Phastlane: A Rapid Transit Optical Routing Network

Mark J. Cianchetti, Joseph C. Kerekes, and David H. Albonesi
Computer Systems Laboratory
Cornell University
{mjc96,jck46,dha7}@cornell.edu

ABSTRACT

Tens and eventually hundreds of processing cores are projected to be integrated onto future microprocessors, making the global interconnect a key component to achieving scalable chip performance within a given power envelope. While CMOS-compatible nanophotonics has emerged as a leading candidate for replacing global wires beyond the 22nm timeframe, on-chip optical interconnect architectures proposed thus far are either limited in scalability or are dependent on comparatively slow electrical control networks.

In this paper, we present Phastlane, a hybrid electrical/optical routing network for future large scale, cache coherent multicore microprocessors. The heart of the Phastlane network is a low-latency optical crossbar that uses simple predecoded source routing to transmit cache-line-sized packets several hops in a single clock cycle under contentionless conditions. When contention exists, the router makes use of electrical buffers and, if necessary, a high speed drop signaling network. Overall, Phastlane achieves 2X better network performance than a state-of-the-art electrical baseline while consuming 80% less network power.

Categories and Subject Descriptors

C.1.2 [**Computer Systems Organization**]: Multiprocessors–Interconnection architectures

General Terms

Design, Performance

Keywords

Nanophotonics, Optical Interconnects, Interconnection Networks, Multicore

1. INTRODUCTION

As the microprocessor industry moves to integrating tens of cores on a single die, the global interconnect becomes a critical performance bottleneck. The ITRS Roadmap [17] projects that metal interconnects will become inadequate to meet the speed and power dissipation requirements of highly scaled ICs beyond 22nm and lists CMOS-compatible optical interconnect technology as a possible solution.

The potential advantages of optical interconnects include high-speed signal propagation, high bandwidth density through time and wavelength division multiplexing (TDM and WDM), and low cross-talk between signal paths [12]. In recent years, significant advances in CMOS-compatible optical components [1, 13, 14, 15, 20, 22] have brought the technology closer to commercial viability. As a result, several teams have proposed detailed architectural designs for multicore chips with integrated optical technology [7, 8, 18, 21], and a number of microprocessor manufacturers are investigating silicon photonic devices and architectures for global on-chip communication in future multicore chips.

However, optical technology has several drawbacks. Optical logic gates [24] and storage (e.g., buffers [23]) are far from mature; thus, control must be implemented in the electrical domain, and buffering likewise must be performed electrically. The lack of a practical multi-layer photonic interconnect scheme analogous to multi-layer wiring means that waveguides must cross [3]. The resulting signal losses can lead to impractically high input power requirements if the system is not carefully architected.

As a result, many proposed optical interconnect architectures are bus-based. For example, the Cornell hybrid electrical/optical interconnect architecture [8] comprises an optical ring that assigns unique wavelengths per node in order to implement a multibus. Every bus cycle, the contents of the buses are optically received, converted to electrical signals, and then handled by logic in the electrical domain (decoded, etc.). The HP Corona crossbar architecture [21] is in fact numerous multiple writer, single reader buses routed in a snake pattern among the nodes.

In both of these approaches, the physical topology (ring or snake) is chosen to avoid waveguide crossings, and a bus approach prevents control functionality from limiting data transmission speed. The Columbia optical network [18] is one of the few that proposes on-chip optical switches. The network consists of a 2D grid of optical waveguides with optical resonators at intersecting points to perform turns. An electrical sub-network sets up the switches in advance of data transmission and tears down the network thereafter. Once the path is set up, communication proceeds between source and destination. The small number of waveguides in each channel limits the number of crossings. Moreover, the setup is done in advance so that the control circuitry does not limit transmission speed. However, the network must transmit a large amount of data to amortize the relatively high latency of the electrical setup/teardown network, making the network unsuitable for a typical cache coherent shared memory system where the unit of transfer is a cache line.

ISCA'09, June 20–24, 2009, Austin, Texas, USA.

In this paper, we present *Phastlane*[1], a 2D on-chip grid of optical crossbar switches targeted to future cache coherent multicore microprocessors. Each Phastlane router carefully balances crossing losses with bandwidth density and employs simple predecoded source routing. These features enable high speed transmission across multiple hops in a single cycle while using less power than a state-of-the-art electrical network. If an output port is blocked, or the distance to be traversed is too long, the switch receives, converts, and electrically buffers the packet for later delivery. Each switch also sets up a unique optical return path to immediately signal the source if a packet is eventually dropped. Compared to an aggressively designed electrical baseline, Phastlane achieves an overall 2X network speedup with 80% lower power consumption.

The rest of this paper is organized as follows. In the next section, we describe the overall Phastlane architecture, including the use of optical resonators and simple control to route packets within a Phastlane router. Then in Section 3 we evaluate the latency, optical power, and area tradeoffs for different architectural parameters and technology scaling scenarios. In Section 4 we discuss our evaluation methodology and we present our results in Section 5. Finally, we discuss related work in Section 6, and we conclude and present possibilities for future research in Section 7.

2. PHASTLANE ARCHITECTURE

One advantage of on-chip silicon photonics is its low latency transmission over distances long enough to amortize the costs of modulation, detection, and conversion. In 16nm technology, the distance beyond which optics achieves lower delay than optimally repeatered wires is expected to be 1-2mm [4], making optical transmission profitable for even single hop network traversals. Our goal, therefore, was to architect an optical switch network that matches the latency and bandwidth of a state-of-the-art electrical network at short distances, exploits the ability of optics to traverse multiple hops in a single cycle in the case of no contention, and uses a cache line as the unit of transfer. Meeting these goals required simplicity in the control path. In particular, we opted for dimension-order routing, fixed-priority arbitration, and simply dropping a packet when buffer space is unavailable. Although these choices impact network efficiency, they permit optical data transmission over long distances to be minimally impeded by control circuitry.

Our design targets cache coherent multicore processors in the 16nm generation with tens to hundreds of cores and a highly-interleaved, main memory using multiple on-chip memory controllers. High bandwidth density and low latency are simultaneously met using WDM to pack many bits into each waveguide and simple predecoded source routing and fixed priority arbitration.

The optical components of the Phastlane 8x8 mesh network are located on a separate chip integrated into a 3D structure with the processor die. Figure 1 shows one of the 64 nodes of the Phastlane network. The node includes one or more processing cores, a two-level cache hierarchy, a memory controller (MC), and the electrical components of the router. The 64 MCs are interleaved on a cache line basis with high bandwidth serial optical links – like those proposed for Corona [21] – connecting each MC to off-chip DRAM.

2.1 Router Microarchitecture

Figure 2 shows a portion of the optical components of a single Phastlane router. Only a fraction of the input and output waveguides and circuitry are shown for clarity. Resonator/receiver pairs at each of the four (N, S, E, and W) input ports receive packets

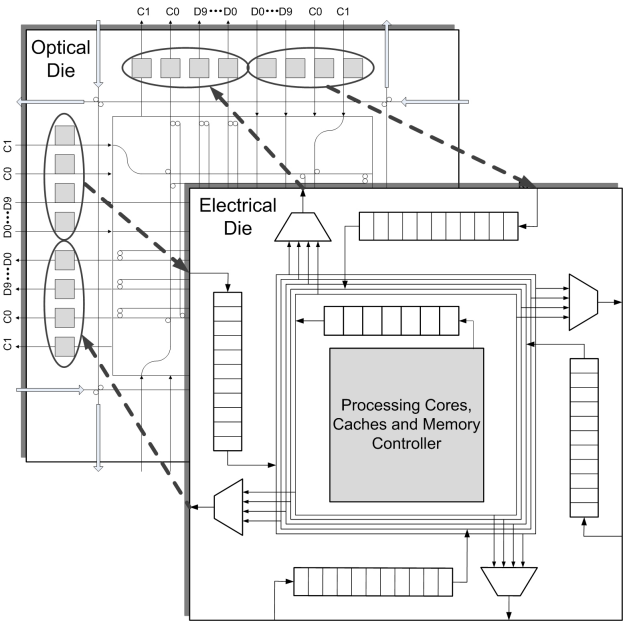

Figure 1: Overall diagram of a Phastlane router showing the optical and electrical dies, including optical receiver and driver connections to the electrical input buffers and output multiplexers. The input buffers capture incoming packets only when they are blocked from an optical output port.

that are either destined for this node or that are blocked. Transmitter/modulator pairs at each output port drive packets from the local node buffer or from one of the input port buffers. Incoming packets that turn left or right pass through the resonators located inside the router to the coupled perpendicular waveguides.

Unlike the Columbia approach [18], Phastlane has no electrical setup/teardown network. Rather, precomputed control bits for each router are optically transmitted in separate waveguides in parallel with the data, and these bits are used to implement simple dimension-order routing and fixed priority arbitration. Each packet consists of a single flit, which contains a full cache line (64 bytes) of Data, the Address, Operation Type and Source ID bits, Error Detection/Correction and miscellaneous bits, and Router Control bits for each of the intermediate routers as well as the destination router. Ten waveguides (D0-D9 in Figure 2) assuming 64-way WDM[2] transmit the entire packet with the exception of the Router Control, which is evenly divided between two additional waveguides (C0 and C1) as shown in Figure 3. The Router Control consists of Straight, Left, Right, Local, and Multicast routing control bits for each of the up to 14 routers that may be traversed in the 8x8 network. The first three bits map to the three possible output ports. The Local bit indicates whether the router should accept the packet for its local node. The Multicast bit indicates a multicast operation as discussed in Section 2.1.4.

Returning to Figure 2, consider a packet arriving at the S input port. The C0 waveguide contains the five control bits for this router on wavelengths $\lambda_1 - \lambda_5$ (Group 1), and up to six other sets of control bits on $\lambda_6 - \lambda_{35}$ (Groups 2-7). All of the C0 bits are received by the resonator/receiver pairs shown on the C0 S input port. The Group 1 control bits are used to route the packet through the

[1]Analogous to a Massachusetts Turnpike Fastlane toll booth, unblocked packets can rapidly transit through a Phastlane router.

[2]In Section 3, we investigate tradeoffs in varying the number of wavelengths as well as other parameters.

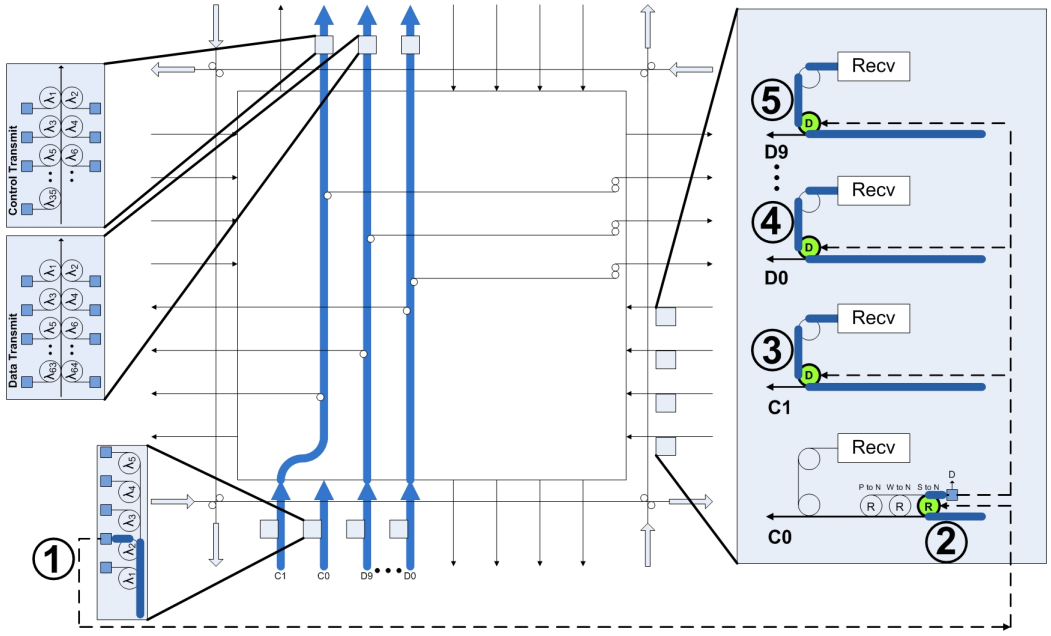

Figure 2: Phastlane optical switch, showing a subset of the signal paths for an incoming packet on the S port and the process of receiving an incoming blocked packet on the E input port.

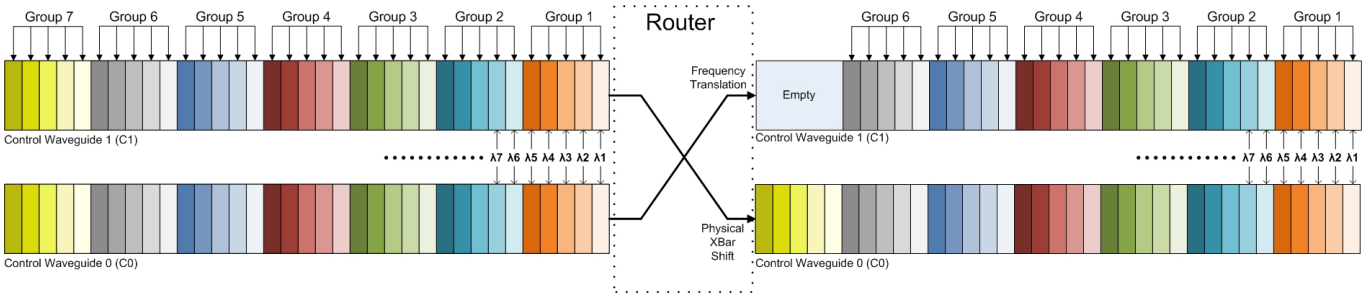

Figure 3: C0 and C1 control waveguides. As inputs, they together hold up to 14 groups of five control bits for each router. The Group 1 bits in the C0 waveguide are used to route the packet through the current router. On exiting the router, the Group 2-7 bits are frequency translated to the Group 1-6 positions and output on the C1 waveguide, while the C1 waveguide is physically shifted to the C0 position at the output port.

switch while the remaining control bits are frequency translated as described below. If the Group 1 Local bit is set, resonator/receiver pairs on D0-D9 and C1 are activated to receive the packet. Otherwise, the packet enters the router and continues on the straightline path towards the N output port. The next set of resonators are activated by the Left bit while the last set are activated by the Right bit. If neither of these are set, the Straight bit is set and the packet exits through the N port. As shown in Figure 3, the C1 waveguide is physically shifted to assume the C0 position at the corresponding output port. The remaining $\lambda_6 - \lambda_{35}$ control bits in C0 are frequency translated to $\lambda_1 - \lambda_{30}$ and are transmitted on the C1 waveguide of the selected output port. This physical shift and frequency translation lines up the control fields for subsequent routers.

Since the straightline paths through the router have priority over turns[3], the C0 Group 1 Straight bit from the S port, when set, blocks incoming packets from the E and W ports from exiting through the N port. For example, if the Right bit for the E input port is set, then this packet must be received or dropped – depending on the available buffer space – to avoid contention with the packet traveling from the S input to the N output. The resonator/receiver pairs labelled ① and ② in Figure 2 detect this situation and receive the E input packet (③-⑤). The Group 1 Straight bit from the S input port (①) activates ② which receives the set Group 1 Right bit off the C0 waveguide on the E input port. This received bit in

[3]While fixed-priority arbitration is inherently unfair, for dimension-order routing where a single turn is required, we found that a more complicated scheme such as round-robin yielded no performance advantage over fixed-priority, while increasing crossbar latency.

turn activates ③-⑤ which receive the packet on the E input port, preventing it from contending with the packet traveling from S to N. By using predecoded fields to directly control turn resonators and to receive lower priority packets, data transmission through the router crossbar is minimally disrupted by control complexity. This characteristic permits low latency transmission through the switch.

2.1.1 *Electrical Buffers and Arbitration*

Each router has five sets of buffers in the electrical domain, four corresponding to the N, S, E, and W input ports and one for the local node (Figure 1). A newly arriving blocked packet is received, translated, and placed in the corresponding buffer if there is space. Buffered packets have priority for output ports over newly arriving packets. A rotating priority arbiter selects up to four packets from these queues to transmit to the four output ports. Any incoming packets that conflict with a buffered packet for an output port are received and buffered if there is space. When no buffered packet competes for an output port then the aforementioned fixed-priority scheme determines the winner among the newly arriving packets.

2.1.2 *Drop Signal Return Path*

Phastlane's simplified optical-based control approach leads to dropping packets if an output port is blocked and an input buffer is full. In order to be able to rapidly signal a dropped packet condition, depending on the situation, one of three actions are taken when a packet arrives at an intermediate router:

- The packet is not blocked; in this case, the router registers the received and translated Straight, Left, and Right bits in order to set up a drop signal return path in the next cycle in case the packet is eventually dropped;
- The packet is blocked but the input port buffer is not full; in this case, the router receives, translates, and buffers the packet and assumes responsibility for its delivery;
- The packet is blocked and the input port buffer is full; in this case, the packet is dropped and the router transmits an asserted Packet Dropped signal and the router's Node ID on the return path output port in the next cycle.

The network includes return paths for signaling the source that its packet was dropped by a particular node[4]. The source may be the original sender of the packet or an intermediate router that buffered the packet (second scenario above). As a packet moves through the network, each router registers the C0 Group 1 Straight, Left, and Right bits. In the next cycle, each router uses these signals to activate the correct return path in case a drop condition needs to be communicated to the source. The router that drops the packet transmits an asserted Packet Dropped signal and its six-bit Node ID on the return path waveguide. These signals propagate through the return path constructed by each router back to the source. The source takes appropriate action (e.g., backoff and resend) upon receiving the Packet Dropped signal. If a source does not receive a Packet Dropped signal in the cycle immediately following transmission, then either the packet arrived at its destination or an interim node has assumed responsibility for its delivery.

The circuitry for constructing this path is straightforward given the predecoded control fields. Referring again to Figure 2, the large arrows show the return path input and output ports. Return paths flow in the opposite direction that packets travel through the router. For example, a packet that entered the N port and exited the E port would have the return path shown in the upper right corner of the router activated in the following cycle. The latched value of the Group 1 Left input from the N port controls the resonator shown in that corner, which makes a return path connection between the E and N ports. If the packet was dropped at this router, then transmitter/modulator pairs connected to the N return path output transmit the seven-bit optical signal in the following cycle.

[4]By definition, each return path is unique and cannot overlap with the return path of any other packet in the same cycle.

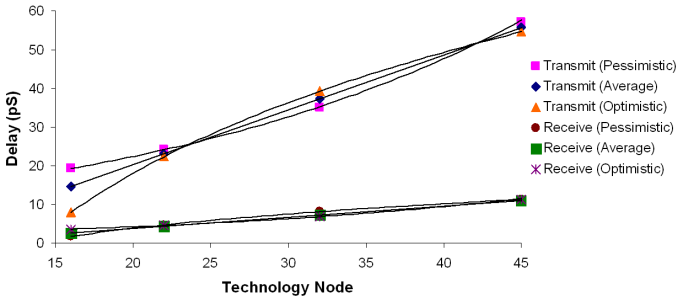

Figure 4: Optimistic, average, and pessimistic scaling trends for transmit and receive delays.

2.1.3 *Pipelined Transmission in Large Networks*

For large networks, such as the 8x8 mesh that we investigate, single cycle corner-to-corner transmission is infeasible at high network clock rates. For these networks, the transmission is pipelined in multiple cycles, using interim nodes to buffer the packet. In our network at 16nm under average scaling assumptions (Section 3), five hops can be traversed in one cycle when taking into account the worst-case situation of contention at every router and late arrival of the packet compared to competing packets. For transmissions requiring more than five hops, the source picks the nodes five and ten hops away along dimension order as interim destinations. The Local bits for the interim nodes and the final destination are set. Each interim node detects that their Local bit is set and places the packet in the input buffer if there is room, and otherwise drops the packet. For the former case, upon detecting that another Local bit is set, it assumes responsibility for sending the packet to either the next interim node or the final destination. If the packet is blocked and buffered in an intermediate node before reaching an interim node, the intermediate node may choose to bypass the original interim node and send the packet further (perhaps directly to its destination). It does so by modifying the Local bits of the packet.

2.1.4 *Multicast Operations*

In a snoopy cache-coherent system, L2 miss requests and coherence messages such as invalidates are broadcast to every node. In Phastlane, a broadcast consists of multiple multicast packets. Multicast packets have a set Multicast bit in the 5-bit router control field. The broadcasting node sends up to 16 multicast messages (eight if it is located on the top or bottom rows of the network).

For a given router, if the Group 1 Multicast bit is set but the Local bit is not, the router receives a portion of the power transmitted on the input lines through separate broadcast resonator/receivers. Since only a portion of the power is extracted, the packet continues through the selected output port to the next router in the absence of contention. If the Group 1 Local bit is set, the packet is received through the local receive resonator. If the Group 1 Multicast bit is also set, it delivers it to the local node. Otherwise, this router is

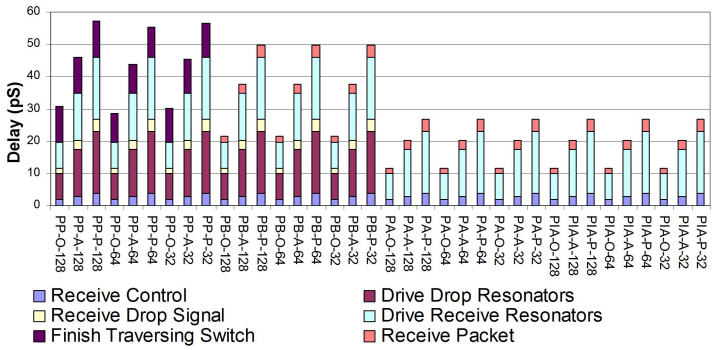

Figure 5: Component delays of the critical paths (PP, PB, PA, and PIA) through the Phastlane router under different scaling assumptions (Optimistic, Average and Pessimistic).

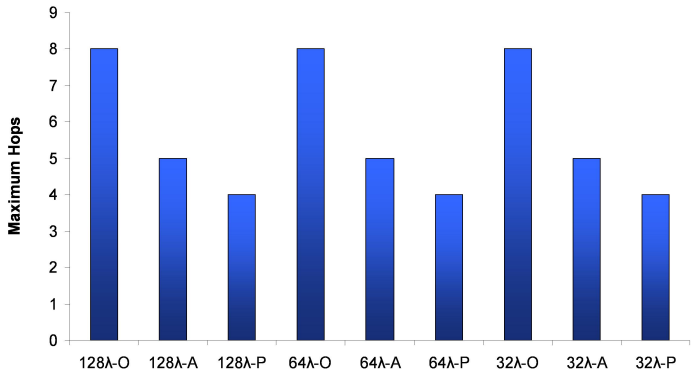

Figure 6: Maximum number of hops a packet can travel in a single 4GHz cycle for different number of wavelengths and scaling assumptions.

merely an interim node for a multicast packet. In this case, it either drops the packet if it has no buffer space available, or buffers the packet and assumes responsibility for completing the multicast. If neither bit is set, it simply routes the packet without receiving it.

If a multicast packet is dropped, the source examines the Node ID of the dropped packet return path and determines which nodes already received the multicast message. It clears the Multicast bits for these nodes for the resent packet.

3. ROUTER DESIGN SPACE EXPLORATION

In this section, we investigate the latency, area, and optical power tradeoffs in designing the Phastlane architecture discussed in the previous section while varying microarchitectural parameters under different technology scaling assumptions.

3.1 Latency

To evaluate router latency tradeoffs, we derive optimistic, average, and pessimistic scaling assumptions for the optical component delays at 16nm (Figure 4). We start with the analysis of Kirman et al. [8] in which each of the optical transmit and receive components were scaled from 45nm to 22nm. We create optimistic, average, and pessimistic scaling scenarios for 16nm by curve fitting the Kirman et al. data to the logorithmic, linear, and exponential functions. Figure 4 shows the aggregate transmit and receive scaling trends. At 16nm, the transmit and receive delays range from 8.0-19.4ps and 1.8-3.7ps, respectively. The waveguide propagation delay is assumed to remain constant at 10.45ps/mm [8].

We calculate the number of hops that a packet can travel through the optical network in a 4GHz processor cycle by analyzing the various potential critical delay paths through the router. Figure 5 breaks down the delays for the following internal router operations:

Packet Pass (PP): A packet passes to a router output port. Assuming that other packets contend for the output port, we determined that the critical path involves the incoming packet removing these contending packets by forcing them to be received at their input port as discussed in Section 2.1. The delay breaks down as follows: (a) receiving the packet Router Control bits; (b) driving the C0 Group 1 resonators of the blocked packets; (c) the signal from (b) driving the receive resonators of the blocked packets, thereby clearing the output port; and (d) traversing the remainder of the switch.

Packet Block (PB): A packet gets blocked and buffered at the switch. The delay is similar to the Packet Pass situation, except that the time to traverse the switch is replaced by the time to receive the blocked packet.

Packet Accept (PA) and *Packet Interim Accept (PIA)*: A packet is accepted at its destination or an interim node. The overall delay is composed of the time to (a) receive the C0 control signals; (b) drive the receive resonators; and (c) receive the packet.

From Figure 5, we observe that the number of wavelengths has little impact on delay and that most of the delay involves driving the resonators. The time to pass through the router exceeds the packet block time. Accepting a packet is the fastest of these three operations. Other delays, such as creating write-enable signals for buffering if a packet is dropped, accepted or interim accepted, were determined to be less critical than these other path delays.

Based on this analysis, we determined that the longest network delay occurs when a packet is injected at the source node, travels the maximum number of hops, and is then accepted. If X is the number of routers between the source and destination, then there will be X Packet Pass delays and $X+1$ inter-router waveguide link delays. By including the delay to drive the modulators at the source, the Packet Accept delay at the destination, and register overhead and clock skew, we can solve for X and determine how many hops can be traversed in one clock. Figure 6 shows that for optimistic, average, and pessimistic scaling assumptions, eight, five, and four hops, respectively, can be traversed in a 4GHz clock cycle independent of the number of wavelengths. We demonstrate, however, in Section 5 that the Phastlane network performance is relatively insensitive to the degree of component delay scaling to 16nm.

3.2 Peak Optical Power

With a fixed packet size, the peak electrical power dissipated by the Phastlane network does not substantially change with parameters such as the number of wavelengths or the maximum number of hops that can be traversed in a cycle. However, the peak optical power can vary considerably as the number of wavelengths and maximum distance are varied for different crossing efficien-

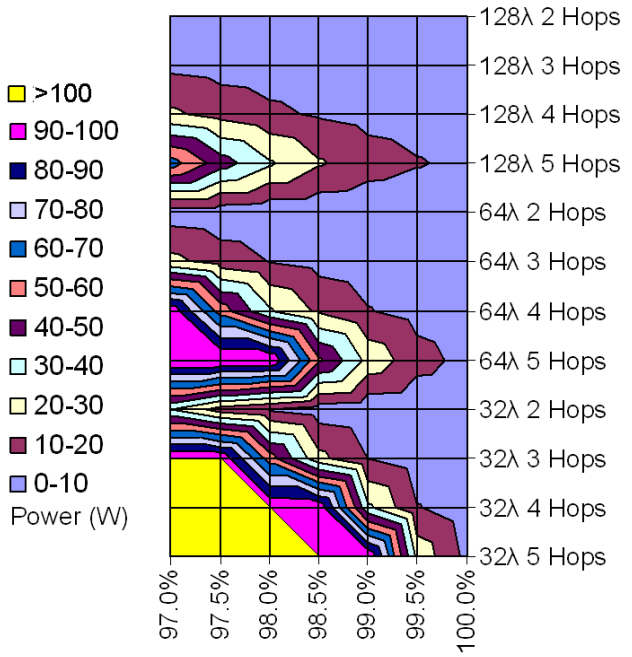

Figure 7: Contour plot of the peak optical power as a function of the crossing efficiency, the number of wavelengths, and the maximum number of hops that can be traversed in a cycle. With more hops, more input optical power is required for packet transmission.

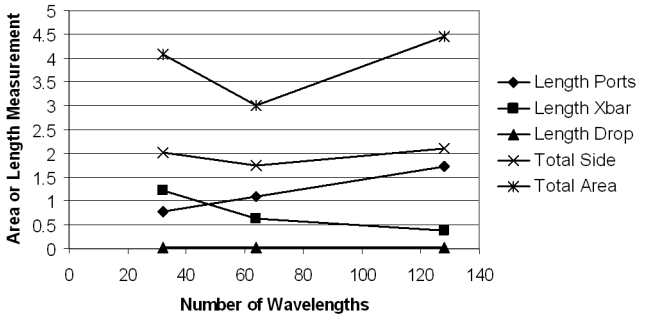

Figure 8: Impact of the number of wavelengths on different router area components and the total area. The best balance of port length and internal router length occurs at 64 wavelengths.

Flits Per Packet	1 (80 Bytes)
Packet Payload WDM	64
Packet Payload Waveguides	10
Routing Function	Dimension-Order
Packet Control Bits	70
Packet Control WDM	35
Packet Control Waveguides	2
Buffer Entries in NIC	50
Max Hops Per Cycle	4, 5, or 8
Node Transmit Arbitration	Rotating Priority
Network Path Arbitration	Fixed Priority

Table 1: Optical network configuration.

cies. Figure 7 shows a contour plot of these relationships. The peak optical power – the maximum optical power that can occur in a single cycle – occurs when every input port in every router simultaneously receives a multicast packet from its nearest neighbor, and all of these packets turn in the same direction (right or left) to an open output port. At the same time, all return paths are being used to signal a dropped packet and all buffers are full and arbitrating for an open output port in the next cycle. This situation creates the maximum number of crossings and activated components. As shown in Figure 7, with 32 wavelengths, due to the excessive number of crossings per router, the network requires either very high crossing efficiency (at least 99%) or a limit on the maximum distance (2-3 hops) to keep the peak optical power to a reasonable value. Limiting the maximum distance reduces the optical power due to fewer total crossings and fewer multicast resonators that extract power from the packet. By moving to 64 wavelengths, a four-hop network requires a peak 32W of optical power at 98% crossing efficiency, while moving to 128 wavelengths permits a five-hop network for the same 32W of power. However, we demonstrate in Section 5 that a five-hop network achieves marginally better performance than a four-hop one. Therefore, a better tradeoff when increasing the number of wavelengths from 64 to 128 is to maintain a four-hop network and reduce the required peak optical input power, e.g., from 32W to 15W with 98% crossing efficiency.

3.3 Area

For cost reasons, the optical component die should not exceed the area of the processor die; otherwise, the latter will need to artifically increase in size in order to line up the related components. Moreover, the electrical components of the router, such as the resonator drivers and receiver amplifiers, should only marginally increase the area of the processor die.

To estimate the area of the processor die, we adopted the methodology of Kumar et al. [10]. For a single processor core with 64KB L1 caches, a 2MB L2 cache, and a Memory Controller the total area is approximately 3.5mm^2. For two cores and four cores sharing an L2 cache, the area is approximately 4.5mm^2 and 6.5mm^2, respectively.

For the optical router, the number of wavelengths impacts router area in two ways. First, the total number of waveguides and turn resonators decreases linearly as the number of wavelengths increases. However, the length of the input ports increases linearly since more resonator/receiver pairs must be attached to the same waveguide. Figure 8 shows how these two factors trade off. The area "sweet spot" is realized with 64 wavelengths for our packet size. For a single core with private L1 and L2 caches, we estimate that 64 wavelengths are necessary to match the area of the processor die. With larger dual and quad core nodes, 32 or 128 wavelengths will also meet die size constraints. The transmitters and receivers require a negligibly small area on the electrical die.

From this analysis, we arrived at the configuration shown in Table 1. The routing bits are packaged in two waveguides using 35-way WDM, while the payload occupies ten waveguides using 64-way WDM. In Section 5, we explore the performance and power of four-hop, five-hop, and eight-hop networks given different scaling scenarios. First, we present our evaluation methodology.

Flits per Packet	1 (80 Bytes)
Routing Function	Dimension-Order
Number of VCs per Port	10
Number of Entries per VC	1
Wait for Tail Credit	YES
VC_Allocator	ISLIP [11]
SW_Allocator	ISLIP [11]
Total Router Delay	2 or 3 cycles
Input Speedup	4
Output Speedup	1
Buffer Entries in NIC	50

Table 2: Baseline electrical router parameters.

Benchmark	Experimental Data Set
Barnes	64 K particles
Cholesky	tk29.O
FFT	4 M particles
LU	2048x2048 matrix
Ocean	2050x2050 grid
Radix	64 M integers
Raytrace	balls4
Water-NSquared	512 molecules
Water-Spatial	512 molecules
FMM	512 K particles

Table 3: SPLASH2 benchmarks and input data sets.

4. EVALUATION METHODOLOGY

To evaluate our proposed optical network, we developed a cycle-accurate network packet simulator that models components down to the flit-level. The simulator generates traffic based on a set of input traces that designate per node packet injections. All network components and functionality described in Section 2 are fully modeled, including finite buffering in the network-interface controller. In order to do a power comparison with the electrical baseline, we also model dynamic power consumption and static leakage power in a manner similar to [8].

We evaluate the electrical baseline network using a modified version of Booksim [5] augmented with dynamic and static leakage power models. The models use CACTI for buffers, and [2] for all other components. We also integrated finite NIC buffering as well as Virtual Circuit Tree Multicasting [6] to perform packet broadcasts. Finally, we changed Booksim to input the same trace files used for our optical simulator.

The electrical baseline is an aggressive router optimized for both latency and bandwidth. The router assumes a virtual-channel architecture with the parameters shown in Table 2. In order to perform a fair performance comparison with our optical configurations, we assume both low latency and high saturation bandwidth for the electrical network. We reduce serialization latency by using a packet size of one flit, the same as in Phastlane. Doing so also gives no bandwidth density advantage to the optical network. We further assume that pipeline speculation and route-lookahead [19] reduce the per hop router latency of the baseline electrical router to 2-3 cycles for every flit. Finally, we assume that the electrical baseline can accept an input flit on each input port each cycle. These flits do not require the cross-bar and instead can be directly accepted by the processor one cycle after the flit enters the router.

Simulated Cache Sizes	32KB L1I, 32KB L1D, 256KB L2
Actual Cache Sizes	64KB L1I, 64KB L1D, 2MB L2
Cache Associativity	4 Way L1, 16 Way L2
Block Size	32B L1, 64B L2
Memory Latency	80 cycles

Table 4: Cache and memory controller parameters.

We evaluate SPLASH2 benchmarks and synthetic traffic workloads. By varying the injection rates of the synthetic benchmarks, we obtain saturation bandwidth and average packet latencies. We created SPLASH2 traces using the SESC simulator [16]. Each benchmark was run to completion with the input sets shown in Table 3. The modeled system consists of 64 cores with private L1 and L2 caches. Each core is 4-way out-of-order and has the cache and memory parameters shown in Table 4. As is typical when using SPLASH2 for network studies, the cache sizes are reduced to obtain sufficient network traffic. Finally, we assume a 16nm technology node operating at a 4GHz processor and network clock with a supply voltage of 1.0V.

5. RESULTS

In this section, we compare the performance and power consumption of the four-hop, five-hop, and eight-hop optical networks – which correspond to pessimistic, average, and optimistic scaling assumptions – to the baseline electrical network with a three cycle latency. We also consider optical configurations with more buffering. While the baseline four-hop optical configuration has 10 buffers at each of the input ports as well as the local node output, we also evaluate buffer sizes of 32, 64, and infinite. Finally, we evaluate an electrical router with a very aggressive two cycle latency. The processor and network frequency are assumed to be 4GHz for all configurations.

We first evaluate average packet latency and saturation bandwidth using the synthetic workloads. The results, shown in Figure 9, highlight the significantly lower latency of the optical networks compared to the conventional electrical networks, even the two-cycle design, while providing slightly better saturation bandwidth. The Phastlane network achieves approximately 5-10X lower latency than the electrical networks. Moreover, for these traffic patterns, the four-hop and five-hop networks provide about the same network latency as the faster eight-hop network.

Figure 10 shows network speedup for the SPLASH2 benchmarks. For six of the benchmarks, the optical four-hop network achieves a network speedup of over 1.5X (and by over 2.8X for three benchmarks) compared to the electrical network. The five-hop and eight-hop networks perform marginally better than the four-hop network; this result indicates that a pessimistic scaling of the optical components will not dramatically impact performance. While overall, the optical configurations far outperform the baseline electrical network, the performance of Barnes, Cholesky, Ocean, and FMM is highly sensitive to the amount of buffering. With Ocean and FMM in particular, limited buffering results in many dropped packets. These packets that are ultimately dropped steal resources from other packets, and also must be retransmitted, which impacts network performance. The four-hop network requires about 64 buffers to match the baseline electrical network performance with Ocean and about 32 buffers with FMM. This result highlights a weakness of our simplified network control: with insufficient buffering, some traffic patterns may lead to many dropped packets that saturate the

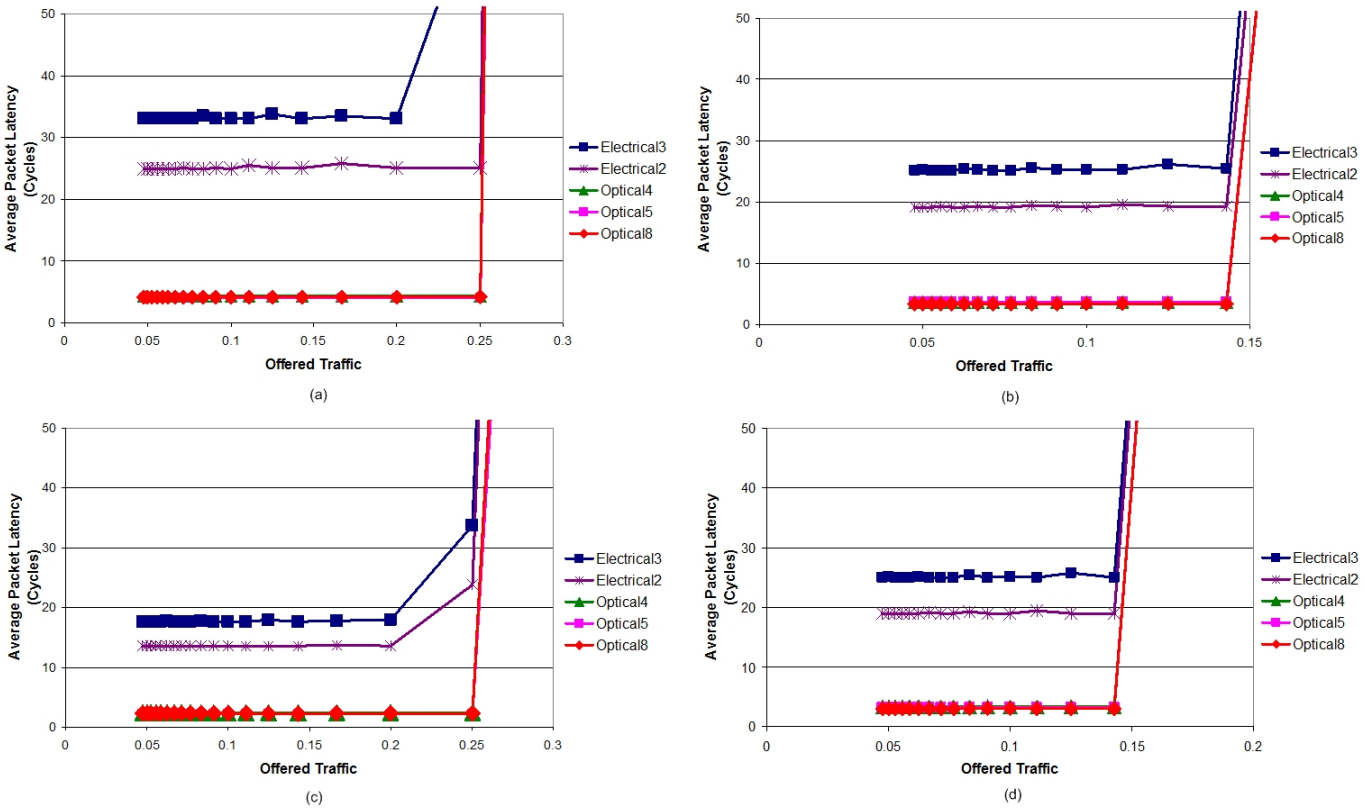

Figure 9: Average packet latency as a function of injection rate for (a) Bit Comp, (b) Bit Reverse, (c) Shuffle, and (d) Transpose.

network. Future work will investigate more sophisticated buffer management schemes to reduce buffering requirements.

The optical configurations are far more power efficient than the electrical network for all SPLASH2 benchmarks. Figure 11 shows that the power consumption of the four-hop and five-hop optical networks is at least 70% less than that of the electrical network for all benchmarks. In addition, the average optical power is far less than the peak calculated in Section 3.2. The eight-hop network consumes much more power than the four-hop and five-hop networks, especially for benchmarks with multicast transfers that take advantage of the additional per-cycle distance. While packets are buffered less often which reduces the electrical power, the average transmit power increases sharply due to additional crossing losses and the additional receivers to drive.

The four-hop network appears to be the best design choice when considering network performance, average electrical+optical power, and peak optical power. Overall, the four-hop network achieves a 2X network speedup over the electrical network with 80% lower power consumption.

6. RELATED WORK

Several on-chip interconnect architectures have been proposed that leverage CMOS-compatible photonics for future multicore microprocessors. Kirman et al. [8] propose a hierarchical interconnect for communication among 64 cores in 32nm technology. A group of four cores and a shared L2 cache communicates with four others groups through an electrical switch. The four 16-processor nodes in turn communicate using an optical ring that implements a bus protocol. Each node writes to the bus using a unique wavelength, which obviates the need for arbitration, and information is read by coupling a percentage of the power from each signal.

Vantrease et al. [21] also propose optical buses for communication among 256 cores in 16nm technology. Similar to [8], multiple cores are grouped as a node and communicate through an electrical sub-network. Inter-node communication occurs through a set of multiple-writer, single-reader buses (one for each node) that together form a crossbar. Optical arbitration resolves conflicts for writing a given bus. An optical token travels around a special arbitration waveguide, and a node reads and removes the token before communicating with its intended target. Chip-to-chip serial optical links communicate with main memory modules that are divided among the network nodes.

Perhaps the closest work to ours is the optical 2D network proposed by Shacham et al. [18]. Data transfer occurs through a grid of waveguides with resonators at crosspoints for turns. Control is handled by an electrical set-up/tear-down network. To enable data transfer, a packet is sent on the electrical network which moves toward the destination and reserves the optical switches along its route. When this path is established, the source transfers data at high bandwidth using the optical network. Finally, a packet is sent in the electrical network to tear-down the established path.

Kumar et al. [9] propose Express Virtual Channels to reduce packet latency in an electrical router beyond techniques such as lookahead routing and pipeline bypassing and speculation [19]. Packets within these channels can bypass the router pipeline.

Our approach leverages elements of each of these prior propos-

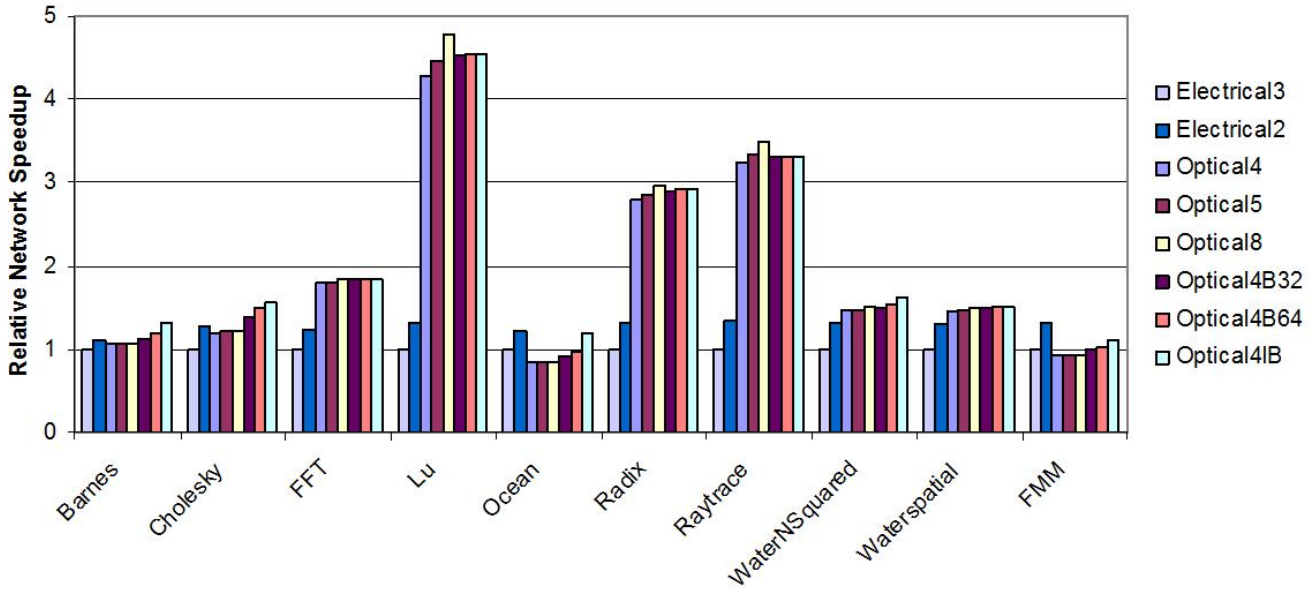

Figure 10: Network speedup of the optical network configurations with four, five, and eight hops relative to the baseline electrical network. Optical4B32 and Optical4B64 have 32 and 64 buffer entries, respectively, compared to 10 for Optical4, while Optical4IB has infinite buffering. Results for the two cycle electrical router are also shown.

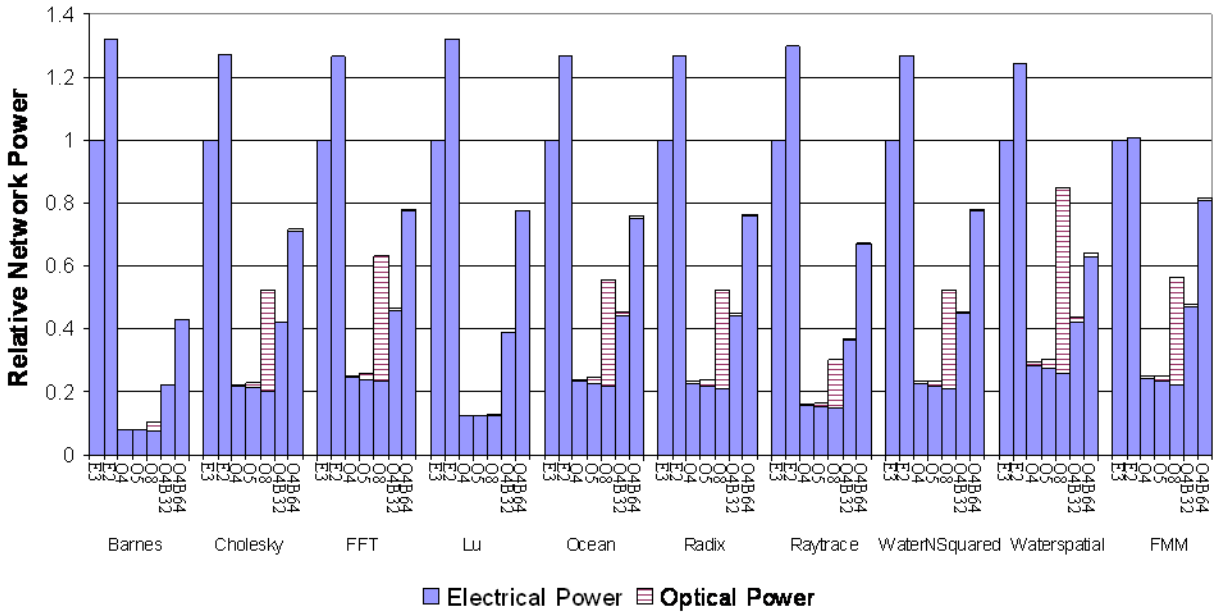

Figure 11: Network power for the optical network configurations compared to the electrical networks.

als. Like Shacham et al., we use a grid of waveguides with turn resonators, but there are several important distinctions between our proposals, some of which are due to differences in data payload size. We rely on only WDM to pack a narrow packet into one cycle, while they use WDM and TDM to achieve very high bandwidth transfer of a much greater amount of data. We optically send control along with the data to set up the router switches on the fly rather than use a slower electrical control network. Like Kirman et al. and Vantrease et al., we target snoopy cache-coherence multicore systems, but our networks are quite different (switch-based rather than bus-based). Finally, as with Express Virtual Channels, we seek to reduce packet latency but we do without special dedicated express

lanes. Rather, we use simple control to exploit the capability of optics to travel multiple hops in a single cycle.

7. CONCLUSIONS

With the integration of tens and eventually hundreds of cores on a microprocessor die, the global interconnect becomes a critical performance bottleneck. CMOS-compatible optical technology has emerged as a potential solution to this problem beyond the 22nm timeframe. Yet, current architectural proposals are limited in scalability or require comparably slow electrical control networks.

In this paper, we present Phastlane, a routing network that exploits the low latency of nanophotonics to permit packets to traverse several hops in the network under contentionless conditions. Phastlane uses simple optical-level, source-based, router control to avoid the control path from becoming a latency bottleneck. In cases of contention, packets are received and electrically buffered or dropped and retransmitted under buffer-full conditions. On a set of ten SPLASH2 benchmarks, Phastlane achieves 2X better network performance while consuming 80% less network power. For future work, we plan to investigate alternatives to the drop network and simple rotating priority arbitration of the electrical buffers.

Acknowledgements

The authors thank the anonymous reviewers for their helpful comments; Rajeev Dokania, Meyrem Kirman, Nevin Kirman, and Sasikanth Manipatruni for information and guidance; and Paula Petrica, Matt Watkins, and Jonathan Winter for help preparing this paper. This research was supported by NSF grants CNS-0708788, CCF-0732300, CCF-0541321, and CCR-0304574; a National Grid Graduate Fellowship; a Semiconductor Research Corporation Education Alliance Grant; and equipment donations from Intel.

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):2867–2869, 2004.

[2] J. Balfour and William Dally. Design Tradeoffs for Tiled CMP On-Chip Networks. In *International Conference on Supercomputing*, June 2008.

[3] W. Bogaerts, P. Dumon, D. V. Thourhout, and R. Baets. Low-Loss, Low-Cross-Talk Crossings for Silicon-on-Insulator Nanophotonic Waveguides. *Optics Letters*, 32(19):2801–2803, 2007.

[4] G. Chen, Hui Chen, Mikhail Haurylau, Nicholas Nelson, Philippe Fauchet, Eby Friedman, and David Albonesi. Predictions of CMOS Compatible On-Chip Optical Interconnect. In *Proceedings of the International Workshop on System Level Interconnect*, April 2005.

[5] W. Dally and B. Towles. Principles and Practices Of Interconnection Networks. Morgan Kaufmann, 2007.

[6] N. E. Jerger, L.-S. Peh, and M. Lipasti. Virtual Circuit Tree Multicasting: A Case for On-Chip Hardware Multicast Support. In *Proc. 35th IEEE/ACM International Symposium on Computer Architecture*, June 2008.

[7] A. Joshi, C. Batten, Y.-J. Kwon, S. Beamer, I. Shamim, K. Asanović, and V. Stojanović. Silicon-Photonic Clos Networks for Global On-Chip Communication. In *Third ACM/IEEE International Symposium on Networks-on-Chip*, May 2009.

[8] N. Kirman, M. Kirman, R. K. Dokania, J. F. Martinez, A. B. Apsel, M. A. Watkins, and D. H. Albonesi. Leveraging Optical Technology in Future Bus-based Chip Multiprocessors. In *Proc. IEEE/ACM 39th Annual International Symposium on Microarchitecture*, Dec. 2006.

[9] A. Kumar, Li-Shiuan Peh, Partha Kundu, and Niraj Jha. Express Virtual Channels: Towards the Ideal Interconnection Fabric. In *Proc. 34th IEEE/ACM International Symposium on Computer Architecture*, June 2007.

[10] R. Kumar, Dean Tullsen, and Norman Jouppi. Core Architecture Optimization for Heterogeneous Chip Multiprocessors. In *Parallel Architectures and Compilation Techniques*, Sept. 2006.

[11] N. McKeown. The iSLIP Scheduling Algorithm for Input-Queued Switches. *ACM Transactions on Networking*, 7(2):188–201, 1999.

[12] D. Miller. Rationale and Challenges for Optical Interconnects to Electronic Chips. *Proceedings of the IEEE*, 88(6):728–749, 2000.

[13] M. Paniccia, V. Krutul, R. Jones, O. Cohen, J. Bowers, A. Fang, and H. Park. A Hybrid Silicon Laser. White Paper, Intel Corporation, 2006.

[14] H. Park, Y. hao Kuo, A. W. Fang, R. Jones, O. Cohen, M. J. Paniccia, and J. E. Bowers. A Hybrid AlGaInAs-Silicon Evanescent Preamplifier and Photodetector. *Optics Express*, 15(21):13539–13546, 2007.

[15] K. Preston, P. Dong, B. Schmidt, and M. Lipson. High-Speed All-Optical Modulation Using Polycrystalline Silicon Microring Resonators. *Applied Physics Letters*, 92(15):151104, 2008.

[16] J. Renau, B. Fraguela, J. Tuck, W. Liu, M. Prvulovic, L. Ceze, S. Sarangi, P. Sack, K. Strauss, and P. Montesinos. Sesc simulator. http://sesc.sourceforge.net, 2005.

[17] Semiconductor Industry Association. International Technology Roadmap for Semiconductors, 2008.

[18] A. Shacham, Keren Bergman, and Luca Carloni. On the Design of a Photonic Network-on-chip. In *First International Symposium on Networks-on-Chip*, May 2007.

[19] L. Shiuan Peh and William Dally. A Delay Model and Speculative Architecture for Pipelined Routers. In *International Symposium on High-Performance Computer Architecture*, Jan. 2001.

[20] J. Tatum. VCSELs for 10 GB/s Optical Interconnects. *Broadband Communications for the Internet Era Symposium Digest, 2001 IEEE Symposium on Emerging Technologies*, Sept. 2001.

[21] D. Vantrease, Robert Schreiber, Matteo Monchiero, Moray McLaren, Normal Jouppi, Marco Fiorentino, A. David, Nathan Binkert, Raymond Beausoleil, and Jung Ho Ahn. Corona: System Implications of Emerging Nanophotonic Technology. In *Proc. 35th IEEE/ACM International Symposium on Computer Architecture*, June 2008.

[22] T. Woodward and A. Krishnamoorthy. 1-Gb/s Integrated Optical Detectors and Receivers in Commercial CMOS Technologies. *IEEE Journal on Selected Topics in Quantum Electronics*, 5(2):146–156, 1999.

[23] F. Xia, L. Sekaric, and Y. Vlasov. Ultracompact Optical Buffers on a Silicon Chip. *Nature Photonics*, 1:65–71, 2006.

[24] Q. Xu and M. Lipson. All-Optical Logic Based on Silicon Micro-Ring Resonators. *Optics Express*, 15(3):924–929, 2007.