2/2
Ld/St queue entries	24/24
Branch penalty (cycles)	7 (min.)
Store forward delay (cycles)	2
Branch predictor,	16K-entry,
(Hybrid of GAg + Bimodal)	14b GHR
BTB size / RAS entries	2048 / 24

Memory Subsystem	
Cache sizes for SPLASH-2 [57]	(1) 64×8KB L1, 16×128KB L2, 1×16MB L3 (2) 64×8KB L1, 16×256KB L2, 1×16MB L3
Cache associativity	4-way L1, 8-way L2, 16-way L3
Cache access latencies	2 IL1/DL1, 8 L2, 56 L3 cycles
Writeback/Replacement policy	WT DL1, WB L2 and L3
Block size	64 bytes
MSHR entries	8 IL1/DL1, 32 L2, 12 L3 (per bank)
IL1/DL1 Cache ports	1/3
L2/L3 Cache banks	8/8
L2 Cache coherence protocol	MESI-based
System bus	64 bits, 2GHz
Memory controllers	4
L3 off-chip bandwidth	4×64GB/s
Memory bus bandwidth	4×32GB/s
Memory RT from controllers	320 cycles

Table 5: Summary of the modeled CMP system. In the table, GHR, MSHR, RAS, and RT stand for global history register, miss status holding register, return address stack, and minimum round-trip time, respectively. Cycle counts are in processor cycles.

Electrical Topology						
Snoop Requests /Bus clk	**Area (mm^2)**		**Power (W)**			
	Switches/ Routers	Wiring	Switches/ Routers	Wiring	Total On-chip (α=1)	Total On-chip (α=0.5)
2	1.47	15.9	1.42	13.40	14.82	8.08
4	1.66	22.81	1.68	19.23	20.91	11.29

Table 6: Area and power characterization of two possible topologies for the baseline electrical bus, with two and four snoop requests per bus cycle, respectively. Total on-chip power is the sum of all electrical power components. Dynamic power components in switching and wiring columns assume α=1. For α=0.5, only the total sum is provided.

4.2 Applications

We use eleven applications from the SPLASH-2 suite [57] (our simulator currently does not support *volrend*). Their description, as well as their input parameters, are given in Table 7. We use MIPS binaries compiled with -O3 optimization level. We fast-forward the initialization part of the applications (at which point we start modeling timing and collecting statistics) and run them to completion.

Bandwidth Characterization

Figure 4 plots histograms, for both cache configurations (Table 5), of the average number of bus requests per processor cycle, sampled at 1,000-cycle intervals, and assuming infinite bus (but not memory) bandwidth, and one- and eight-bus-cycle address and data buses, respectively.

The results show that, for the studied applications, the address/snoop bus bandwidth requirements generally stay at or below 1.5 req./processor cycle. Naturally, bandwidth demand is generally shifted to lower values in the larger L2 cache configuration. Relatively speaking, considering the results with 256KB L2s, the most bandwidth-hungry applications are *barnes*, *cholesky*, *FFT*, *LU*, *ocean*, *radiosity*, and *raytrace*, which have significant periods of execution with bandwidth demand greater than 0.5 requests per processor cycle.

4.3 Results

Figure 5 shows performance results for H-4x1A1D, H-4x1A2D, and H-4x1A3D, relative to the electrical baseline. Interestingly, in spite of the higher snoop request bandwidth, H-4x1A1D experiences a significant performance degradation in nearly all cases. This is mainly due to its lower per-node data bandwidth (one outgoing port to the optical bus vs. two outgoing ring-ports in electrical baseline). When higher data bandwidth is provided via additional wavelengths (H-4x1A2D and H-4x1A3D), the opto-electrical configurations achieve significant speedups. This is particularly true in the configurations with area-constrained L2 caches, where the opto-electrical buses can accommodate the increased L2 miss rates better, resulting in average (peak) speedups of 1.30 (1.52) for the H-4x1A3D configuration.

To further understand the sources of performance improvement, Figure 6 shows the average latency breakdown (in bus cycles) of bus transactions in the baseline electrical, H-4x1A2D, and H-4x1A3D configurations. (In the plots, the Data Transfer category accounts for the latency spent only on the data network itself, excluding time spent in memory or caches, for example.)

We observe latency advantages for the opto-electrical configurations in both address/snoop and data networks. In the former, effective latency is reduced by 26% on average (23 to 17 bus cycles) when moving from electrical to electro-optical technology (both in the case of area-constrained and area-unconstrained L2 configurations). Recall that, even in the absence of contention, the opto-electrical buses have a latency advantage over our electrical baseline. Moreover, the opto-electrical buses can support twice as much snoop request/response bandwidth as the electrical baseline at similar power and area cost. For some applications with relatively high bandwidth demand (Figure 4), such as *radiosity* and *raytrace*, the savings can be as high as 39%. Our simulation data show that contention at the arbitration phase for these applications is due in part to conflicting requests to the same cache line (in our bus protocol, conflicting requests to a cache line with an outstanding request are deferred). This is amplified indirectly by the extended latency of the outstanding requests in the data network.

Indeed, for the configurations under study, the main overall benefit comes from reduced contention (and thus effective latency) for data transfers. Our simulations show that the data network struggles to supply the bandwidth needed to satisfy these requests. It is in the data network that the availability of extra wavelengths through WDM yields the largest performance improvements. Still, some applications suffer from significant contention in the data network even for H-4x1A3D, leaving room for further improvement. From our simulation data, we identify the main cause to be contention at the L2 cache input ports. Notice that the bandwidth to the caches (and memory controller) is kept unchanged in all

SPLASH-2	Description	Problem size	Global L2 miss %	
			128KB L2s	256KB L2s
Barnes	Evolution of galaxies	16k particles	0.17	0.11
Cholesky	Cholesky factorization kernel	tk29.O	0.40	0.20
FFT	FFT kernel	64k points	0.53	0.23
FMM	N-body problem	16K particles	0.06	0.01
LU	LU kernel	512×512 matrix, 16×16 blocks	0.02	0.02
Ocean	Ocean movements	258×258 ocean	2.97	2.45
Radiosity	Iterative diffuse radiosity method	-room -ae 5000.0 -en 0.05 -bf 0.1	0.84	0.17
Radix	Integer radix sort kernel	radix 32, 1M integers	0.20	0.14
Raytrace	3-D ray tracing	car	1.31	0.69
Water-NSq	Forces and potentials of	512 molecules	0.10	0.05
Water-Sp	water molecules (both)	512 molecules	0.12	0.02

Table 7: Application descriptions and simulated problem sizes. Observed global L2 miss rates (averaged over all L1 and L2 caches) for two set of cache configurations using optimistic (single-bus-cycle address and eight-bus-cycle data transmissions, and no contention) bus (but not memory) are provided for reference.

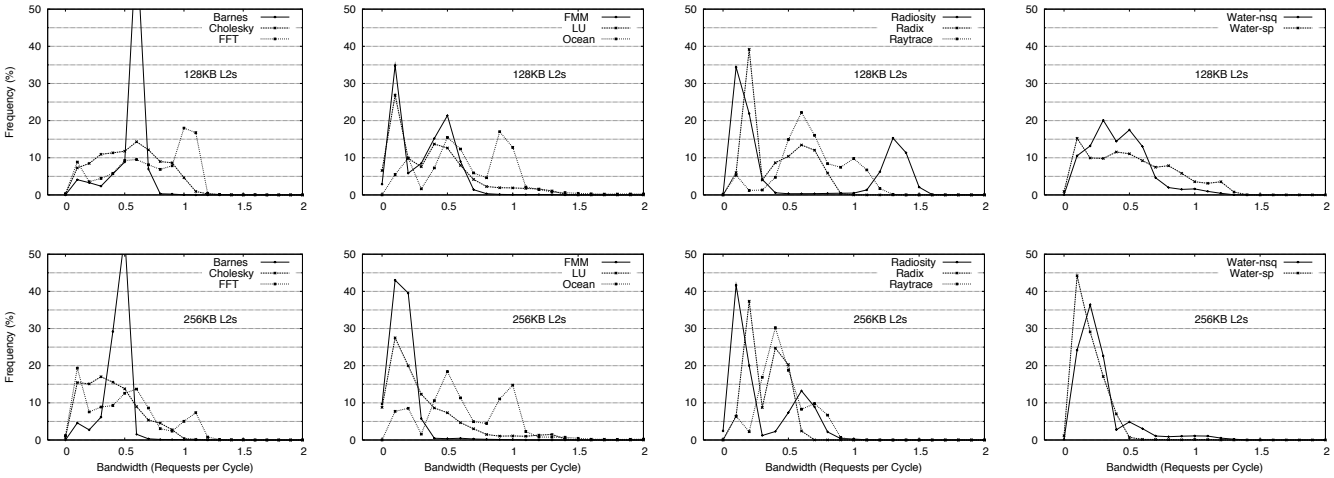


Figure 4: Histograms of the average number of bus requests per processor cycle sampled at 1,000-processor cycle execution intervals. An optimistic bus (single-bus-cycle address and eight-bus-cycle data transmissions, no contention) is used. Top-row plots are obtained using 128KB L2 caches, and bottom-row plots are obtained using 256KB L2 caches.

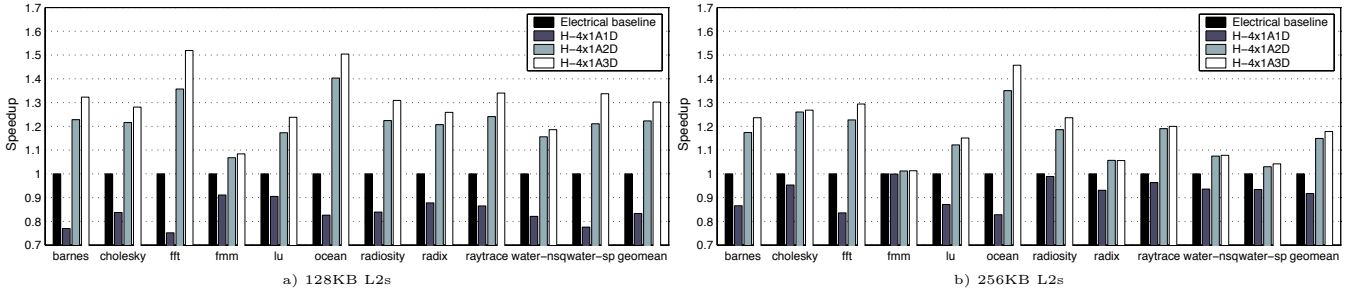
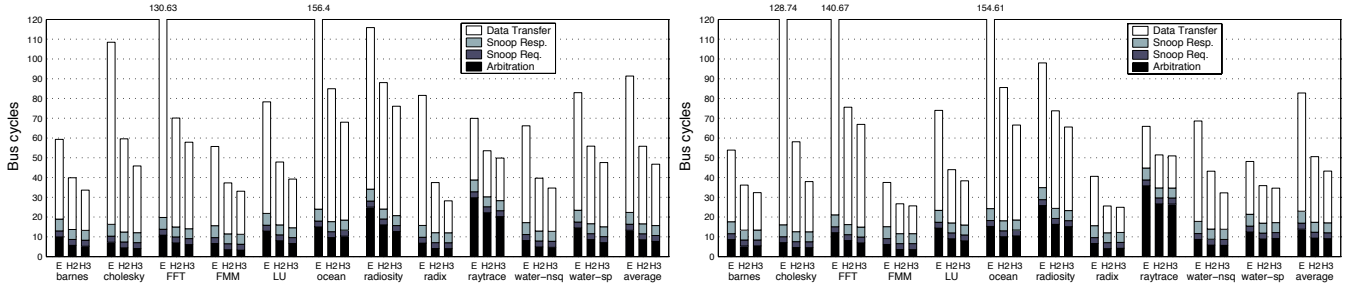


Figure 5: Performance improvements of four-node opto-electrical buses as the number of available wavelengths per node for the data network is varied from one to three. (The address network uses one distinct wavelength per node in all three cases.) Speedups are relative to the baseline electrical interconnect. Results are provided for systems with two different cache configurations (Table 5).

configurations in spite of the increased data bandwidth on the optical loop.

Note that some of the performance gains exhibited by the opto-electrical buses could be given back in exchange for power/area savings. Indeed, for the two opto-electrical configurations, our power/area model indicate that reducing the snoop request/response bandwidth to that of the electrical baseline could lower both power and (more so) metal area cost significantly.

Finally, Table 8 shows parallel efficiencies (relative to a sequential run in the same configuration in each case) for all applications running on the electrical baseline and on H-4x1A{2,3}D. In general, scalability improves with the addition of optical technology. Not surprisingly, those applications that suffer from more contention in the data network tend to exhibit lower parallel efficiencies in all configurations. And it is precisely the scalability of these applications that improves the most with the addition of optical technology.



a) 128KB L2s b) 256KB L2s

Figure 6: Average latency breakdown (in bus cycles) of bus transactions in baseline electrical (E), H-4x1A2D (H2), and H-4x1A3D (H3) buses. In the data network, only the time actually spent in the network is reported.

SPLASH-2	128KB L2s			256KB L2s		
	Baseline	H-4x1A2D	H-4x1A3D	Baseline	H-4x1A2D	H-4x1A3D
Barnes	0.69	0.82	0.89	0.69	0.79	0.84
Cholesky	0.21	0.25	0.26	0.26	0.31	0.31
FFT	0.31	0.41	0.46	0.39	0.47	0.50
FMM	0.52	0.55	0.56	0.59	0.59	0.59
LU	0.37	0.42	0.45	0.44	0.48	0.49
Ocean	0.23	0.31	0.33	0.24	0.32	0.34
Radiosity	0.16	0.19	0.21	0.17	0.21	0.21
Radix	0.72	0.85	0.89	0.87	0.91	0.91
Raytrace	0.24	0.28	0.30	0.22	0.25	0.26
Water-NSq	0.64	0.73	0.75	0.78	0.83	0.83
Water-Sp	0.48	0.57	0.63	0.76	0.78	0.79

Table 8: Parallel efficiencies of the simulated SPLASH-2 applications for the specified configurations.

In summary, our evaluation shows that incorporating optical technology in bus-based CMPs can have a beneficial impact on performance, and that WDM support may be critical to effect this impact in both address/snoop and data networks. The fact that WDM comes at very small additional area and power is encouraging. In the particular design points that we evaluated, the contribution to performance by the data network turned out to be dominant. A more sophisticated design of the data network, perhaps at an additional area and power expense, may allow applications to benefit more from improvements in the address/snoop bus.

5 AREAS FOR FUTURE RESEARCH

Our effort represents only an initial foray into optical interconnects in chip multiprocessors. While the results are promising, the area is a rich one for future research. In this section, we summarize some of the areas that we identify as critical for further investigation. Although there are many issues to be addressed on the materials, device, and manufacturing fronts, we focus on problems that are more interdisciplinary in nature and of interest to systems researchers.

An area of interest involves the best use of optical waveguides and WDM for a particular CMP design. On the one hand, our results indicate that increasing the number of WDM channels alone can have a large performance impact for bus-intensive applications in the CMP configuration studied. On the other hand, other CMP organizations (e.g., a larger number of cores) may require other organizations that present very different power/performance/complexity challenges (e.g., number of modulators, power and bandwidth requirements at the nodes, etc.).

Another interesting area for exploration involves the use of polymer as the waveguide material. As we noted, a major drawback of polymer waveguides at the moment is the lack of a suitable modulator, and VCSELs introduce additional manufacturing cost and complexity. However, the 2.3x improvement in the waveguide refractive index using polymers over silicon is a compelling enough speed advantage to motivate additional research. At the same time, the lower bandwidth density of polymer is likely to require more aggressive levels of WDM than silicon for the same level of interconnect bandwidth. Similar to conventional local, intermediate, and global electrical interconnects, in which different layers vary in their propagation and bandwidth density, the same can be potentially achieved in the optical domain using a hybrid system of fast, wider-channel polymer, and slower, but narrow-channel silicon waveguide layers. An open research question is how to architect CMP systems of hundreds of processors that best exploit such a heterogeneous optical interconnect structure.

The performance improvements that can be obtained using optics is limited by how far the technology penetrates into the bus protocol. In our hierarchical approach, for instance, a fraction of the bus latency is addressed via optics, but there is still a large fraction that remains entirely electrical. We believe that there is interesting joint research at the bus protocol and optics component fronts that can be undertaken to address this shortcoming.

Temperature management of the optical components is also a critical systems-level issue. Optical modules are very sensitive to temperature variations and require either active or passive optical control to maintain stable device operation [56]. With more of the responsibility for microprocessor temperature management moving to the system level, microarchitects may need to craft means for dynamically maintaining viable optical component operating temperatures. Furthermore, while our initial effort has focused on performance optimization within given power and area constraints, there are certainly power-aware optimizations that can be devised for on-chip optical interconnects (as previously investigated for off-chip optical systems [16]).

Finally, we believe that there are other interesting optical network topologies that can be potentially explored. Our loop bus has many advantages, including permitting the use of a simple snoopy-based protocol. Other topologies, including flat switch-type networks, are certainly possible, although buffering is difficult since it requires translation between op-

tical and electrical components (which may call for bufferless approaches).

6 RELATED WORK

Haurylau et al. [21] extract the delay, bandwidth density, and power requirements that the optical interconnect components must meet in order for on-chip optical interconnects to be comparable with their electrical counterparts. Similarly, Chen et al. [13] project the performance characteristics of future optical devices and then compare the optical and electrical interconnect paths in terms of delay, bandwidth density, and power. They estimate that, for a unit distance at 32nm technology, the delay of an optical interconnect would be approximately 2.2 times faster than an electrical wire. Further they show that, at the same technology node, optical interconnects consume less power, but have lower bandwidth density than their electrical counterparts due to their wider pitches (assuming polymer waveguides).

Kobrinski et al. [29] investigate optical clock distribution and optical global signaling and compare these with their electrical counterparts. They find little power, jitter, or skew improvements from using optics in clock distribution. However, they conclude that by using WDM, optics can be beneficial for global signaling in terms of high bandwidth and low latency. Chen et al. [15] compare four different technologies (electrical, 3D, optical, and RF) for on-chip clock distribution. They also show that because most of the skew and power of clock signaling arise in local clock distribution, there is no significant skew and power advantages of the new technologies, including the optical solution.

Connor [40] reviews the optical interconnect technologies and opto-electronic devices for inter- and intra-chip interconnects, followed by an EDA design flow methodology for optical link designs. The work describes an optical clock distribution network implementation and finds, through circuit simulation, that such a realization can consume significantly less power (five times lower power in case of 64-node H-three at 5GHz) than its electrical counterpart. The work also proposes a behavioral model of a 4x4 crossbar-like data network, based on wavelength routing that connects four masters to four slaves. However, they do not evaluate its performance in a system.

On-chip transmission-line-based interconnects have also been proposed as alternatives to traditional global wires. These interconnects make use of very wide metal wires so that signals propagate in the high frequency LC domain at near the speed of light [11]. While they do not require any new process to implement, one of their major drawbacks is that they have very low bandwidth due to the large wire width required, which may not be suitable to realize a wide inter-processor interconnect.

There have been many proposals for off-chip optical interconnects targeting shared or distributed memory multiprocessors. We comment on some recent efforts. Louri et al. [36, 37] propose snoopy address sub-interconnects where an optical token circulates around the processors to provide arbitration to transmit the requests through an H-tree like fully optical interconnect. This approach requires modification of the coherence protocol. Webb et al. [55] focus on optical network implementations in large scale distributed shared memory systems. They propose the use of an optical crossbar (implemented using free space optics) for intra-cluster connections, and either crossbar or a point-to-point hypercube optical interconnect that has less connectivity for the inter-cluster connections. Finally, Chen et al. [16], through detailed power models of optical components and network simulation, explore the design space of power-aware opto-electronic off-chip networks. They propose several techniques to dynamically control the power in such networks, achieving significant power savings. Their analysis is performed using both VCSEL-based modulation and off-chip laser source feeding multiple-quantum-well (MQW) modulators, finding the VCSEL-based solution slightly more power-performance efficient. Note that we assume (on-chip) ring-resonator-based PIN modulators that generally have favorable characteristics over MQW modulators.

Burger and Goodman [10], in an attempt to exploit the high-bandwidth broadcasting capability of optical interconnects (particularly when free-space optics is used), propose a new execution model to reduce serial overheads within a parallel program by having the serial code performed redundantly at any node of a massively parallel multiprocessor/multicomputer system allocated to the program.

Nelson et al. [39] evaluate the performance improvement of replacing global point-to-point electrical wires between the unified front-end and multiple back-ends of a large-scale clustered multithreaded (CMT) processor, where the back-ends are spread across the die, spatially interleaved with caches due to thermal constraints.

Our work is distinct from these prior efforts by being the first to investigate the design trade-offs, performance benefits, and power and area costs, of integrating CMOS-compatible optical interconnect technology into CMPs.

7 CONCLUDING REMARKS

We have investigated the integration of CMOS-compatible optical technology into the cache-coherent bus of a future CMP. By carefully modeling the speed, area, and power characteristics of electrical and recently-developed optical components, and projecting to 32nm technology, we determine that a hierarchical bus consisting of both optical and electrical levels yields significant performance within reasonable power and area constraints. Our approach exploits wave division multiplexing technology (WDM) to provide each node on the optical bus with unique wavelength(s), which are used to build a high-bandwidth multi-way bus. This speeds up several protocol operations, especially data transfer and arbitration.

In the course of our work, we identify several critical areas for future interdisciplinary research, among them exploring additional ways to exploit WDM, the use of both polymer and silicon waveguides, the design of bus protocols and optical components that permit further replacement of wires with optics, dynamic temperature management of the optical components, and the exploration of alternative network topologies. Overall, on-chip optical interconnects is a rich area for future research, one with great potential to address the global interconnect limitations of future CMPs.

ACKNOWLEDGMENTS

This work was supported in part by NSF CAREER awards CCF-0545995 (Martínez) and CCF-0347649 (Apsel); NSF awards CNS-0509404 and CCF-0429922 (Martínez); NSF award CCF-0304574 (Albonesi); an IBM Faculty Award (Martínez); two Intel Foundation graduate fellowships (Meyrem and Nevin Kırman); a NSF graduate fellowship (Watkins); and gifts from Intel, Analog Devices, and SAIC.

REFERENCES

[1] V.R. Almeida, C.A. Barrios, R.R. Panepucci, M. Lipson, M.A. Foster, D.G. Ouzounov, and A.L. Gaeta. All-optical switching on a silicon chip. *Optics Letters*, 29(24):2867, December 2004.

[2] H.B. Bakoglu. *Circuits, Interconnections, and Packaging for VLSI*. Addison-Wesley, Menlo Park, CA, 1990.

[3] K. Banerjee and A. Mehrotra. Power dissipation issues in interconnect performance optimization for sub-180nm designs. In *Symposium on VLSI Circuits Digest of Technical Papers*, pages 12–15, Honolulu, June 2002.

[4] C. A. Barrios, V. R. de Almeida, and M. Lipson. Low-power-consumption short-length and high-modulation-depth silicon electrooptic modulator. *Journal of Lightwave Technology*, 21(4):1089–1098, April 2003.

[5] L. A. Barroso and M. Dubois. The performance of cache-coherent ring-based multiprocessors. In *International Symposium on Computer Architecture*, pages 268–277, San Diego, CA, May 1993.

[6] A. F. Benner, M. Ignatowski, J. A. Kash, D.M. Kuchta, and M. B. Ritter. Exploitation of optical interconnects in future server architectures. *IBM Journal of Research and Development*, 49(4/5):755, July–September 2005.

[7] M. A. Blake, S. M. German, P. Mak, A. E. Seigler, and G. A. Huben. Bus protocol for a switchless distributed shared memory computer system. United States Patent #6,988,173 B2, International Business Machines Corporation, January 2006.

[8] S. Borkar. Low power design challenges for the decade. In *Conference on Asia South Pacific Design Automation*, pages 293–296, Yokohama, Japan, January–February 2001.

[9] S.Y. Borkar, P. Dubey, K.C. Kahn, D.J. Kuck, H. Mulder, S.S. Pawlowski, and J.R. Rattner. Platform 2015: Intel processor and platform evolution for the next decade. Technical report, Intel White Paper, March 2005.

[10] D. Burger and J. R. Goodman. Exploiting optical interconnects to eliminate serial bottlenecks. In *Proceedings of the Third International Conference on Massively Parallel Processing Using Optical Interconnections*, pages 106–113, October 1996.

[11] R. T. Chang, N. Talwalkar, P. Yue, and S. S. Wong. Near speed-of-light signaling over on-chip electrical interconnects. *IEEE Journal of Solid-State Circuits*, 38(5):834–838, May 2003.

[12] A. Charlesworth. The Sun Fireplane system interconnect. In *ACM/IEEE Conference on Supercomputing*, pages 1–14, Denver, CO, November 2001.

[13] G. Chen, H. Chen, M. Haurylau, N. Nelson, D. Albonesi, P. M. Fauchet, and E.G. Friedman. Electrical and optical on-chip interconnects in scaled microprocessors. In *International Symposium on Circuits and Systems*, pages 2514–2517, Kobe, Japan, May 2005.

[14] G. Chen, H. Chen, M. Haurylau, N. Nelson, P. M. Fauchet, E.G. Friedman, and D. Albonesi. Predictions of CMOS compatible on-chip optical interconnect. In *International Workshop on System-Level Interconnect Prediction*, pages 13–20, San Francisco, CA, April 2005.

[15] K.-N. Chen, M. J. Kobrinsky, B. C. Barnett, and R. Reif. Comparisons of conventional, 3-D, optical, and RF interconnects for on-chip clock distribution. *IEEE Transactions on Electron Devices*, 51(2):233–239, February 2004.

[16] X. Chen, L.-S. Peh, G.-Y. Wei, and Y.-K. Prucnal. Exploring the design space of power-aware opto-electronic networked systems. In *International Symposium on High-Performance Computer Architecture*, pages 120–131, San Francisco, CA, February 2005.

[17] J. Crow. Terabus Objectives and Challenges, C2COI Kickoff Meeting, http://www.darpa.mil/mto/c2oi/kick-off/Crow_Terabus.pdf, 2003.

[18] D. E. Culler and J. P. Singh. *Parallel Computer Architecture: A Hardware/Software Approach*. Morgan Kaufmann Publishers, San Francisco, CA, first edition, 1999.

[19] J. D. Davis, J. Laudon, and K. Olukotun. Maximizing CMP throughput with mediocre cores. In *International Conference on Parallel Architectures and Compilation Techniques*, Saint Louis, MO, September 2005.

[20] S. R. Deshpande. Method and apparatus for achieving correct order among bus memory transactions in a physically distributed SMP system. United States Patent #6,779,036, International Business Machines Corporation, August 2004.

[21] M. Haurylau, H. Chen, J. Zhang, G. Chen, N.A. Nelson, D.H. Albonesi, E.G. Friedman, and P.M. Fauchet. On-chip optical interconnect roadmap: Challenges and critical directions. In *2nd International Conference on Group IV Photonics*, pages 17–19, Antwerp, Belgium, September 2005.

[22] R. Ho. *On-Chip Wires: Scaling and Efficiency*. Ph.D. dissertation, Dept. of Electrical Engineering, Stanford University, August 2003.

[23] R. Ho, W. Mai, and M. A. Horowitz. The future of wires. *Proceedings of the IEEE*, 89(4):490–504, April 2001.

[24] Intel White Paper. *Next Leap in Microprocessor Architecture: Intel Core Duo*, 2006.

[25] The ITRS Technology Working Groups, http://public.itrs.net. *International Technology Roadmap for Semiconductors (ITRS) 2005 Edition*.

[26] S. Fields J. M. Tendler, S. Dodson. POWER4 system microarchitecture. Technical report, IBM White Paper, October 2001.

[27] P. Kapur, G. Chandra, and K.C. Saraswat. Power estimation in global interconnects and its reduction using a novel repeater optimization methodology. In *IEEE/ACM Design Automation Conference*, pages 461–466, New Orleans, LA, June 2002.

[28] P. Kapur and K. C. Saraswat. Comparisons between electrical and optical interconnects for on-chip signaling. In *International Interconnect Technology Conference*, pages 89–91, Burlingame, CA, June 2002.

[29] M. Kobrinsky, B. Block, J-F. Zheng, B. Barnett, E. Mohammed, M. Reshotko, F. Robertson, S. List, I. Young, and K. Cadien. On-chip optical interconnects. *Intel Technology Journal*, 08(02), May 2004.

[30] R. Kumar, V. Zyuban, and D. M. Tullsen. Interconnections in multi-core architectures: Understanding mechanisms, overheads and scaling. In *International Symposium on Computer Architecture*, pages 408–419, Madison, Wisconsin, June 2005.

[31] J. Laudon and D. Lenoski. The SGI Origin: A ccNUMA highly scalable server. In *International Symposium on Computer Architecture*, pages 241–251, Denver, CO, June 1997.

[32] D. Lenoski, J. Laudon, T. Joe, D. Nakahira, L. Stevens, A. Gupta, and J. Hennessy. The DASH prototype: Logic overhead and performance. *IEEE Transactions on Parallel and Distributed Systems*, 4(1):41–61, January 1993.

[33] A. F. J. Levi. Fiber-to-the-Processor and Other Challenges for Photonics in Future Systems, http://asia.stanford.edu/events/Spring05/slides/050421-Levi.pdf, 2005.

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

[35] A. Liu, R. Jones, L. Liao, D. Samara-Rubio, D. Rubin, O. Cohen, R. Nicolaescu, and M. Paniccia. A high-speed silicon optical modulator based on a metal-oxide-semiconductor capacitor. *Nature*, 427:615–618, February 2004.

[36] A. Louri and A. K. Kodi. Parallel optical interconnection network for address transactions in large-scale cache coherent symmetric multiprocessors. *IEEE Journal of Selected Topics on Quantum Electronics*, 9(2):667–676, March–April 2003.

[37] A. Louri and A. K. Kodi. An optical interconnection network and a modified snooping protocol for the design of large-scale symmetric multiprocessors (SMPs). *IEEE Transactions on Parallel and Distributed Systems*, 15(12):1093–1104, December 2004.

[38] D. A. Miller. Rationale and challenges for optical interconnects to electronic chips. *Proceedings of the IEEE*, 88(6):728–749, June 2000.

[39] N. Nelson, G. Briggs, M. Haurylau, G. Chen, H. Chen, D.H. Albonesi, E.G. Friedman, and P.M. Fauchet. Alleviating thermal constraints while maintaining performance via silicon-based on-chip optical interconnects. In *Workshop on Unique Chips and Systems*, Austin, Texas, March 2005.

[40] Ian O'Connor. Optical solutions for system-level interconnect. In *International Workshop on System-Level Interconnect Prediction*, pages 79–88, Paris, France, February 2004.

[41] University of Illinois at Urbana-Champaign. `http://sesc.sourceforge.net`, 2005.

[42] A. Pappu and A. Apsel. Electrical isolation and fan-out in intra-chip optical interconnects. In *International Symposium on Circuits and Systems*, pages II–533–6 Vol. 2, Vancouver, Canada, May 2004.

[43] A. M. Pappu and A. B. Apsel. Analysis of intrachip electrical and optical fanout. *Applied Optics*, 44(30):6361–6372, October 2005.

[44] A. M. Pappu and A. B. Apsel. A low power, low delay TIA for on-chip applications. *Conference on Lasers and Electro-Optics*, 1:594–596, May 2005.

[45] P. Rabiei, W. H. Steier, C. Zhang, and L. R. Dalton. Polymer micro-ring filters and modulators. *Journal of Lightwave Technology*, 20(11):1968–1975, November 2002.

[46] A. Rahman and R. Reif. System-level performance evaluation of three-dimensional integrated circuits. *IEEE Transactions on Very Large Scale Integrated Systems*, 8(6):671–678, December 2000.

[47] H. Shah, P. Shiu, B. Bell, M. Aldredge, N. Sopory, and J. Davis. Repeater insertion and wire sizing optimization for throughput-centric VLSI global interconnects. In *IEEE/ACM International Conference on Computer Aided Design*, pages 280–284, San Jose, CA, November 2002.

[48] R. A. Soref and B. R. Bennett. Electrooptical effects in silicon. *IEEE Journal on Quantum Electronics*, 23(1):123–129, January 1987.

[49] S.J. Souri, K. Banerjee, A. Mehrotra, and K.C. Saraswat. Multiple Si layer ICs: Motivation, performance analysis, and design implications. In *IEEE/ACM Design Automation Conference*, pages 213–220, Los Angeles, CA, June 2000.

[50] K. Strauss, X. Shen, and J .Torrellas. Flexible snooping: Adaptive forwarding and filtering of snoops in embedded-ring multiprocessors. In *International Symposium on Computer Architecture*, Boston, MA, June 2006.

[51] D. Tarjan, S. Thoziyoor, and N. P. Jouppi. Cacti 4.0. Technical Report HPL-2006-86, HP Laboratories Palo Alto, June 2006.
http://quid.hpl.hp.com:9081/cacti/.

[52] J. Tatum. VCSELs for 10 GB/s optical interconnects. In *IEEE Emerging Technologies Symposium on BroadBand Communications for the Internet Era*, pages 58–61, Richardson,TX, September 2001.

[53] J. M. Tendler, J. S. Dodson, J. S. Fields, H. Le, and B. Sinharoy. POWER4 system microarchitecture. *IBM Journal of Research and Development*, 46(1):5–25, January 2002.

[54] H.-S Wang, L.-S. Peh X. Zhu, and S. Malik. Orion: A power-performance simulator for interconnect networks. In *International Symposium on Microarchitecture*, pages 294–305, Istanbul, Turkey, November 2002.

[55] B. Webb and A. Louri. A class of highly scalable optical crossbar-connected interconnection networks (SOCNs) for parallel computing systems. *IEEE Transactions on Parallel and Distributed Systems*, 11(5):444–458, May 2000.

[56] S. M. Weiss, M. Molinari, and P.M. Fauchet. Temperature stability for silicon-based photonic band-gap structures. *Applied Physics Letters*, 83(10):1980–1982, September 2003.

[57] S. C. Woo, M. Ohara, E. Torrie, J. P. Singh, and A. Gupta. The SPLASH-2 programs: Characterization and methodological considerations. In *International Symposium on Computer Architecture*, pages 24–36, Santa Margherita Ligure, Italy, June 1995.

[58] T. K. Woodward and A. V. Krishnamoorthy. 1-Gb/s integrated optical detectors and receivers in commercial CMOS technologies. *IEEE Journal of Selected Topics on Quantum Electronics*, 5(2):146–156, March–April 1999.

[59] T. Yin, A. M. Pappu, and A. B. Apsel. Low-cost, high-efficiency, and high-speed SiGe phototransistors in commercial BiCMOS. *IEEE Photonics Technology Letters*, 18(1):55–57, January 2006.

[60] I. Young. Intel introduces chip-to-chip optical I/O interconnect prototype. *Technology@Intel Magazine*, pages 3–7, April 2004.

Appendix A: Core Frequency Estimation

If we set core frequencies based simply on the maximum transistor switching capability projected by ITRS [25], processor frequencies would be unrealistically high (e.g. 22.98GHz at 32nm). Indeed, once we factor in power constraints, feasible frequency levels are much lower. We now extrapolate a trend of future CMP core frequencies that respects such power limitations.

Borkar [8] provides a trend of the leakage power (as a fraction of total power consumption at high temperature) for generations down to 50nm technology (Figure 7). Using exponential curve fitting, we obtain the value for our 32nm target. These values, however, assume that no leakage reduction technique, which are expected to reduce leakage by 2.5 times or more in future technologies [8], is applied. Consequently, we assume a 2.5 times reduction in leakage for 50nm and smaller feature sizes (Figure 7).

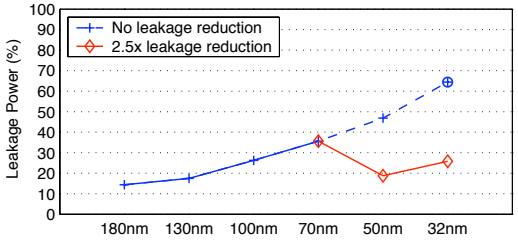


Figure 7: Leakage power (% of total power) projections, taken from Borkar [8] for up to 50nm technology node, and extended to 32nm using exponential curve fitting. Leakage power percentages assuming 2.5 times leakage reduction for 50nm and 32nm technologies are also plotted.

Technology	65nm	45nm	32nm
P_{TOT} (W)	189	198	198
C_g $(E-16F/\mu m)$	6.99	7.35	6.28
V_{dd} (V)	1.1	1	0.9
Frequency (GHz)	4.00	4.40	4.08

Table 9: Summary of ITRS [25] parameters used to calculate the processor frequencies at different technology nodes.

Using the ITRS-projected maximum total power (189W for 65nm, and 198W for subsequent technologies) and the above leakage power projections, we obtain the peak dynamic power. The consumed dynamic power on a chip can be expressed using a basic formula as follows:

$$P_D = V_{dd}^2 C_g W_g f \sum_i A_i k_i$$

where V_{dd} is the power supply voltage, C_g is the total gate capacitance per micron device width $(F/\mu m)$, W_g is the minimum transistor width (μm), f is the core frequency, A_i is the switching activity factor for each capacitive circuit node in the processor, and k_i is the ratio of circuit node capacitance to the minimum NMOS transistor gate capacitance, which depends on the circuit topology and transistor sizing, as well as wire capacitance.

We use ITRS projections to set V_{dd} and C_g. W_g decreases by the scaling factor. In the case of $A_i k_i$, we do not use absolute values. Indeed, by assuming that the number of cores, caches, etc. are doubled with each generation while still retaining the circuit structure, we can reasonably assume that the number of circuit nodes also doubles with each generation, thus doubling the sum with each generation. (We also benefit from the fact that local wire capacitance also scales, resulting in the relative ratio of local wire capacitance to minimum gate capacitance to remain constant.)

Following these trends, substituting the known parameters (Table 9) in the formula, and assuming a 4GHz core frequency at 65nm, we find that the core frequency remains approximately constant in subsequent technologies (Table 9, bottom row). This finding is in agreement with Intel's projections in [9]. Thus, in our 32nm CMP model, we assume a processor core frequency of 4GHz.