Building manycore processor-to-DRAM networks using monolithic silicon photonics

Ajay Joshi[+], Christopher Batten[+], Vladimir Stojanović[+] and Krste Asanović[‡]
[+]{joshi, cbatten, vlada}@mit.edu
Massachusetts Institute of Technology, Cambridge, MA,
[‡]krste@eecs.berkeley.edu
University of California, Berkeley, CA

Modern embedded systems already include dozens of cores on a single die and this number is expected 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 and FB-DIMM, are not expected to meet bandwidth demands with reasonable power consumption and packaging cost. We propose to solve this *manycore-to-memory bandwidth* problem by combining monolithic silicon photonic technology that supports *dense wavelength-division multiplexing* (DWDM) and a hybrid opto-electrical processor-memory network architecture.

Monolithic silicon photonic technology

Unlike existing approaches, we use the standard bulk CMOS flow to design new photonic devices. Figure 1 illustrates these components 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 photonic waveguide. This photonic waveguide is designed in the poly-Si layer on top of the shallow trench isolation. The Si substrate under the waveguide is etched away after chip fabrication [1], to form an air gap, which provides good optical cladding. The photonic waveguide carries the light past a series of resonant ring modulators [2], which modulate the intensity of the light at resonant wavelength. For our target system, double-ring filters enable 128 wavelengths per ring in free spectral range. 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 [2] to "drop" the corresponding wavelength from the waveguide into a local photodetector. These photodetectors make use of SiGe PMOS, and operate in 1200-1300 nm range. Although not shown in Figure 1, we can simultaneously send information in the reverse direction using different wavelengths (λ_3, λ_4) coupled into the same waveguide on chip B and received by chip A. Our analysis suggests that the total energy overhead for the various electrical back-end components of this photonic link will be less than 250 fJ/b (150 fJ/b for signaling and 100 fJ/b for heating), which is 1-2 orders of magnitude less than state-of-the-art photonic devices [3].

Processor-memory network architecture

Previous approaches have used Si-photonics for intra-chip communication [4,5]. In this work, we focus on developing a unified on-chip/off-chip photonic network to address the *manycore-to-memory bandwidth* problem. To help navigate the large design space of interconnect networks in future power-constrained systems, we developed analytical models that connect the component energy-models with network performance metrics like ideal throughput (T_{ideal}) and the zero-load latency (*ZLL*). Our target system consists of 256 cores designed at 22 nm, runs at 2.5 GHz, has a power budget of 20 W for on-chip/off-chip communication and has a large number of DRAM modules. We (optimistically) project that electrical off-chip I/O in the 22 nm node will be around 5 pJ/b while our photonic technology decreases the off-chip I/O cost to around 250 fJ/b.

Mesh topology

The mesh topology is chosen as our baseline on-chip network for its simplicity, use in practice [6] and reasonable efficiency [7]. Separate request and response mesh networks connect cores with on-chip memory access points (AP), which have dedicated I/O channels to DRAM blocks. The DRAM address space is cache-line interleaved across APs to balance the load and give good average-case performance. To maximize throughput, and minimize bottlenecks in both on-chip mesh and I/O channels, the throughput of the on-chip mesh is balanced with the throughput of off-chip I/O by choosing appropriate mesh link widths. The analytical models suggest that for fixed power budget, using photonic off-chip I/O with a simple on-chip mesh topology increases throughput by ≈ 5x compared to electrical I/O at similar latency. However, the 20x difference in energy efficiency between photonic and electrical I/O implies that there is still room for improvement.

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 energy inefficient on-chip electrical network to reach the appropriate AP. Hence, we augment the electrical mesh topology with a low energy cost photonic global crossbar between groups of cores and DRAM modules (Figure 2). Every group of cores has an

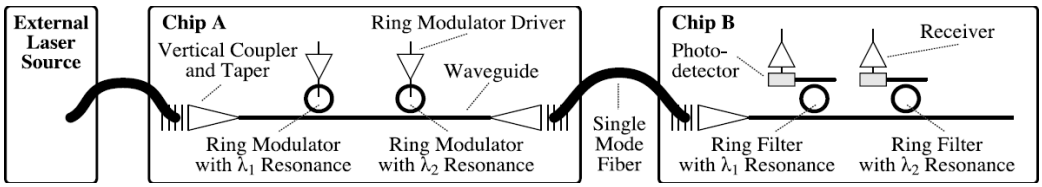

Figure 1: Photonic link using WDM.

independent AP to each DRAM module. Messages from a core reach an AP and then quickly move across the crossbar and arbitrate with messages from other groups at the global crossbar switch to access the DRAM module.

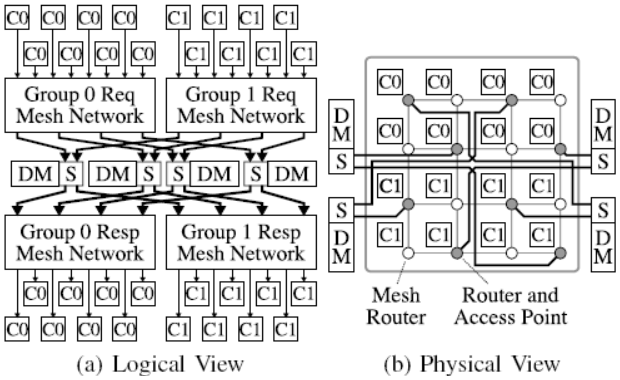

Figure 2. Mesh augmented with a global crossbar (Ci = Cores in group i, S = global crossbar switch, DM = DRAM module).

At 5 pJ/b I/O energy cost, adding a global crossbar with groups has little impact on system throughput, while for I/O energy cost of 250 fJ/b, grouping improves throughput by ≈2–3x. Combining the 5x throughput due to lower I/O energy of photonics and the 2–3x improvement from grouping, an opto-electrical global crossbar yields ≈10–15x better throughput than a simple mesh with electrical I/O. A 30% reduction in *ZLL* is also obtained at 250 fJ/b when a global crossbar with grouping is used due to reduction in the hop count in the group sub-mesh.

Simulation results

The analytical results helped narrow down the design space and set the network parameters for detailed simulations. To quantify the effects of routing protocols and contention we have developed a cycle accurate network simulation framework. The modeled system includes 256 cores, 16 DRAM modules and 256b message sizes. All mesh networks use dimension-ordered routing and wormhole flow control. 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.

Figure 3 shows the average latency as a function of injection rate for nine configurations. We consider over-provisioning the on-chip mesh to better balance the expected achievable throughput. The simulated latencies and peak offered bandwidths are different from those obtained using analytical models due to additional micro-architectural pipeline latencies, dimension-ordered routing, and realistic flow-control. In contrast to the results from analytical models, grouping significantly improved throughput for electrical configurations. This is primarily because our analytical model assumes a large number of DRAM modules while our simulated system models a more realistic 16 DRAM modules, resulting in a less uniform traffic distribution. Additional simulations not shown in Figure 3 indicated that over-provisioning the electrical network for photonic configurations either decreased performance or had no effect. The best-case optical Og16x1 configuration can achieve a throughput of 9 Kb/cycle or 22 Tb/s, which is ≈ 8-10x better than the best electrical configuration (Eg4x2), while also slightly reducing the memory access latency.

Future work

Moving forward we plan to explore other network topologies like torus, mesh with express lanes, etc. and look at other benchmark suites like HPEC, HPC, etc. to explore the design space. In addition, L2 cache will be incorporated in the network architecture to create a more realistic design.

Acknowledgement

We would like to thank the various device groups at MIT for helping us in design, post-fabrication processing and characterization of photonic devices.

References

[1] C. Holzwarth et al. Localized substrate removal technique enabling strong-confinement microphotonics in bulk Si CMOS processes. Proc. CLEO 2008.

[2] J. Orcutt et al. Demonstration of an electronic photonic integrated circuit in a commercial scaled bulk CMOS process. Proc. CLEO 2008.

[3] A. Narasimha et al. A fully integrated 4_10Gb/s DWDM optoelectronic transceiver in a standard 0.13 µm CMOS SOI. JSSC, 42(12):2736–2744, Dec 2007.

[4] A. Shacham et al.. Photonic NoC for DMA communications in chip multiprocessors. Symp. on High-Performance Interconnects, pages 29–36, Sep 2007.

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

[6] D. Wentzlaff et al. On-chip interconnection architecture of the Tile processor. IEEE Micro, pp. 15–21, Sep-Oct 2007.

[7] J. Balfour et al. Design tradeoffs for tiled CMP on-chip networks. Int'l Conf. on Supercomputing, Jun 2006.

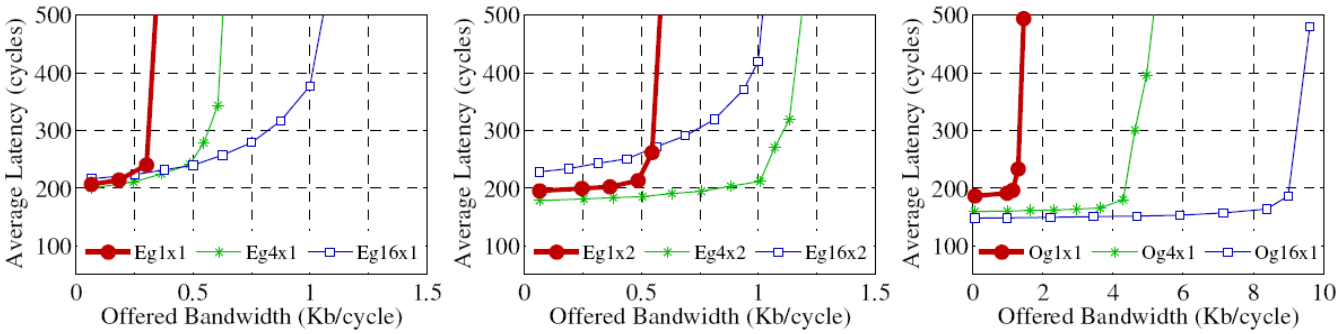

Figure 3: Simulation results for nine topology configurations (E: electrical, O: optical, g{1,4,16}: num of groups, x{1,2}: Over-provisioning factor (OPF) = ideal mesh throughput / ideal IO throughput).