Building Manycore Processor-to-DRAM Networks with Monolithic Silicon Photonics

Christopher Batten[1], Ajay Joshi[1], Jason Orcutt[1], Anatoly Khilo[1], Benjamin Moss[1]
Charles Holzwarth[1], Miloš Popović[1], Hanqing Li[1], Henry Smith[1], Judy Hoyt[1]
Franz Kärtner[1], Rajeev Ram[1], Vladimir Stojanović[1], Krste Asanović[2]

[1] Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology, Cambridge, MA

[2] Department of Electrical Engineering and Computer Science
University of California, Berkeley, CA

Abstract

We present a new monolithic silicon photonics technology suited for integration with standard bulk CMOS processes, which reduces costs and improves opto-electrical coupling compared to previous approaches. Our technology supports dense wavelength-division multiplexing with dozens of wavelengths per waveguide. Simulation and experimental results reveal an order of magnitude better energy-efficiency than electrical links in the same technology generation. Exploiting key features of our photonics technology, we have developed a processor-memory network architecture for future manycore systems based on an opto-electrical global crossbar. We illustrate the advantages of the proposed network architecture using analytical models and simulations with synthetic traffic patterns. For a power-constrained system with 256 cores connected to 16 DRAM modules using an opto-electrical crossbar, aggregate network throughput can be improved by ≈8–10× compared to an optimized purely electrical network.

1. Introduction

Modern embedded, server, graphics, and network processors already include tens to hundreds of cores on a single die and this number will surely continue to increase over the next decade. Corresponding increases in main memory bandwidth, however, are also required if the greater core count is to result in improved application performance. Projected future enhancements of existing electrical DRAM interfaces, such as XDR [14] and FB-DIMM [18], are not expected to supply sufficient bandwidth with reasonable power consumption and packaging cost. We are attempting to meet this *manycore bandwidth challenge* by combining monolithic silicon photon-ics with an optimized processor-memory network architecture.

Existing approaches to on-chip photonic interconnect have required extensive process customizations, some of which are problematic for integration with manycore processors and memories. In contrast, our approach has been to develop new photonic devices that utilize the existing material layers and structures in a standard bulk CMOS flow. Apart from preserving the massive investment in standard fabrication technology, monolithic integration also reduces the area and energy costs of interfacing electrical and optical components. Our technology focuses on supporting *dense wavelength-division multiplexing* (DWDM), packing dozens of wavelengths onto the same waveguide, to provide further improvements in area and energy efficiency. In Section 2 we describe our technology and present experimental results from photonic devices fabricated in a standard 65 nm bulk CMOS process.

We leverage DWDM to develop a new high-performance and energy-efficient processor-memory network, which would not be feasible with conventional electrical interconnect. Our architecture is based on an *opto-electrical global crossbar* implemented with a combination of on-chip/off-chip photonic interconnect for high-density, high-throughput, long-range transport, and electrical interconnect for fast routing, efficient buffering, and short-range transport. A key feature of our architecture is that photonic links are not only used for inter-chip communication, but also provide cross-chip transport to off-load global on-chip electrical wiring. Section 3 uses analytical modeling and simulation to illustrate the potential advantages of an opto-electrical global crossbar. For our target system with 256 cores and 16 independent DRAM modules we observe a ≈8–10× improvement in throughput compared with pure electrical systems under
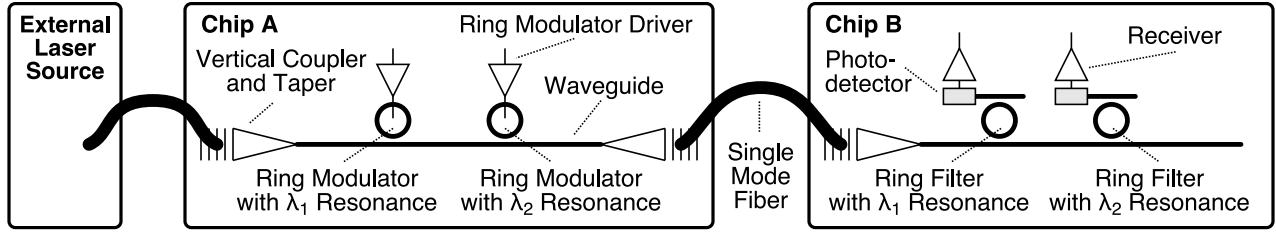
Figure 1: Photonic link with two point-to-point channels implemented with wavelength division multiplexing

(a) Waveguide with air gap

(b) Double-ring resonant filter
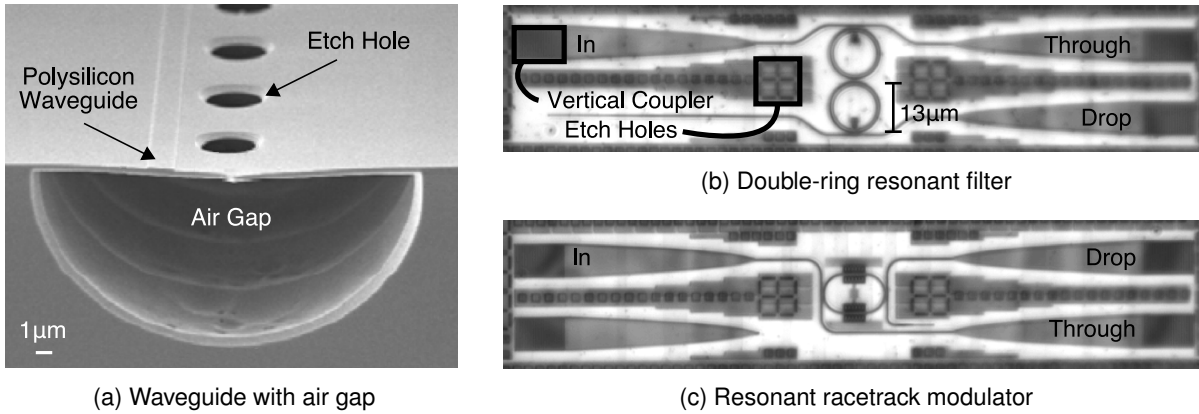
(c) Resonant racetrack modulator

Figure 2: SEM images of photonic devices: (a) cross-section of poly-Si waveguide over SiO$_2$ film with an air gap etched into the silicon substrate [6]; (b) double-ring filter, resonant wavelength is filtered to *drop* port while all other wavelengths continue to *through* port; (c) racetrack modulator, without charge injection the resonant wavelength is filtered to *drop* port while all other wavelengths continue to *through* port, with charge injection the resonant frequency changes such that no wavelengths are filtered to *drop* port.

similar energy constraints. Section 4 describes in more detail how we assemble the various photonic and electrical components to actually implement the desired network architecture while minimizing area overhead and off-chip laser power.

2. Photonic Technology

Figure 1 illustrates the components of our photonic technology using a simple WDM link. Light from an off-chip two-wavelength (λ_1,λ_2) laser source is carried by an optical fiber and arrives perpendicular to the surface of chip A, where a vertical coupler steers the light into an on-chip waveguide. The waveguide carries the light past a series of transmit drivers. Each transmitter uses a resonant ring modulator [5, 11, 13] tuned to a different wavelength to modulate the intensity (on-off keying) of the light passing by at that wavelength. Modulated light continues through the waveguide, exits chip A through a vertical coupler into another fiber, and is then coupled into a waveguide on chip B. On chip B, each of the two receivers use a tuned resonant ring filter [13, 20] to "drop" the corresponding wavelength from the waveguide into a local photodetector. The photodetector turns

absorbed light into current, which is sensed by the electrical receiver. Although not shown in Figure 1, we can simultaneously send information in the reverse direction by using another external laser source producing different wavelengths coupled into the same waveguide on chip B and received by chip A.

In the rest of this section, we briefly describe how we design each of the photonic devices to work around the limitations of a commercial sub-100 nm bulk CMOS process. We use our experiences with a 65 nm test chip [13] and our feasibility studies for a prototype 45 nm process to extrapolate photonic device parameters for our target 22 nm technology node. We also describe the electrical circuits required to interface with our photonic devices, before concluding this section with a summary of the energy efficiency of a complete optical link.

2.1. Laser

Due to the indirect Si bandgap, there are no known high-efficiency laser sources in Si, so all proposed silicon-photonic technologies use off-chip laser sources. We also use an external laser source to supply continuous wavelengths which are then modulated on-die. The laser

power does not directly enter the total power budget of the chip, but has to be on the order of few watts to keep the system cost low.

2.2. Photonic Waveguides

The waveguide is the most fundamental photonic component since all other passive structures on the chip (resonators, couplers, splitters, etc.) are made from the same material. Previously, photonic waveguides have been made either using the silicon body as a core in a silicon-on-insulator (SOI) process, with custom thick buried oxide (BOX) as cladding [5], or by depositing amorphous silicon [9] or silicon-nitride [2] on top of the interconnect stack. These approaches either require significant process changes to a standard bulk CMOS flow (or even a thin BOX SOI flow) or have high thermal isolation properties (like thick BOX SOI), which are unacceptable in manycore processors where effective thermal conduction is needed to mitigate the high-power density common in manycore computation.

To avoid process changes, we designed our photonic waveguides in the poly-Si layer on top of the shallow-trench isolation in a standard CMOS bulk process [13]. Unfortunately, the shallow-trench oxide is too thin to form an effective cladding and shield the core from optical mode leakage losses into the silicon substrate. We have developed a novel self-aligned post-processing procedure to etch away the silicon substrate underneath the waveguide forming an air gap [6]. When the air gap is more than 5 μm deep it provides a very effective optical cladding. Figure 2a shows an SEM cross-sectional image of a preliminary test chip used to experiment with various approaches to etching the air gap.

2.3. Resonant Filters

To pack a large number of wavelengths per waveguide we require resonant ring filters with high frequency selectivity. Frequency roll-off can be increased by cascading multiple rings [20]. Figure 2b shows a double-ring filter including the vertical couplers and tapers used to test the filter [13].

The stability of the filter resonance and roll-off due to process variations (line-edge roughness and lithographic precision) is a major concern. Our experimental results indicate that poly-Si height and width control is sufficient to provide stable ring frequencies within 100 GHz bands [13]. In addition to variations in ring geometry, ring resonance is also sensitive to temperature. Fortunately, the etched air gap under the ring provides thermal isolation from the thermally conductive substrate, and we add in-plane poly-Si heaters inside the ring to improve heating efficiency. Thermal simulations suggest that the
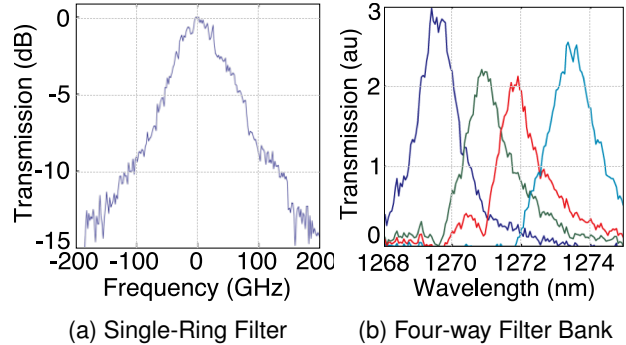

(a) Single-Ring Filter (b) Four-way Filter Bank

Figure 3: Experimental results for single-ring filters implemented in a bulk 65 nm CMOS test chip [13]

heating budget for the whole optical link will not exceed 100 fJ/b which is significantly lower than existing thermal tuning solutions [11].

We use our 65 nm bulk CMOS test chip to help estimate the number of wavelengths which can be multiplexed onto the same waveguide [13]. Figure 3a shows the measured transfer function for a single-ring filter, and Figure 3b shows the measured transfer characteristics for a four-way filter bank where each filter is tuned to a different wavelength. These results show a 200 GHz wavelength separation for single-ring filters with 2.7 THz free spectral range (FSR), indicating that at least 12 wavelengths in each direction can be multiplexed on one waveguide. By using double-ring filters with smaller radius (4 THz FSR) we can pack up to 64 wavelengths per waveguide at a tighter 60 GHz spacing. By interleaving different wavelengths traveling in opposite directions (which helps mitigate interference) we can possibly have up to 128 wavelengths per waveguide.

2.4. Modulators

The pioneering work of Soref [17] showed that the free-carrier plasma dispersion effect can change the refractive index of silicon. This effect was used to provide phase shift in branches of Mach-Zehnder (MZ) modulators [12] and more recently to change the resonance of forward-biased minority charge-injection ring modulators [11, 13]. The waveguide in a MZ or ring modulator is designed as a PIN diode, with the waveguide core acting as the undoped intrinsic region of the diode charged under a high-injection regime to realize the free carrier-plasma effect. Due to their smaller size (3–10 μm radius), ring modulators have much lower power consumption (1 pJ/b [5, 11]) compared to MZ modulators, which have lengths in millimeters and dissipate 30–60 pJ/b [12]. Lacking silicon waveguides in our process, we create PIN diodes by doping the edges of the poly-Si waveguide [13], forming a lateral diode with undoped poly-Si as the intrinsic region.

Figure 2c shows our resonant racetrack modulator. Our device simulations indicate that with poly-Si carrier lifetimes of 0.1–1 ns it is possible to achieve sub-200 fJ/b efficiency at up to 10 Gb/s speeds when advanced driver circuits are used. With a 4 μm waveguide pitch and 128 wavelengths per waveguide, this results in a data rate density of 320 Gb/s/μm, or approximately 128× the achievable data rate density of optimally repeated global on-chip electrical interconnect [8].

2.5. Photodetectors

While high-efficiency epitaxial Ge photodetectors have been demonstrated in a customized SOI process [5], the lack of pure Ge presents a challenge for mainstream bulk CMOS processes. We use the embedded SiGe (20–30% Ge), typically used for the p-MOSFET transistor source/drain regions, to create a photodetector operating at around 1200 nm. Simulation results show good capacitance (<1 fF/μm) and dark current of (<10 fA/μm) at near-zero bias conditions, but the sensitivity of the structure needs to be improved to meet our system specifications. In advanced process nodes, the responsivity and speed will improve through better coupling between the waveguide and the photodetector in scaled device dimensions, and increased percentage of Ge for device strain.

2.6. Electrical Back-end Components

Table 1 shows the estimated energy costs of the electrical back-end for the optical link (drivers, receivers, and clocking) using a predictive technology model for the 22 nm node [22]. The dominant source of energy consumption is the modulator driver, followed by optical receiver and clocking circuits. Driver circuits can be designed to tightly control the injection of charge into the modulator diode and provide low-power and high-modulation bandwidth operation. To avoid robustness and power issues from distributing a clock to hundreds of phase-locked loops on a manycore processor chip, we

Table 1: Estimated energy of photonic components

Component	Energy (fJ/b)	Cap (fF)
Serializer	1.5	6
Pre-Driver	19.0	36
Push-Pull Modulator	70.0	24
Analog Receiver Front End	40.0	
Flip-Flop Sampling & Monitoring	12.0	
Deserializer	1.5	6
Optical Clocking Source	2.0	4
Clock Phase Control	12.0	
Total	**158.0**	

propose implementing an optical clock delivery scheme similar to [4] but using a simpler, single-diode receiver with duty-cycle correction.

2.7. Energy Efficiency of Full Photonic Link

Photonic network performance is directly related to the energy efficiency of the devices used in the photonic link. Our analysis in this section suggests that the total electrical energy for our photonic link will be around 250 fJ/b (150 fJ/b signaling and 100 fJ/b heating) with an additional 300 fJ/b for external laser power. This is 1–2 orders of magnitude lower than state-of-the-art photonic devices [5, 11, 12]. Our technology achieves this energy efficiency while supporting DWDM with dozens of wavelengths per waveguide resulting in a bandwidth density of up to 320 Gb/s/μm.

Energy-efficient DWDM is enabled by (1) *monolithic integration* of photonic devices into an advanced CMOS process (smaller device parasitics, smaller capacitance of circuits driving photonic devices), (2) *innovative device design* (efficient thermal tuning through etch-undercut isolation, energy-efficient modulator, and SiGe photo-detector), and (3) *custom circuit design* (monolithic integration allows advanced modulator driver and receiver circuits, such as equalizer-controlled modulator current injection and energy-efficient, regenerative receiver structures).

3. Network Architecture

The challenge when designing a network architecture is to turn the raw link-level benefits of energy-efficient DWDM photonics into system-level performance improvements. Previous approaches have used photonics for intra-chip circuit-switched networks with very large messages [16], intra-chip bus networks for processor-to-L2 cache bank traffic [10], and general-purpose inter-chip links [15]. In this work. we focus on using photonics to implement processor-to-DRAM networks, as we believe main memory bandwidth will be a key bottleneck in future manycore systems.

Global crossbars are theoretically attractive for processor to memory networks since they have minimal network diameter, are non-blocking, and can achieve high throughput. Unfortunately, implementing an electrical global crossbar between hundreds of cores and tens of DRAM modules is impractical, due to area and energy inefficiencies. Implementing a purely photonic global crossbar is also difficult since this would require optical switching and arbitration. In this section, we argue for a hybrid opto-electrical global crossbar to exploit the advantages of each medium: photonic interconnect for compact, low-energy, and high-throughput transport, and

electrical interconnect for fast switching, efficient buffering, and local transport.

3.1. Analytical Model

Our target system for the 22 nm node includes 256 cores running at 2.5 GHz with a large number of DRAM modules. We predict this system will be power constrained as opposed to area constrained, i.e., although there will be abundant on-chip wiring resources (and to some extent off-chip I/O pins) it will not be possible to drive them all without exceeding the chip's thermal and power delivery envelope. To compare across a range of network architectures, we assume a combined power budget for the on-chip network and off-chip I/O, and individually optimize each architecture's distribution of power between on-chip and off-chip interconnect.

To help navigate the large design space, we have developed analytical models that connect component energy-models with performance metrics like *ideal throughput* and the *zero-load latency* for each of the candidate topologies. The ideal throughput is the maximum aggregate observed bandwidth that all cores can sustain under a uniform random traffic pattern with ideal flow-control and perfectly balanced routing. The zero-load latency is the average latency (including both hop latency and serialization latency) of a memory request and corresponding response under a uniform random traffic pattern with no contention in the network.

Analytical energy models for electrical and photonic implementations of on-chip interconnect and off-chip I/O were based on our insights in Section 2, previous work on optimal on-chip electrical interconnect [8], as well as gate-level analysis derived from the Orion models [19] and adapted for our 22 nm technology. We constrained our design space exploration by requiring the sum of on-chip network energy and off-chip I/O energy to not exceed 20 W (8 nJ/cycle at 2.5 GHz).

3.2. Mesh Topology

From the wide variety of possible topologies for processor-memory networks, we selected the mesh topology shown in Figure 4 for our baseline network owing to its simplicity, use in practice [7, 21], and reasonable efficiency [1]. We also examined concentrated mesh topologies with four cores per mesh router [1]. Two logical networks separate requests from responses to avoid protocol deadlock, and we implement each logical network with a separate physical network. Some of the mesh routers include an *access point* (AP) which uses off-chip I/O to connect that router to a single DRAM module. Cores send requests through the request mesh to the appropriate AP, which then forwards requests to the DRAM module.
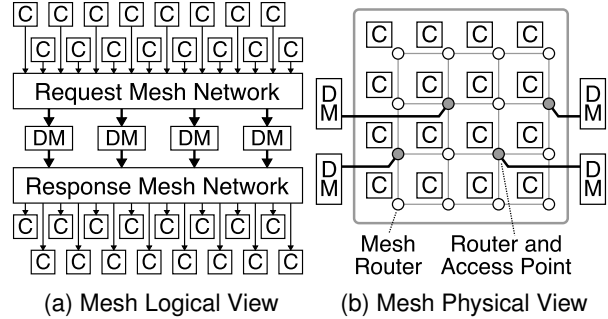

(a) Mesh Logical View (b) Mesh Physical View

Figure 4: Mesh (C = Core, DM = DRAM module)

Responses are sent back to the AP, through the response mesh, and eventually to the original core. The DRAM address space is cache-line interleaved across APs to balance the load and give good average-case performance. Our model is largely independent of whether the actual DRAM memory controller logic is located next to the AP, at the edge of the chip, or off-chip near the DRAM module.

Figure 5 shows the tension between on-chip network and off-chip I/O energy, and its impact on the theoretical system throughput and zero-load latency. For all three subfigures, the independent variable is the router-to-router channel bitwidth in the on-chip meshes. We estimate that the router-to-router channel energy is 43 fJ/b [8] and the electrical off-chip I/O energy is 5 pJ/b. Figure 5a shows that for small mesh channel bitwidths the on-chip network consumes little energy and most of the energy can be spent on off-chip I/O, while larger mesh channel bitwidths leave less energy for off-chip I/O. Figure 5b shows that for mesh channel bandwidths below 23 b/cycle, the system throughput is limited by the mesh, while beyond 23 b/cycle, the energy-starved off-chip I/O becomes the bottleneck. Finally, Figure 5c shows that for small mesh channel bitwidths, mesh serialization latency dominates total latency, while for larger bitwidths, serialization latency at the off-chip I/O interface dominates total latency.

In theory, to maximize throughput, we should choose a mesh channel bitwidth which balances the ideal throughput of the mesh with the off-chip I/O. For example, in Figure 5b throughput is maximized when the mesh channel bitwidth is 23 b/cycle. In practice, however, it is much more difficult to achieve ideal throughput in meshes than in off-chip I/O, due to multi-hop contention and load balancing issues. We can increase the *overprovisioning factor* (OPF) of the mesh network in an effort to better balance the expected achievable throughput. For example, Figure 5b shows the mesh channel bitwidth corresponding to an OPF of one, two, and four. We will investigate the impact of OPF on the achievable throughput in Section 3.4.
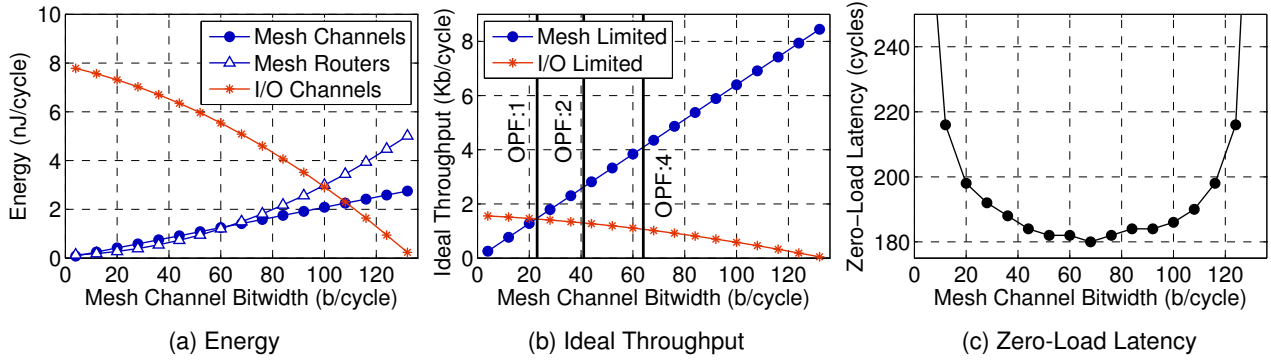
(a) Energy (b) Ideal Throughput (c) Zero-Load Latency

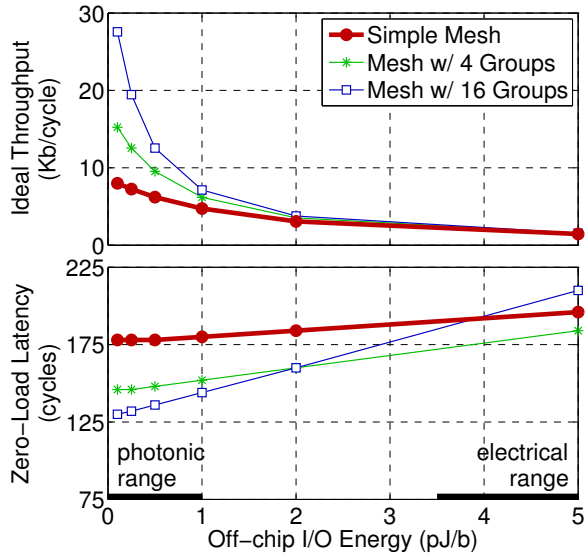**Figure 5: Various metrics vs. mesh channel bitwidth with 8 nJ/cycle constraint for on-chip mesh and off-chip I/O**

Figure 6: Ideal throughput and zero-load latency as a function of off-chip I/O energy efficiency

(a) Logical View (b) Physical View

Figure 7: Mesh augmented with a global crossbar (C_i = Core in group i, S = global crossbar switch, DM = DRAM module)

3.3. Mesh with Global Crossbar Topology

Although using photonics to implement energy-efficient off-chip I/O channels improves performance, messages still need to use the on-chip electrical network to reach the appropriate AP and this global on-chip communication is a significant bottleneck. System throughput can be further improved by moving this global traffic from energy-inefficient mesh channels onto energy-efficient global channels. To this end, we augment the electrical mesh topology with a global crossbar between groups of cores and DRAM modules.

Figure 7 illustrates an example of a global crossbar with two groups of cores. Every group of cores has an independent AP to each DRAM module so that each message need only traverse its local group sub-mesh to reach an appropriate AP. Messages then quickly move across the crossbar and arbitrate with messages from other groups at the *global crossbar switch* before actually accessing the DRAM module. Figure 7b shows the crossbar channels implemented using off-chip I/O and the global crossbar switches located off-chip near the DRAM module, which helps reduce the power density of the pro-

Figure 6 plots the ideal throughput and zero-load latency as a function of the energy efficiency of the off-chip I/O with an OPF of one. We (optimistically) project that electrical off-chip I/O in the 22 nm node will be around 5 pJ/bit while our photonic technology decreases the off-chip I/O cost to around 250 fJ/bit. Focusing on the bold simple mesh line, we can see that decreasing the off-chip I/O channel energy increases the ideal throughput with a slight reduction in the zero-load latency. This is because more energy-efficient off-chip I/O means there is more energy available for the on-chip network resulting in an overall higher system throughput. These analytical results provide some intuition that using photonic off-chip I/O with a simple on-chip mesh topology can increase throughput by $\approx 5\times$ at similar latency. However, the $20\times$ difference in energy efficiency between photonic and electrical off-chip interconnect implies that there still might be room for improvement.
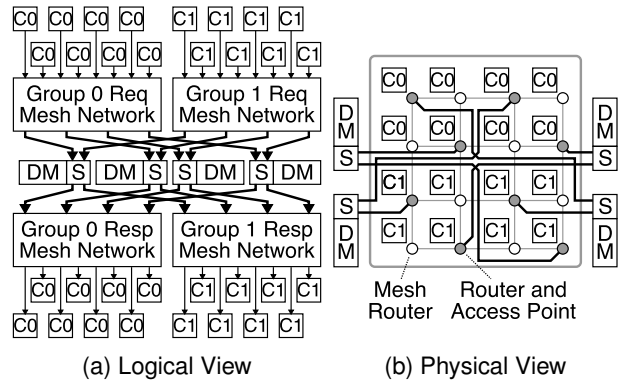
cessor chip and enables multi-socket configurations to easily share the same DRAM modules. It is important to note that this topology is not a full crossbar between cores and DRAM modules but instead connects *groups* of processors and DRAM modules. The group sub-meshes provide electrical buffering and arbitration for the APs and the switches provide electrical buffering and arbitration for the DRAM modules.

Figure 6 shows that for off-chip I/O energies in the electrical range adding a global crossbar has little impact on system throughput. Adding groups moves the mesh limited throughput curve in Figure 5b up and to the left but does not significantly change the I/O limited throughput curve, and the shallow slope of the I/O limited throughput curve limits overall performance gains. Improved off-chip I/O energy efficiency gives steeper I/O limited throughput curves and thus better exploits increased mesh throughput from grouping. Figure 6 shows that for off-chip I/O energies in the photonic range adding groups can improve throughput by $\approx$2–3$\times$ over a simple mesh with the same I/O energy cost by moving global on-chip communication onto the energy-efficient photonic links. Combining the 5$\times$ throughput increase from the raw I/O energy-efficiency of photonics and the 2–3$\times$ improvement from grouping, an opto-electrical global crossbar theoretically yields $\approx$10–15$\times$ better throughput than a simple mesh with electrical I/O.

Adding a global crossbar can reduce hop latency as well since a message needs only a few hops in the group sub-mesh before using the low-latency crossbar. Unfortunately, the energy constraint means that for some configurations (e.g. a 16 group crossbar with 5 pJ/b off-chip I/O energy) the crossbar channels become quite narrow, significantly increasing the serialization latency and the overall zero-load latency. Figure 6 shows that a global crossbar with 250 pJ/b off-chip I/O energy can reduce the zero-load latency by 30% compared to a simple mesh.

3.4. Simulation Results

The analytical results helped guide our design space exploration, but to more accurately evaluate the performance of the various topologies we used a detailed cycle-accurate micro-architectural simulator which models pipeline latencies, router contention, message fragmentation, credit-based flow control, and serialization overheads. The modeled system includes 256 cores and 16 DRAM modules with two-cycle mesh routers, one-cycle mesh channels, four-cycle global crossbar channels, and 100-cycle DRAM array access latency. All mesh networks use dimension-ordered routing and wormhole flow control [3]. We constrain all configurations to have an equal amount of network buffering, measured as total number of bits. For this work we use a synthetic

uniform random traffic pattern at a configurable injection rate. Due to the cache-line interleaving across APs, we believe this traffic pattern is representative of many bandwidth-limited applications. All request and response messages are 256 b, which is a reasonable average assuming a load/store network with 64 b addresses and 512 b cache lines. We use warmup, measure, and wait phases of several thousand cycles and an infinite source queue to accurately examine the latency at a given injection rate [3].

Table 2 shows the simulated configurations and the corresponding mesh and off-chip I/O channel bitwidths as derived from the analysis in the previous section with a total power budget of 20 W. The E configurations use a simple mesh with electrical off-chip I/O while the O configurations use photonic off-chip I/O. The first three configurations keep the OPF constant while varying the number of groups and the simulation results are shown in Figure 8a. These simulations show a significantly greater improvement in peak throughput due to grouping than predicted by the analytical model in Figure 6. Although this is partially due to realistic flow-control and routing, the primary discrepancy is that our analytical model assumes a large number of DRAM modules (APs distributed throughout the mesh) while our simulated system models a more realistic 16 DRAM modules (APs positioned in the middle of the mesh) resulting in a less uniform traffic distribution.

We can overprovision the mesh network to help these configurations better approach their theoretical peak throughput. The tradeoff is that overprovisioning increases the mesh energy resulting in less energy for the off-chip I/O and an overall lower peak throughput (see Figure 5b). The hope is that the higher achievable throughput outweights the reduction in peakt throughput. Overprovisioning is less useful as we increase the number of groups since each group submesh network becomes smaller and the number of APs per group increases. The remaining configurations in Table 2 vary the OPF as a

Table 2: Simulated configurations

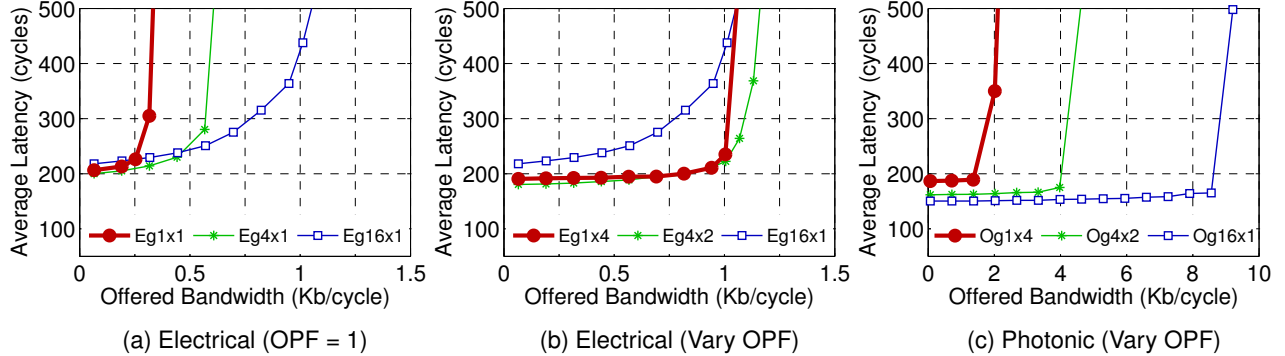| Config Name	Num Groups	OPF	Mesh Channel b/cycle	Xbar Channel b/cycle
Eg1x1	1	1	23	92
Eg4x1	4	1	12	24
Eg16x1	16	1	7	7
Eg1x4	1	4	64	64
Eg4x2	4	2	23	23
Og1x4	1	4	128	128
Og4x2	4	2	115	115
Og16x1	16	1	76	76

Figure 8: Simulation results for various topology configurations (see Table 2)

function of the number of groups. This conveniently results in setting the mesh channel bitwidth equal to the off-chip I/O bitwidth reducing implementation complexity. Figure 8b shows that varying the OPF results in all three electrical configurations achieving a throughput of 1–1.2 Kb/cycle (2.5–3.0 Tb/s). The Eg16x1 configuration's poorer performanace is due to its small flit size (many flits per message) resulting in increased congestion within the group submesh networks.

Figure 8c shows the performance of the photonic networks. Just replacing the off-chip I/O with photonics in a simple mesh topology results in a 2× improvement in throughput. However, the real benefit of photonic interconnect only becomes apparent when we augment the simple mesh with an opto-electrical global crossbar. The Og16x1 configuration can achieve a throughput of 9 Kb/cycle (22 Tb/s), which is an ≈8-10× improvement over the best electrical configuration (Eg4x2) at the same latency. The photonic configurations also provide a slight reduction in the zero-load latency.

Although results are not shown, we also investigated a concentrated mesh topology with one mesh router for every four cores [1]. Concentration decreases the total number of routers (which decreases the hop latency) at the expense of increased energy per router. Concentrated mesh configurations had similar throughput as the configurations in Figure 8b with slightly lower zero-load latencies. Concentration had little impact when combined with photonic off-chip I/O. We also investigated the effect of message fragmentation and found that it did not change the general trends of our results.

4. Full System Description

In this section we describe in more detail how we use our photonic technology and network architecture to implement a target system with 256 cores and 16 independent DRAM modules. We assume a core frequency of 2.5 GHz and a die size of 400 mm². Based on the analysis in the previous section we choose an electri-

cal mesh with a 16-group opto-electrical global crossbar. Since each group has one global crossbar channel to each DRAM module, there are a total of 256 processor-memory channels with one *photonic access point* (PAP) per channel. We use our energy-constrained analytical model and factor in various practical implementation issues to help determine an appropriate mesh bandwidth (64 b/cycle/channel) and off-chip I/O bandwidth (64 b/cycle/channel) which gives a total peak bisection bandwidth of 16 Kb/cycle (40 Tb/s).

Figure 9 shows the physical design of our target system. An external laser with optical power waveguides distributes multi-wavelength light across the chip. PAPs modulate this light to multiplex global crossbar channels onto *vertical waveguides* which connect to the *ring filter matrix* in the middle of the chip. The ring filter matrix aggregates all of the crossbar channels destined for the same DRAM module onto a small number of *horizontal waveguides*. These horizontal waveguides are then connected to the *global crossbar switch* via optical fiber. The switch converts the photonic channel back into the electrical domain for buffering and arbitration. Responses use light traveling in the opposite direction to return along the same optical path. The global crossbar uses credit-based flow control (piggybacked onto response messages) to prevent PAPs from overloading the buffering in the global crossbar switch.

Since each ring modulator operates at 10 Gb/s, we need 16 ring modulators per PAP and 16 ring filters per connection in the matrix to achieve our target of 64 b/cycle/channel. Since each waveguide can support up to 64 λ in one direction we need a total of 64 vertical waveguides and 64 horizontal waveguides. Due to the 30 mW non-linearity limit in waveguides, we need one optical power waveguide per vertical waveguide. We position waveguides together wherever possible to help amortize the overheads associated with our etched air gap technique. To ease system integration, we envision using a single optical ribbon with 64 fibers coupled to the
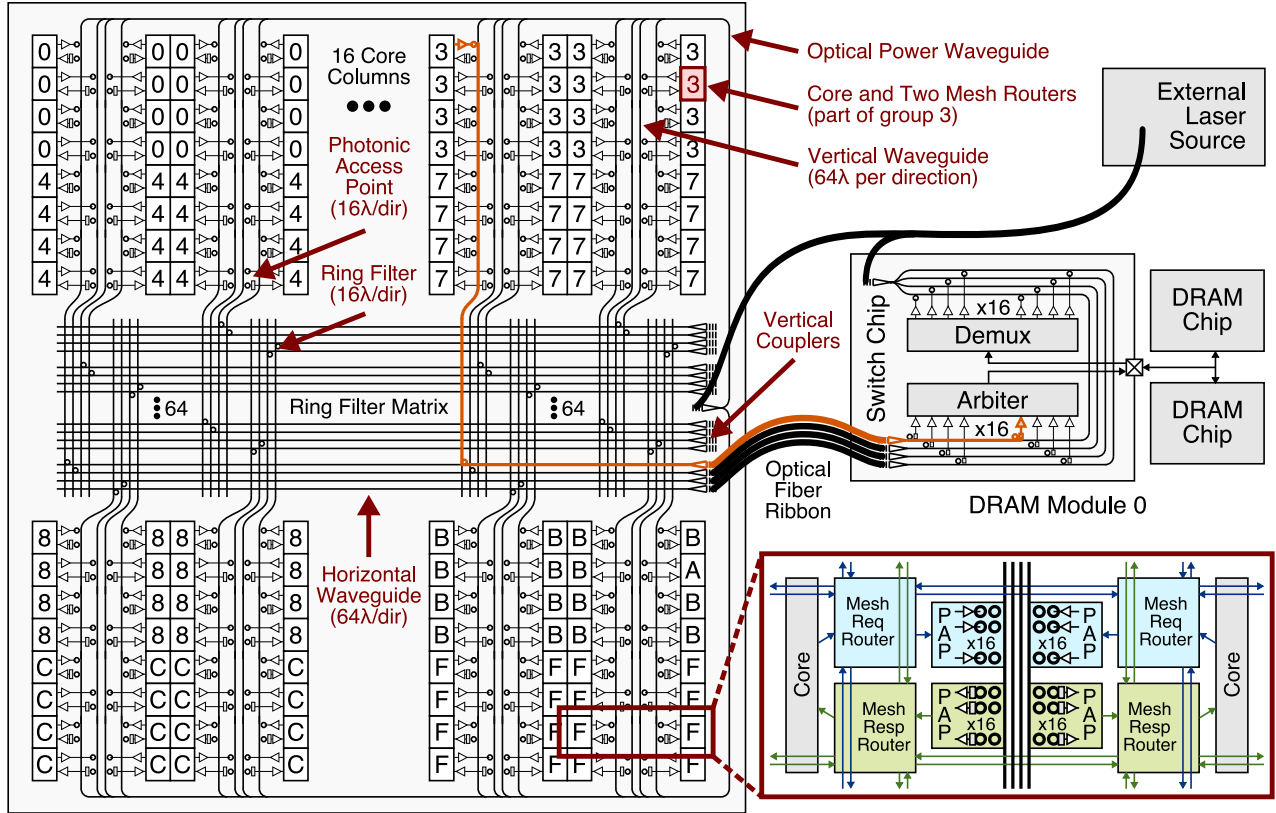
Figure 9: Target system with 256 cores, 16 DRAM modules, and 16 group opto-electrical crossbar. Each core is labeled with a hexadecimal number indicating its group. For simplicity the electrical mesh channels are only shown in the inset, each ring in the main figure actually represents 16 double rings modulating or filtering 16 different wavelengths, and each optical power waveguide actually represents 16 waveguides (one per vertical waveguide). The global crossbar request channel which connects group 3 to DRAM module 0 is shown in orange.

64 horizontal waveguides. Fibers are then stripped off in groups of four to connect to each global crossbar switch.

We now estimate the area overhead and required laser power for the opto-electrical global crossbar. Each waveguide is 0.5 μm wide on a 4 μm pitch, and each air gap requires an additional 20 μm for etch holes and alignment margins. We use two cascaded 10 μm diameter rings for all modulators and filters. Although waveguides can be routed at minimum pitch, they require additional spacing for the rings in the PAPs and ring filter matrix. The total chip area overhead for the optical power, vertical, and horizontal waveguides is between 5% and 10%. Table 10 shows an estimated optical power budget for each photonic component. A preliminary analysis of our technology suggests that our network topology is most sensitive to the losses in numerous waveguide crossings and on-chip waveguide traversal. While we can meet the waveguide crossing loss with our optimized waveguide crossing design, we are currently investigating the trade-offs between surface and bulk loss to minimize the overall waveguide loss and achieve the targets in Table 10.

Figure 10: Optical power budget

Component	Each (dB)	Total (dB)
Coupler	1	3
Splitter	0.2	1
Non-Linearity	1	1
Filter (to through node)	0.01	3.2
Modulator Insertion	0.5	0.5
Waveguide Crossing	0.05	3.2
Waveguide (per cm)	1	4
Optical Fiber (per cm)	0.5e-5	0
Filter (to drop node)	1.5	3
Photodetector	0.1	0.1
Receiver Sensitivity		-20 dBm
Power per Wavelength		-1 dBm
Total Laser Power		6.5 W

5. Conclusion

Although processor chips are projected to integrate hundreds of cores in the near future, memory bandwidth predictions are much bleaker. In this paper, we introduced both a new photonic technology and an application of this technology to meet the *manycore bandwidth challenge*. Our photonic technology will enable monolithic integration into mainstream sub-100 nm CMOS process flows. Based on simulations and experimental results from our 65 nm test chip, we estimate that we can achieve energy-efficient *dense wavelength-division multiplexing* DWDM provides bandwidth densities on the order of 320 Gb/s/μm at only 250 fJ/bit resulting in an order of magnitude improvement over optimized electrical interconnect. We leverage this photonic technology to implement an *opto-electrical global crossbar* between small groups of cores and DRAM modules. Simulation results of our target system with 256 cores and 16 DRAM modules show a ≈8–10× improvement in network throughput compared to an optimistic baseline electrical system under similar energy constraints.

Acknowledgments

References

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

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

[3] W. Dally and B. Towles. *Principles and Practices of Interconnection Networks*. Morgan Kaufmann, 2004.

[4] C. Debaes et al. Receiver-less optical clock injection for clock distribution networks. *Journal of Selected Topics in Quantum Electronics*, 9(2):400–409, Mar-Apr 2003.

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

[6] C. Holzwarth et al. Localized substrate removal technique enabling strong-confinement microphotonics in bulk Si CMOS processes. *Conf. on Lasers and Electro-Optics*, 2008.

[7] Y. Hoskote et al. A 5-GHz mesh interconnect for a teraflops processor. *IEEE Micro*, 27(5):51–61, Sep-Oct 2007.

[8] B. Kim and V. Stojanovic. Equalized interconnects for on-chip networks: Modeling and optimization framework. *Int'l Conf. on Computer Aided Design*, Nov 2007.

[9] L. Kimerling et al. Electronic-photonic integrated circuits on the CMOS platform. *Proceedings of the SPIE*, 6125, Mar 2006.

[10] N. Kirman et al. Leveraging optical technology in future bus-based chip multiprocessors. *Int'l Symp. on Microarchitecture*, pages 492–503, Dec 2006.

[11] M. Lipson. Compact electro-optic modulators on a silicon chip. *Journal of Selected Topics in Quantum Electronics*, 12(6):1520–1526, Nov-Dec 2006.

[12] A. Narasimha et al. A fully integrated 4×10Gb/s DWDM optoelectronic transceiver in a standard 0.13 μm CMOS SOI. *Journal of Solid State Circuits*, 42(12):2736–2744, Dec 2007.

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

[14] D. Pham et al. Overview of the architecture, circuit design, and physical implementation of a first-generation Cell processor. *Journal of Solid State Circuits*, 41(1):179–196, 2006.

[15] C. Schow et al. A <5mW/Gb/s/link, 16×10Gb/s bi-directional single-chip CMOS optical transceiver for board level optical interconnects. *Int'l Solid-State Circuits Conf.*, pages 294–295, Feb 2008.

[16] A. Shacham and other. Photonic NoC for DMA communications in chip multiprocessors. *Symp. on High-Performance Interconnects*, pages 29–36, Sep 2007.

[17] R. Soref. Silicon-based optoelectronics. *Proceedings of the IEEE*, 81:1687–1706, 1993.

[18] P. Vogt. Fully buffered DIMM (FB-DIMM) server memory architecture: Capacity, performance, reliability, and longevity. *Intel Developer Forum*, Feb 2004.

[19] H. Wang, L. Peh, and S. Malik. Power-driven design of router microarchitectures in on-chip networks. *Int'l Symp. on Microarchitecture*, pages 105–116, Dec 2003.

[20] M. Watts et al. Design, fabrication, and characterization of a free spectral range doubled ring-resonator filter. *Conf. on Lasers and Electro-Optics*, 1:269–272, May 2005.

[21] D. Wentzlaff et al. On-chip interconnection architecture of the Tile processor. *IEEE Micro*, 27(5):15–21, Sep-Oct 2007.

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