Building Many-Core Processor-to-DRAM Networks with Monolithic CMOS Silicon Photonics

Christopher Batten
Ajay Joshi
Jason Orcutt
Anatol Khilo
Benjamin Moss
Charles W. Holzwarth
Miloš A. Popović
Hanqing Li
Henry I. Smith
Judy L. Hoyt
Franz X. Kärtner
Rajeev J. Ram
Vladimir Stojanović
Massachusetts Institute of Technology

Krste Asanović
University of California, Berkeley

SILICON PHOTONICS IS A PROMISING TECHNOLOGY FOR ADDRESSING MEMORY BANDWIDTH LIMITATIONS IN FUTURE MANY-CORE PROCESSORS. THIS ARTICLE FIRST INTRODUCES A NEW MONOLITHIC SILICON-PHOTONIC TECHNOLOGY, WHICH USES A STANDARD BULK CMOS PROCESS TO REDUCE COSTS AND IMPROVE ENERGY EFFICIENCY, AND THEN EXPLORES THE LOGICAL AND PHYSICAL IMPLICATIONS OF LEVERAGING THIS TECHNOLOGY IN PROCESSOR-TO-MEMORY NETWORKS.

●●●●●● Modern embedded, server, graphics, and network processors already include tens to hundreds of cores on a single die, and this number will continue to increase over the next decade. Corresponding increases in main memory bandwidth are also required, however, if the greater core count is to result in improved application performance. Projected enhancements of existing electrical DRAM interfaces are not expected to supply sufficient bandwidth with reasonable power consumption and packaging cost. To meet this many-core memory bandwidth challenge, we are combining monolithic CMOS silicon photonics 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 many-core processors and memories.[1,2] In contrast, we are developing new photonic devices that use the existing material layers and structures in a standard bulk CMOS flow. In addition to preserving the massive investment in standard fabrication technology, monolithic integration reduces the area and energy costs of interfacing electrical and optical components. Our technology supports dense wavelength-division multiplexing (DWDM) with dozens of wavelengths packed onto the same waveguide to further improve area and energy efficiency.

The challenge when designing a photonic chip-level network is to turn the raw link-level benefits of energy-efficient DWDM photonics into system-level performance improvements. Previous approaches have used photonics for intrachip circuit-switched networks with very large messages,[3] intrachip

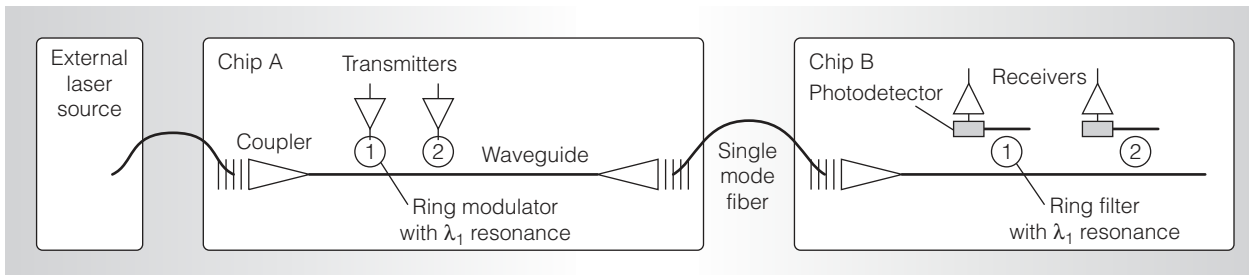Published by the IEEE Computer Society   0272-1732/09/$26.00 © 2009 IEEE

Figure 1. Two point-to-point photonic links implemented with wavelength division multiplexing.

crossbar networks for processor-to-L2 cache bank traffic,[4,5] and general-purpose interchip links.[6] Since main-memory bandwidth will be a key bottleneck in future many-core systems, this work considers leveraging photonics in processor-to-DRAM networks. We propose using a local meshes to global switches (LMGS) topology that connects small meshes of cores on-chip to global switches located off-chip near the DRAM modules. Our optoelectrical approach implements both the local meshes and global switches electrically and uses seamless on-chip/off-chip photonic links to implement the global point-to-point channels connecting every group to every DRAM module. A key feature of our architecture is that the photonic links are not only used for interchip communication, but also to provide cross-chip transport to off-load intrachip global electrical wiring.

A given logical topology can have many different physical implementations, each with different electrical, thermal, and optical power characteristics. In this work, we describe a new ring-filter matrix template as a way to efficiently implement our optoelectrical networks. We explore how the quality of different photonic devices impacts the area overhead and optical power of this template. As an example of our vertically integrated approach, we identified waveguide crossings as a critical component in the ring-filter matrix template, and this observation served as motivation for the photonic device researchers to investigate optimized waveguide crossing structures.

We have applied our approach to a target system with 256 cores and 16 independent DRAM modules. Our simulation results show that silicon photonics can improve throughput by almost an order of magnitude compared to pure electrical systems under similar power constraints. Our work suggests that the LMGS topology and corresponding ring-filter matrix layout are promising approaches for turning the link-level advantages of photonics into system-level benefits.

Photonic technology

Although researchers have proposed many types of devices for chip-scale optical networks, the most promising approach uses an external laser source and small energy-efficient ring resonators for modulation and filtering. Figure 1 illustrates the use of such devices to implement a simple wavelength-division multiplexed link. An optical fiber carries light from an off-chip laser source to chip A, where it is coupled into an on-chip waveguide. The waveguide routes the light past a series of transmitters. Each transmitter uses a resonant ring modulator tuned to a different wavelength to modulate the intensity of the light passing by at that wavelength. Modulated light continues through the waveguide, exits chip A into another fiber, and is then coupled into a waveguide on chip B. This waveguide routes the light by two receivers that use a tuned resonant ring filter to "drop" the corresponding wavelength from the waveguide into a local photodetector. The photodetector turns absorbed light into current, which is amplified by the electrical portion of the 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.

We have developed a novel approach that implements these devices in a commercial

....
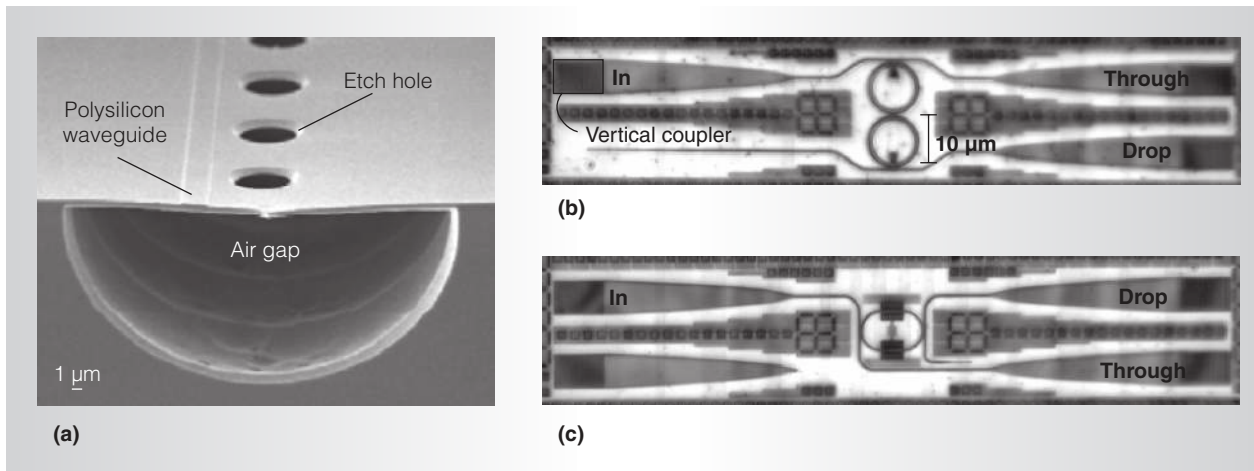
HOT INTERCONNECTS

Figure 2. Photonic devices implemented in a standard bulk CMOS process. Waveguides are implemented in poly-Si with an etched air gap to provide optical cladding (a).[7] Cascaded rings are used to filter the resonant wavelength to the ''drop'' port while all other wavelengths continue to the ''through'' port (b). Ring modulators use charge injection to modulate a single wavelength: without charge injection the resonant wavelength is filtered to the ''drop'' port while all other wavelengths continue to the ''through'' port; with charge injection, the resonant frequency changes such that no wavelengths are filtered to the ''drop'' port (c).

sub-100-nm bulk CMOS process.[7,8] This allows photonic waveguides, ring filters, transmitters, and receivers to be monolithically integrated with hundreds of cores on the same die, which reduces cost and increases energy efficiency. We use our experiences with a 65-nm test chip and our feasibility studies for a prototype 32-nm process to extrapolate photonic device parameters for our target 22-nm technology node.

Previously, researchers implemented photonic waveguides using the silicon body as a core in a silicon-on-insulator (SOI) process with custom thick buried oxide (BOX) as cladding,[2] or by depositing additional material layers (such as silicon nitride) on top of the interconnect stack.[1] To avoid process changes, we designed our photonic waveguides in the polysilicon (poly-Si) layer on top of the shallow trench isolation in a standard CMOS bulk process (see Figure 2a). Unfortunately, the shallow-trench oxide is too thin to form an effective cladding and to shield the core from optical mode leakage losses into the silicon substrate. Hence, we developed a novel self-aligned postprocessing procedure to etch away the silicon substrate underneath the waveguide forming an air gap.[7] When the air gap is more than 5 μm deep, it provides an effective optical cladding.

For this work, we assume up to eight waveguides can use the same air gap with a 4-μm waveguide pitch.

We use poly-Si resonant ring filters for modulating and filtering different wavelengths (see Figure 2b). The ring radius determines the resonant frequency, and we cascade rings to increase the filter's selectivity. The ring's resonance is also sensitive to temperature and requires some form of active thermal tuning. Fortunately, the etched air gap under the ring provides isolation from the thermally conductive substrate, and we add in-plane poly-Si heaters inside most rings to improve heating efficiency. Thermal simulations suggest that we will require 40 to 100 μW of static power for each double-ring filter assuming a temperature range of 20 K. We estimate that we can pack up to 64 wavelengths per waveguide at a 60-GHz spacing and that interleaving wavelengths traveling in opposite directions (which helps mitigate interference) could provide up to 128 wavelengths per waveguide.

Our photonic transmitters are similar to past approaches that use minority charge injection to change the resonant frequency of ring modulators.[9] Our racetrack modulator is implemented by doping the edges of a poly-Si ring, creating a lateral PiN diode

with undoped poly-Si as the intrinsic region (see Figure 2c). Due to their smaller size (3 to 10 µm radius), ring modulators can have lower power consumption than other approaches (such as Mach-Zehnder modulators). Our device simulations indicate that with poly-Si carrier lifetimes of 0.1 to 1 ns, it is possible to achieve sub-100 fJ per bit (fJ/b) for random data at up to 10 gigabits per second (Gbps) speeds when using advanced driver circuits. To avoid robustness and power issues from distributing a clock to hundreds of transmitters, we propose implementing an optical clock delivery scheme using a simple single-diode receiver with duty-cycle correction. With a 4-µm waveguide pitch and 64 to 128 wavelengths per waveguide, we can achieve a data rate density of 160 to 320 Gbps/µm, which is approximately two orders of magnitude greater than the data rate density of optimally repeated global on-chip electrical interconnect.[10]

Photonic receivers often use high-efficiency epitaxial Germanium (Ge) photodetectors,[2] but the lack of pure Ge presents a challenge for mainstream bulk CMOS processes. We use the embedded SiGe (20 to 30 percent Ge) in the p-MOSFET transistor source and drain regions to create a photodetector operating at approximately 1,200 nm. Simulation results show good capacitance (less than 1 fF/µm) and dark current (less than 10 fA/µm) at near-zero bias conditions, but the structure's sensitivity must be improved to meet our system specifications. In advanced process nodes, the responsivity and speed should improve through better coupling between the waveguide and the photodetector in scaled device dimensions, and an increased percentage of Ge for device strain. Our photonic receiver circuits would use the same optical clocking scheme as our transmitters, and we estimate that the entire receiver will consume less than 50 fJ/b for random data.

Based on our device simulations and experiments, we estimate the total electrical and thermal on-chip energy for a complete 10-Gbps photonic link (including a doubling-ring modulator and filter at the receiver) to be 100 to 250 fJ/b for random data. In addition to the on-chip electrical power, the external laser's electrical power consumption must also remain in a reasonable range. The light generated by the laser experiences optical losses in each photonic device, which reduces the amount of optical power reaching the photodetector. Different network topologies and their corresponding physical layout result in different optical losses and thus require varying amounts of optical laser power. With current laser efficiencies, generating optical laser power requires three to four times greater electrical laser power. In addition to the photonic device losses, there is also a limit to the total amount of optical power that can be transmitted through a waveguide without large nonlinear losses. In this work, we assume a maximum of 50 mW per waveguide at 1 dB loss. Network architectures with high optical losses per wavelength will need to distribute those wavelengths across many waveguides (increasing the overall area) to stay within this nonlinearity limit.

Many-core processor-to-DRAM network topologies

Monolithic silicon photonics is a promising technology for addressing the many-core memory bandwidth challenge. We present a hybrid optoelectrical approach that targets the advantages of each medium: photonic interconnect for energy-efficient global communication and electrical interconnect for fast switching, efficient buffering, and local communication.

Our target system for this work is a 256-core processor running at 2.5 GHz with tens of DRAM modules. Although such a system will be feasible on a 400-mm^2 die in the 22-nm node, it will likely be power constrained as opposed to area constrained. The system will have abundant on-chip wiring resources and, to some extent, off-chip I/O pins, but 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 we individually optimize each architecture's distribution of power between these two components.

To help navigate the large design space, we developed analytical models that connect component energy models with performance
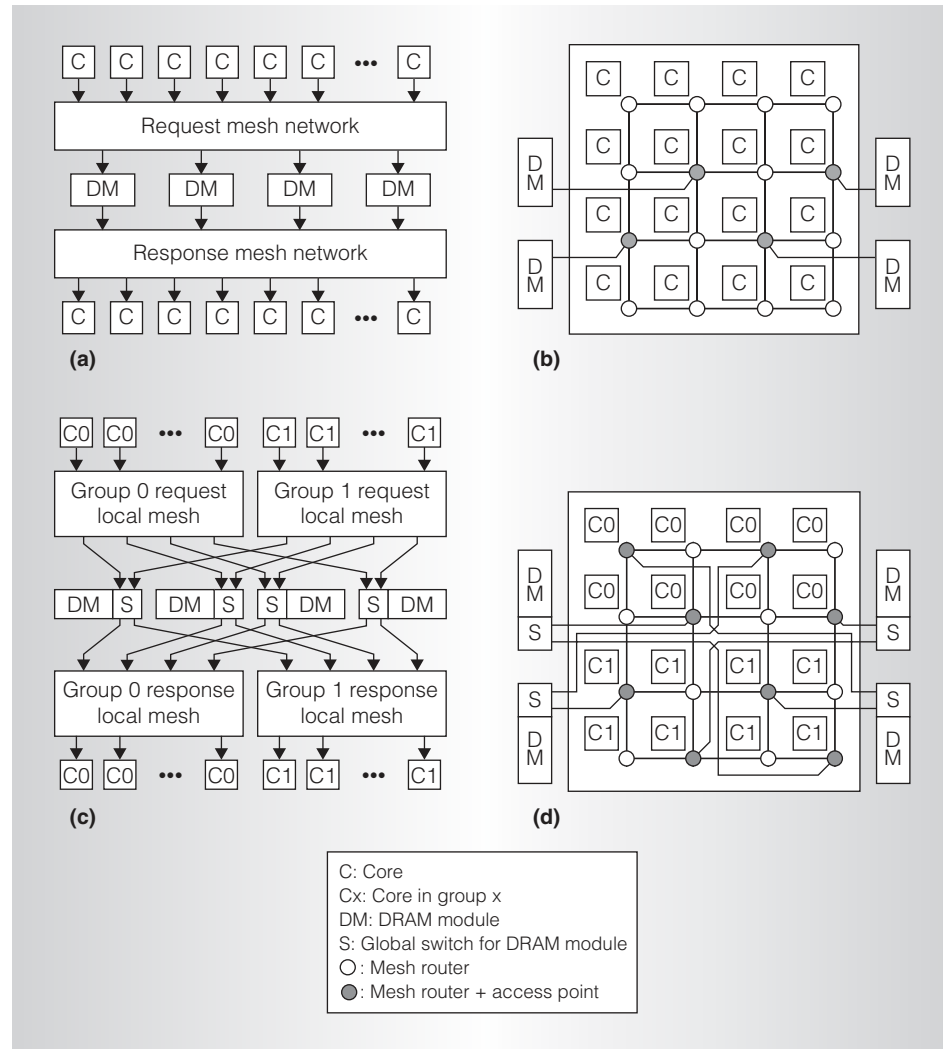
Figure 3. Logical and physical views of mesh and local meshes to global switches (LMGS) topologies: mesh logical view (a), mesh physical view (b), LMGS logical view (c), and LMGS physical view (d).

metrics such as ideal throughput and zero-load latency. The ideal throughput is the maximum aggregate 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 are based on our insights

in the last section, previous work on optimal on-chip electrical interconnect,[10] and a circuit-level analysis for our 22-nm technology.

Mesh topology

From the wide variety of possible topologies for processor-memory networks, we selected the mesh topology in Figures 3a and 3b for our baseline network because of its simplicity, use in practice, and reasonable efficiency. We also examined concentrated mesh topologies with four cores per mesh router.[11] 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, which is the interface between the on-chip network and the channel that connects to a DRAM module. Cores send requests through the request mesh to the appropriate access point, which then forwards requests to the DRAM module. Responses are sent back to the access point, through the response mesh, and eventually to the original core. The DRAM address space is cache-line interleaved across access points to balance the load and give good average-case performance. Our model is largely independent of whether the DRAM memory controller is located next to the access point, at the edge of the chip, or off-chip near the DRAM module.

Figure 4 shows what fraction of the total network power is consumed in the on-chip mesh network as a function of the total network's ideal throughput. To derive this plot, we first choose a bitwidth for the channel between routers in the mesh, then we determine the mesh's ideal throughput. Finally, we assume that the off-chip I/O must have an equal ideal throughput as the on-chip mesh to balance the on-chip and off-chip bandwidths. We use our analytical models to determine the power required by the on-chip mesh and off-chip I/O under uniform random traffic with random data. We assume that an electrical off-chip I/O link in the 22-nm node will require approximately 5 pJ/b at 10 Gbps, while our photonic technology can decrease this to 250 fJ/b. For comparison, our analytical models predict that the mesh router-to-router link energy will be approximately 50 fJ/b. Figure 4 also shows configurations corresponding to 10-, 20-, and 30-W power constraints on the sum of the on-chip network power and off-chip I/O power.

Focusing on the simple mesh line in Figure 4a, we can see that with electrical off-chip I/O approximately 25 percent of the total power is consumed in the on-chip mesh network. The ideal throughput under a 20-W power constraint is approximately 1 kilobit per cycle (Kbits/cycle). Energy-efficient photonic off-chip I/O enables increased off-chip bandwidth, but photonics
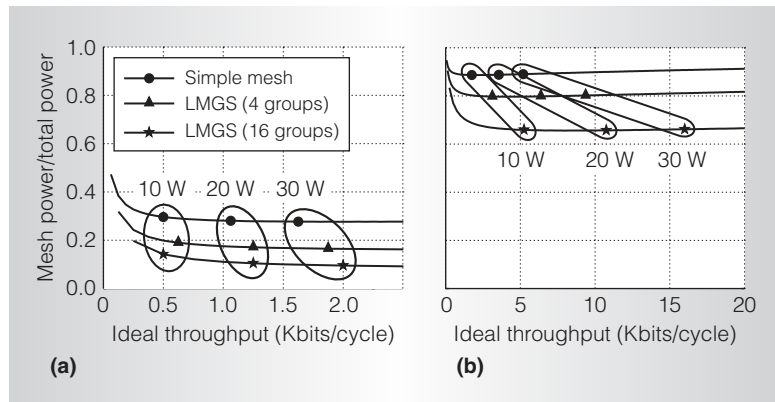

Figure 4. Fraction of total network power consumed in mesh versus ideal throughput: electrical assuming 5 pJ/b (a) and photonic assuming 250 fJ/b (b). Markers show configurations corresponding to 10-, 20-, and 30-W power constraints.

also leaves more energy to improve the on-chip electrical network's throughput. Figure 4b shows that photonics can theoretically increase the ideal throughput under a 20-W power constraint by a factor of 3.5 to about 3.5 Kbits/cycle. With a simple mesh and photonic off-chip I/O, almost all the power is consumed in the on-chip network.

For all the configurations we discuss here, we assume a constant amount of on-chip network buffering as measured by the total number of bits. For example, configurations with wider physical channels have fewer entries per queue. Figure 4 shows that for very small throughputs the power overhead due to a constant amount of buffering starts to outweigh the power savings from narrower mesh channels, so the mesh power starts to consume a larger fraction of the total power.

Figure 5 plots the ideal throughput and zero-load latency as a function of the energy efficiency of the off-chip I/O under a 20-W power constraint. Focusing on the simple mesh line, we can see that decreasing the off-chip I/O link energy increases the ideal throughput with a slight reduction in the zero-load latency. 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 access point, and this on-chip global communication is a significant bottleneck.
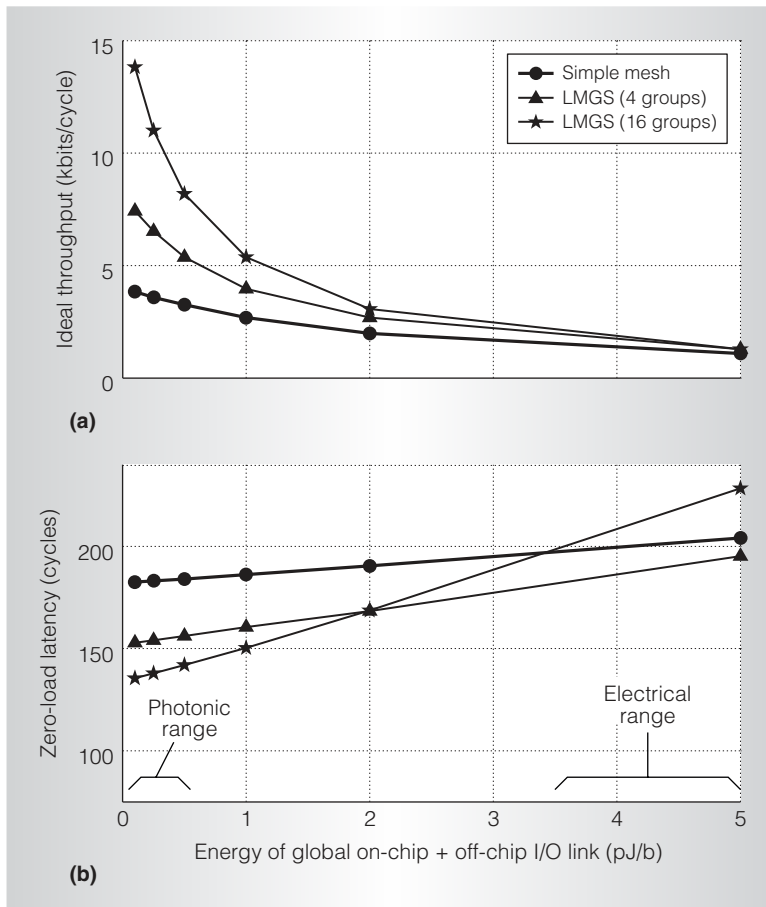
Figure 5. Ideal throughput (a) and zero-load latency (b) under 20-W power constraint.

accessing the DRAM module. As Figure 3d shows, each global point-to-point channel uses a combination of on-chip global links and off-chip I/O links. The global switches are located off-chip near the DRAM module, which helps reduce the processor chip's power density and enables multi-socket configurations to easily share the same DRAM modules.

Figures 4a and 4b show the theoretical performance of the LMGS topology compared to a simple mesh. For both electrical and photonic off-chip I/O, LMGS topologies reduce the fraction of the total power consumed in the on-chip mesh since global traffic is effectively being moved from the mesh network onto the on-chip global and off-chip I/O channels. However, with electrical technology, most of the power is already spent in the off-chip I/O so grouping doesn't significantly improve the ideal throughput. With photonic technology, most of the power is consumed in the on-chip mesh network, so offloading global traffic onto energy-efficient photonic channels can significantly improve performance. This assumes that we use photonics for both the off-chip I/O and on-chip global channels so that we can create seamless on-chip/off-chip photonic channels from each local mesh to each global switch. Essentially, we're exploiting the fact that once we pay to transmit a bit between chips optically, it doesn't cost any extra transceiver energy (although it might increase optical laser power) to create such a seamless on-chip/off-chip link. Under a 20-W power constraint, the ideal throughput improves by a factor of 2.5 to 3 compared to a simple mesh with photonic off-chip I/O. This ultimately suggests almost an order of magnitude improvement compared to using electrical off-chip I/O.

Figure 5b shows that the LMGS topology can also reduce hop latency since a message needs only a few hops in the group submesh before using the low-latency global point-to-point channels. Unfortunately, the power constraint means that for some configurations (such as 16 groups with electrical off-chip I/O), the global channels become narrow, significantly increasing the serialization latency and the overall zero-load latency.

LMGS topology

We can further improve system throughput by moving this global traffic from energy-inefficient electrical mesh channels onto energy-efficient optical channels. Figures 3c and 3d illustrate a LMGS topology that partitions the mesh into smaller groups of cores and then connects these groups to main memory with switches located off-chip near the DRAM modules. Figures 3c and 3d show 16 cores and two groups. Every group of cores has an independent access point to each DRAM module so each message need only traverse its local group submesh to reach an appropriate access point. Messages then quickly move across the global point-to-point channels and arbitrate with messages from other groups at a DRAM module switch before actually
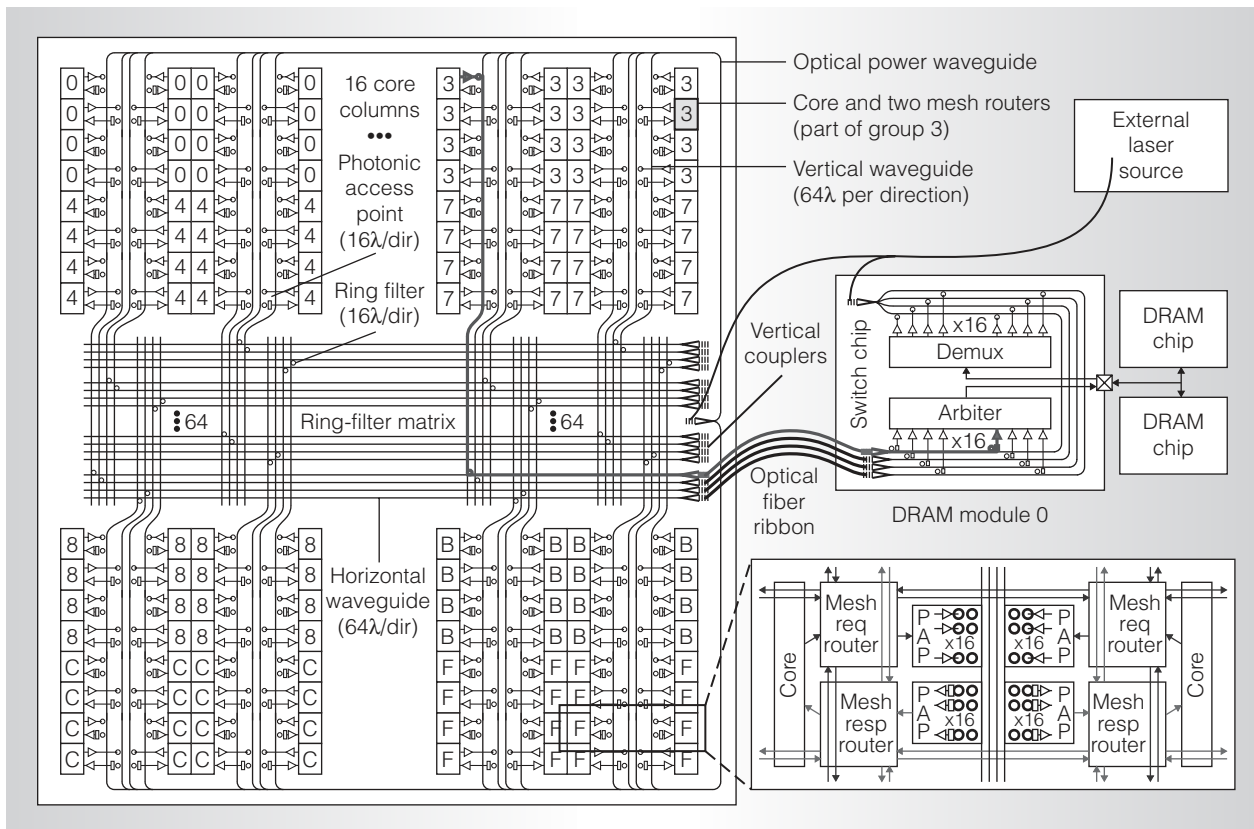
Figure 6. Ring-filter matrix implementation of LMGS topology with 256 cores, 16 groups, and 16 DRAM modules. Each core is labeled with a hex number indicating its group. For simplicity, the electrical mesh channels are only shown in the inset, each ring in the main figure represents 16 double rings, and each optical power waveguide represents 16 waveguides (one per vertical waveguide). The global request channel that connects group 3 to DRAM module 0 is highlighted.

Photonic ring-filter matrix implementation

We have developed a new approach based on a ring-filter matrix for implementing the mesh and LMGS topologies. Figure 6 illustrates the proposed layout for a 16-group, 256-core system running at 2.5 GHz with 16 independent DRAM modules. We assume a 400-mm^2 die implemented in a 22-nm technology. Since each group has one global channel to each DRAM module, there are a total of 256 processor-memory channels with one photonic access point (PAP) per channel. An external laser coupled to on-chip optical power waveguides distributes multiwavelength light to the PAPs located across the chip. PAPs modulate this light to multiplex the global point-to-point channels onto vertical waveguides that connect to the ring-filter matrix in the middle of the chip. The ring-filter matrix aggregates all the channels destined for the same DRAM module onto a small number of horizontal waveguides. These horizontal waveguides are then connected to the DRAM module switch chip via optical fiber. The switch chip converts data on 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 channels use credit-based flow control (piggybacked onto response messages) to prevent PAPs from overloading the buffering in the DRAM module switches.

For the example in Figure 6, we use our analytical model with a 20-W power constraint to help determine an appropriate mesh bandwidth (64 bits/cycle/channel) and off-chip I/O bandwidth (64 bits/cycle/channel), which gives a total peak bisection bandwidth

....
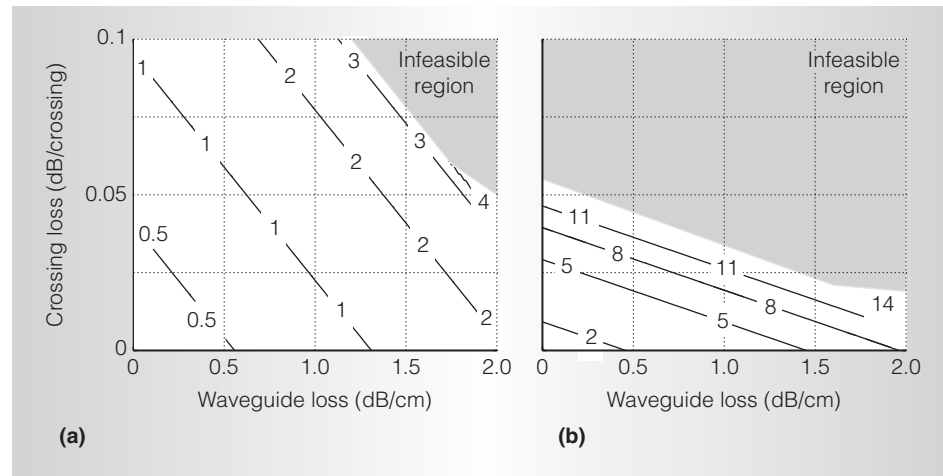
HOT INTERCONNECTS

Figure 7. Optical laser power in watts for 32 bits/cycle global I/O channels (a) and 128 bits/cycle global I/O channels (b), given the following optical loss parameters: 1 dB coupler loss, 0.2 dB splitter loss, 1 dB nonlinearity at 50 mW, 0.01 dB through loss, 1.5 dB drop loss, 0.5 dB modulator insertion loss, 0.1 dB photodetector loss, and −20 dBm receiver sensitivity.

of 16 Kbits/cycle or 40 terabits per second (Tbps) in each direction. Since each ring modulator operates at 10 Gbps, we need 16 ring modulators per PAP and 16 ring filters per connection in the matrix to achieve our target 64 bits/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 50-mW nonlinearity waveguide limit, we need one optical power waveguide per vertical waveguide. We aggregate waveguides 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 64 horizontal waveguides. Fibers are then stripped off in groups of four to connect to each DRAM module switch.

The proposed ring-filter matrix template can be used for different numbers of groups, cores, DRAM modules, and target system bandwidths by simply varying the number of horizontal and vertical waveguides. These different systems will have different optical power and area overheads. Figure 7 shows the optical laser power as a function of waveguide loss and waveguide crossing loss for 16-group networks with both less aggregate bandwidth (32 bits/cycle global I/O channels) and more aggregate bandwidth (128 bits/cycle global I/O channels) than

the system pictured in Figure 6. Higher-quality devices always result in lower total optical power. Systems with higher ideal throughput (see Figure 7b) have quadratically more waveguide crossings, making them more sensitive to crossing losses. Additionally, certain combinations of waveguide and crossing losses result in large cumulative losses and require multiple waveguides to stay within the nonlinearity limit. These additional waveguides further increase the total number of crossings, which in turn continues to increase the power per wavelength, meaning that for some device parameters it is infeasible to leverage the ring-filter matrix template. This type of analysis can be used to drive photonic device research, and we have developed optimized waveguide crossings that can potentially reduce the crossing loss to 0.05 dB per crossing.[12]

We also studied the area overhead of the ring-filter matrix template for a range of waveguide and crossing losses. We assumed 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. Our study found that the total chip area

overhead for the photonic components in the system shown in Figure 6 ranges from 5 to 10 percent depending on the quality of the photonic components. From these results, we can see that although this template provides a compact and well-structured layout, it includes numerous waveguide crossings that must be carefully designed to limit total optical laser power.

Simulation results

To more accurately evaluate the performance of the various topologies, we used a detailed cycle-level microarchitectural simulator that models pipeline latencies, router contention, credit-based flow control, and serialization overheads. The modeled system includes 256 cores and 16 DRAM modules in a 22-nm technology with two-cycle mesh routers, one-cycle mesh channels, four-cycle global point-to-point channels, and 100-cycle DRAM array access latency. All mesh networks use dimension-ordered routing and wormhole flow control. We constrain all configurations to have an equal amount of network buffering, measured in 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 access points, we believe this traffic pattern is representative of many bandwidth-limited applications. All request and response messages are 256 bits, which is a reasonable average assuming a load/store network with 64-bit addresses and 512-bit cache lines. We assume that the flow-control digit (flit) size is equal to the physical channel bitwidth. We use warm-up, measure, and wait phases of several thousand cycles each and an infinite source queue to accurately determine the latency at a given injection rate. We augment our simulator to count various events (such as channel utilization, queue accesses, and arbitration), which we then multiply by energy values derived from our analytical models. For our energy calculations, we assume that all flits contain random data.

Table 1 shows the simulated configurations and the corresponding mesh and off-chip I/O channel bitwidths as derived from the analysis presented earlier in this article with a total power budget of 20 W. We also considered

Name*	Mesh channel width (bits per cycle)	Global I/O channel width (bits per cycle)
Eg1x1	16	64
Eg4x1	8	16
Eg16x1	8	8
Eg1x4	64	64
Eg4x2	24	24
OCg1x1	64	256
OCg4x1	48	96
OCg16x1	48	48
OCg1x4	128	128
OCg4x2	80	80
OAg1x1	64	256
OAg4x1	64	128
OAg16x1	64	64
OAg1x4	128	128
OAg4x2	96	96

Table 1. Simulated configurations.

* The name of each configuration indicates the technology we used to implement the off-chip I/O (E = electrical, OC = conservative 250 fJ/b photonic links, OA = aggressive 100 fJ/b photonic links), the number of groups (g1/g4/g16 = 1/4/16 groups), and the OPF (x1/x2/x4 = OPF of 1/2/4).

various practical issues when rounding each channel bit width to an appropriate multiple of eight. In theory, all configurations should balance the mesh's throughput with the throughput of the off-chip I/O so that neither part of the system becomes a bottleneck. In practice, however, it can be difficult to achieve the ideal throughput in mesh topologies due to multihop contention and load-balancing issues. Therefore, we also consider configurations that increase the mesh network's overprovisioning factor (OPF) in an effort to improve the expected achievable throughput. The OPF is the ratio of the on-chip mesh ideal throughput to the off-chip I/O ideal throughput.

The Eg1x1, Eg4x1, and Eg16x1 configurations keep the OPF constant while varying the number of groups; Figure 8a shows the simulation results. The peak throughput for Eg1x1 and Eg4x1 are significantly less than predicted by the analytical model in Figure 4a. This is due to realistic flow-control and routing and the fact that our analytical model assumes a large number of DRAM modules (access points distributed throughout the mesh) while our simulated system

....
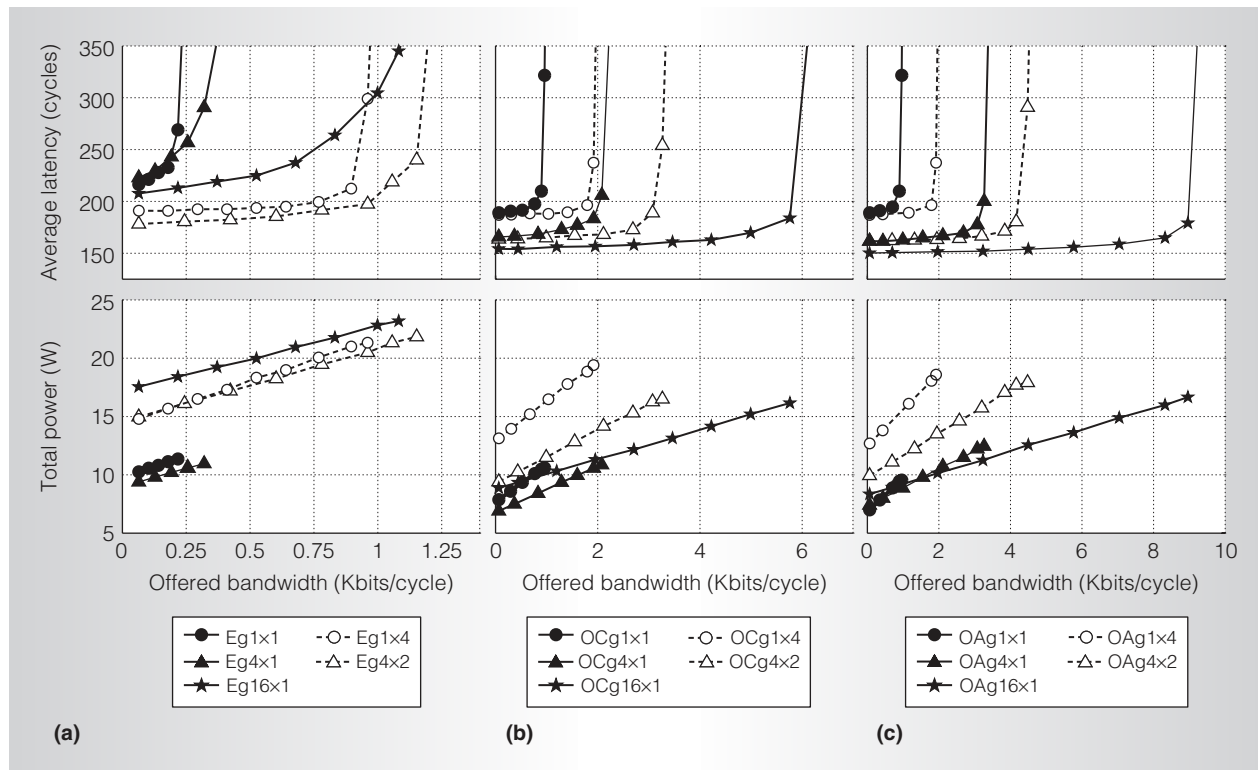
HOT INTERCONNECTS

Figure 8. Simulated performance and power for the topology configurations in Table 1 assuming electrical interconnect (a), conservative photonic interconnect (b), and aggressive photonic interconnect (c).

models a more realistic 16 DRAM modules (access points positioned in the middle of the mesh), resulting in a less uniform traffic distribution. The lower saturation point explains why Eg1x1 and Eg4x1 consume significantly less than 20 W. We investigated various OPF values for all three amounts of grouping and found that the Eg1x4 and Eg4x2 configurations provide the best trade-off. Eg1x4 and Eg4x2 increase the throughput by three to four times over the balanced configurations. Overprovisioning had minimal impact on the 16-group configuration since the local meshes are already small. Overall, Eg4x2 is the best electrical configuration. It consumes approximately 20 W near saturation.

Figures 8b and 8c show the power and performance of the photonic networks. Just replacing the off-chip I/O with photonics in a simple mesh topology (for example, OCg1x4 and OAg1x4) results in a two-times improvement in throughput. However, the full benefit of photonic interconnect only

becomes apparent when we partition the on-chip mesh network and offload more traffic onto the energy-efficient photonic channels. The OAg16x1 configuration can achieve a throughput of 9 Kbits/cycle (22 Tbps), which is approximately an order of magnitude improvement over the best electrical configuration (Eg4x2) at the same latency. The photonic configurations also provide a slight reduction in the zero-load latency. The best optical configurations consume approximately 16 W near saturation. At very light loads, the 16-group configurations consume more power than the other optical x1 configurations. This is because the 16-group configuration has many more photonic channels and thus higher static power overheads due to both leakage and thermal tuning power. The overprovisioned photonic configurations consume higher power since they require much wider mesh channels.

Figure 9 shows the power breakdown for the Eg4x2, OCg16x1, and OAg16x1 configurations near saturation. As expected, most

of the power in the electrical configuration is spent on the global channels connecting the access points to the DRAM modules. By implementing these channels with energy-efficient photonic links, we have a larger portion of our energy budget for higher-bandwidth on-chip mesh networks even after including the overhead for thermal tuning. The photonic configurations consume almost 15 W, leaving 5 W for on-chip optical power dissipation as heat. Ultimately, photonics enables almost an order of magnitude improvement in throughput at similar latency and power consumption.

Although the results are not shown, we also investigated a concentrated mesh topology with one mesh router for every four cores.[11] Concentration decreases the total number of routers (which decreases the hop latency) at the expense of increased energy per router. Concentrated mesh configurations have similar throughput as the configurations in Figure 8a with slightly lower zero-load latencies. Concentration had little impact when combined with photonic off-chip I/O.

O ur work at the network architecture level has helped identify which photonic devices are the most critical and helped establish new target device parameters. These observations motivate further device-level research as illustrated by our work on optimized waveguide crossings. We feel this vertically integrated research approach will be the key to fully realizing the potential of silicon photonics in future many-core processors. MICRO

Acknowledgments

Figure 9. Power breakdown near saturation for the best electrical and optical configurations.

....
References
1. T. Barwicz et al., ''Silicon Photonics for Compact, Energy-efficient Interconnects,'' *J. Optical Networking,* vol. 6, no. 1, 2007, pp. 63-73.
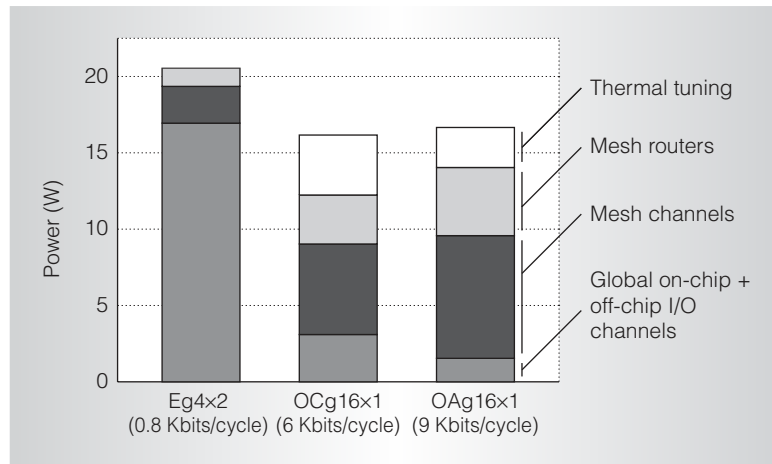
2. C. Gunn, ''CMOS Photonics for High-speed Interconnects,'' *IEEE Micro,* vol. 26, no. 2, 2006, pp. 58-66.

3. A. Shacham et al., ''Photonic NoC for DMA Communications in Chip Multiprocessors,'' *Proc. Symp. High-Performance Interconnects* (HOTI), IEEE CS Press, 2007, pp. 29-38.

4. N. Kirman et al., ''Leveraging Optical Technology in Future Bus-based Chip Multiprocessors,'' *Proc. Int'l Symp. Microarchitecture,* IEEE CS Press, 2006, pp. 492-503.

5. D. Vantrease et al., ''Corona: System Implications of Emerging Nanophotonic Technology,'' *Proc. Int'l Symp. Computer Architecture* (ISCA), IEEE CS Press, 2008, pp. 153-164.

6. C. Schow et al., ''A <5mW/Gb/s/link, 16x10Gb/s Bidirectional Single-chip CMOS Optical Transceiver for Board Level Optical Interconnects,'' *Proc. Int'l Solid-State Circuits Conf.,* 2008, pp. 294-295.

7. C. Holzwarth et al., ''Localized Substrate Removal Technique Enabling Strong-confinement Microphotonics in Bulk Si CMOS Processes,'' *Proc. Conf. Lasers and Electro-Optics* (CLEO), Optical Soc. of America, 2008.

8. J. Orcutt et al., ''Demonstration of an Electronic Photonic Integrated Circuit in a Commercial Scaled Bulk CMOS Process,'' *Proc. Conf. Lasers and Electro-Optics* (CLEO), Optical Soc. of America, 2008.

9. M. Lipson, ''Compact Electro-optic Modulators on a Silicon Chip,'' *J. Selected Topics in*

....

Quantum Electronics, vol. 12, no. 6, 2006, pp. 1520-1526.

10. B. Kim and V. Stojanovíc, ''Characterization of Equalized and Repeated Interconnects for NoC Applications,'' *IEEE Design and Test of Computers,* vol. 25, no. 5, 2008, pp. 430-439.

11. J. Balfour and W. Dally, ''Design Tradeoffs for Tiled CMP On-chip Networks,'' *Proc. Int'l Conf. Supercomputing,* ACM Press, 2006, pp. 187-198.

12. M. Popovíc, E. Ippen, and F. Kärtner, ''Low-Loss Bloch Waves in Open Structures and Highly Compact, Efficient Si Waveguide-crossing Arrays,'' *Proc. 20th Ann. Mtg. of IEEE Lasers and Electro-Optics Society,* IEEE Press, 2007, pp. 56-57.

Christopher Batten is a PhD candidate in the Electrical Engineering and Computer Science Department at the Massachusetts Institute of Technology. His research interests include energy-efficient parallel computer systems and architectures for emerging technologies. Batten has an MPhil in engineering from the University of Cambridge.

Ajay Joshi is a postdoctoral associate in MIT's Research Laboratory of Electronics. His research interests include interconnect modeling, network-on-chip design, high-speed low-power digital design, and physical design. Joshi has a PhD in electrical engineering from the Georgia Institute of Technology.

Jason Orcutt is a PhD candidate in MIT's Electrical Engineering and Computer Science Department. His research interests include device and process design for CMOS photonic integration. Orcutt has an MS in electrical engineering from MIT.

Anatol Khilo is a PhD candidate in MIT's Electrical Engineering and Computer Science Department. His research interests include the design of nanophotonic devices and photonic analog-to-digital conversion. Khilo has an MS in electrical engineering from MIT.

Benjamin Moss is an MS candidate in MIT's Electrical Engineering and Computer Science Department. His research interests include circuit design and modeling for integrated photonic CMOS systems. Moss has a BS in electrical engineering, computer science, and computer engineering from the Missouri University of Science and Technology.

Charles W. Holzwarth is a PhD candidate in MIT's Material Science and Engineering Department. His research interests include the development of nanofabrication techniques for integrated electronic-photonic systems. Holzwarth has a BS in material science and engineering from the University of Illinois at Urbana-Champaign.

Miloš A. Popović is a postdoctoral associate in MIT's Research Laboratory of Electronics. His research interests include the design and fundamental limitations of nanophotonic devices, energy-efficient electronic-photonic circuits, and nano-optomechanical photonics based on light forces. Popović has a PhD in electrical engineering from MIT.

Hanqing Li is a research scientist with MIT's Microsystems Technology Laboratories. His research interests include MEMS fabrication technologies, photonics, energy harvesting, and micro sensors and actuators. Li has a PhD in physics from the University of Nebraska-Lincoln.

Henry I. Smith is a professor in MIT's Electrical Engineering and Computer Science Department and is president of LumArray, a spin-off from MIT developing a maskless photolithography system. He supervises research in nanofabrication technology and applications thereof in electronics, photonics, and materials science. Smith has a PhD from Boston College.

Judy L. Hoyt is a professor in MIT's Electrical Engineering and Computer Science Department and associate director of MIT's Microsystems Technology Laboratories. Her research interests include fabrication and device physics of silicon-based heterostructures and nanostructures, such as high mobility Si and Ge-channel MOSFETs; nanowire FETs; novel transistor structures; and photodetectors for electronic/photonic

integrated circuits. Hoyt has a PhD in applied physics from Stanford University.

Franz X. Kärtner is a professor in MIT's Electrical Engineering and Computer Science Department. His research interests include classical and quantum noise in electronic and optical devices as well as femtosecond lasers and their applications in frequency metrology. Kärtner has a PhD in electrical engineering from Technische Universitat München, Germany.

Rajeev J. Ram is a professor in MIT's Electrical Engineering and Computer Science Department, director of the MIT Center for Integrated Photonic Systems, and associate director of the Research Laboratory of Electronics. His research interests include optoelectronic devices for applications in communications, biological sensing, and energy production. Ram has a PhD in electrical engineering from the University of California, Santa Barbara.

Vladimir Stojanović is an assistant professor in MIT's Electrical Engineering and Computer Science Department. His research interests include high-speed electrical and optical links and networks, communications and signal-processing architectures, and high-speed digital and mixed-signal IC design. Stojanović has a PhD in electrical engineering from Stanford University.

Krste Asanović is an associate professor in the Electrical Engineering and Computer Science Department at the University of California, Berkeley. His research interests include computer architecture, VLSI design, and parallel programming. Asanović has a PhD in computer science from the University of California, Berkeley.

Direct questions and comments about this article to Christopher Batten, University of California, 565 Soda Hall, Berkeley, CA 94704; cbatten@mit.edu.

For more information on this or any other computing topic, please visit our Digital Library at http://computer.org/csdl.