Photonic Interconnection Networks for Multicore Architectures

Nathan Binkert and Marco Fiorentino
Hewlett-Packard Labs, 1501 Page Mill Rd. MS 1177, Palo Alto, CA 94304
binkert@hp.com

Abstract: Silicon nanophotonics provides computer architects with the ability to solve pin bandwidth and cross-chip communication problems. Furthermore, ring resonators can be used to create simple optical circuits to implement low-latency global arbitraion.

OCIS codes: 130.6750, 200.4650.

1. Introduction

We expect that many-core microprocessors will push chip performance from billions to trillions of instructions per second in the coming decade. To support this increased performance, memory and inter-core bandwidths will also have to scale by orders of magnitude. Pin limitations, the energy cost of electrical signaling, and the non-scalability of chip-length global wires are significant bandwidth impediments. In this scenario photonic interconnects offer a number of advantages compared to their electrical counterparts. Photonic interconnects have lower power consumption than electrical interconnects for suitably long links and the length at which photonics is more efficient than electronics has been steadily shrinking. Because in photonic links most of the power is dissipated at the extremities, power consumption is largely independent of the link length. This allows one to build networks using links with a wide array of lengths without repeaters. Photonics therefore enables low-latency network architectures with many equivalent nodes. Use of dense wavelength division multiplexing (DWDM) also increases the bandwidth density thus providing a solution to the problem of limited pin bandwidth. We have also shown [1] that, using a DWDM link architecture with multiple writers and a single receiver, one can build an all-to-all network crossbar within a reasonable power envelope. Here we focus on the advantages that photonics brings to the design of processor networks, processor-to-memory links, and distributed arbitration schemes.

2. Corona

Many-core microprocessors with thousands threads challenge the programmer, compiler, and runtime system to manage the placement and migration of programs and large amounts of data. In our work on the Corona architecture [1], we devised a photonic-based network that enables a tightly coupled, highly parallel system comprising a large number of homogeneous cores and caches. In this application photonics provides a crossbar interconnect that has near-uniform latency, a fair interconnect arbitration protocol, and high (one byte per FLOP or floating point operation) bandwidth between cores and from caches to memory.

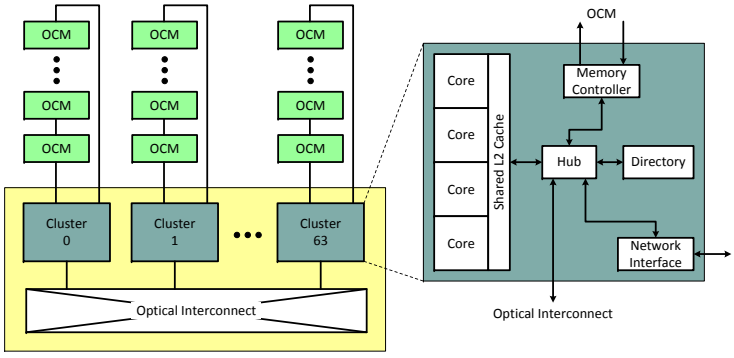

Fig. 1. Architecture Overview

Figure 1 gives a conceptual view of the system. Corona uses silicon nanophotonics in three ways: 1) to attach optically connected memory (described in Section 3) with 10 TB/s of off-chip memory bandwidth; 2) to provide 20 TB/s of global on-chip bandwidth (the optical interconnect in Figure 1); and 3) to implement optical arbitration (described in Section 4.

The Corona architecture is made up of 256 multithreaded in-order cores and is capable of supporting up to 1024 threads simultaneously, providing up to 10 teraflops of computation. The chip core network is formed by links in which multiple senders share a waveguide that is dedicated to a single receiver. This structure allows us to realize an all-to-all crossbar where each node can directly communicate with any other node. The downside is that the access to the shared resource (the waveguide) needs to be arbitrated, our solution is described in Section 4.

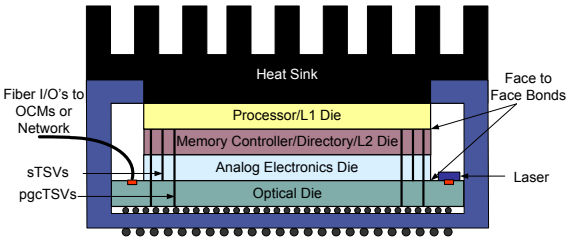

Fig. 2. Corona 3D Package

Figure 2 illustrates the Corona 3D die stack. Most of the signal activity, and therefore heat, are in the top die (adjacent to the heat sink) which contains the clustered cores and L1 caches. The processor die is face-to-face bonded with the L2 cache die, providing direct connection between each cluster and its L2 cache, hub, memory controller, and directory. The bottom die contains all of the optical structures (waveguides, ring resonators, detectors, etc.) and is face-to-face bonded with the analog electronics which contain detector circuits and control ring resonance and modulation.

All of the L2 die components are potential optical communication end points and connect to the analog die by through silicon vias. This strategy minimizes the layout impact since most die-to-die signals are carried in the face-to-face bonds. The optical die is larger than the other die in order to expose a *mezzanine* to permit fiber attachments for I/O and OCM channels and external lasers.

3. Optically Connected Memory (OCM)

Figure 3 shows the 3D stacked OCM module, built from custom DRAM die and an optical die. 3D stacking is used to increase capacity per stack and to minimize the delay and power in the interconnect between the optical fiber and the individual DRAM cells. The high-performance optical interconnect allows a single bank to quickly provide all the data for an entire cache line. In contrast, current electrical memory systems and DRAMs activate many banks on many die on a DIMM, reading out tens of thousands of bits into an open page. However, with highly interleaved memory systems and a thousand threads, the chances of the next access being to an open page are small. Corona's DRAM architecture avoids accessing an order of magnitude more bits than are needed for the cache line, and hence consumes less power in its memory system.

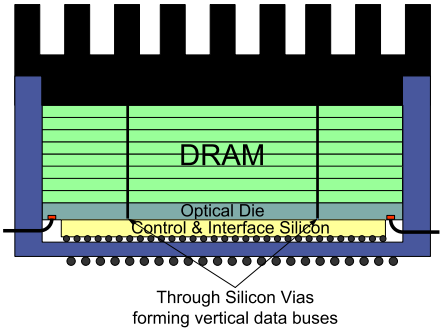

Fig. 3. Optically Connected Memory

One design goal is to scale main memory bandwidth to match the growth in computational power. Maintaining this balance ensures that the performance of the system is not overly dependent on the cache utilization of the application. Our target external memory bandwidth for a 10 teraflop processor is 10 TB/s. Using an electrical interconnect to achieve this performance would require excessive power; over 80 W assuming 1 mW/Gb/s [2] interconnect power. Instead, we use a nanophotonic interconnect that has high bandwidth and low power. The same channel separations and data rates that are used on the internal interconnect network can also be used for external fiber connections. We estimate the interconnect power to be 0.1 mW/Gb/s, which equates to a total memory system power of approximately 8 W.

Each of the 64 memory controllers connects to its external memory by a pair of single-waveguide, 64-wavelength DWDM links. The optical network is modulated on both edges of the clock. Hence each memory controller provides 160 GB/s of off-stack memory bandwidth, and all memory controllers together provide 10 TB/s.

This allows all communication to be scheduled by the memory controller with no arbitration. Each external optical communication link consists of a pair of fibers providing half duplex communication between the CPU and a string of optically connected memory (OCM) modules. The link is optically powered from the chip stack; after connecting to the OCMs, each outward fiber is looped back as a return fiber. Although the off-stack memory interconnect uses the same modulators and detectors as the on-stack interconnects, the communication protocols differ. Communication between processor and memory is master/slave, as opposed to peer-to-peer. To transmit, the memory controller modulates the light and the target module diverts a portion of the light to its detectors. To receive, the memory controller detects light that the transmitting OCM has modulated on the return fiber. Because the memory controller is the master, it can

supply the necessary unmodulated power to the transmitting OCM.

Corona supports memory expansion by adding additional OCMs to the fiber loop (see Figure 1). Expansion adds only modulators and detectors and not lasers, so the incremental communication power is small. As the light passes directly through the OCM without buffering or retiming, the incremental delay is also small, so that the memory access latency is similar across all modules. In contrast, a serial electrical scheme (e.g. FBDIMM) would typically require the data to be resampled and retransmitted at each module, increasing the communication power and access latency.

4. Optical Arbitration

In a system like Corona coordinating the various parts to avoid conflicts in the use of shared resources is a key component to guarantee functionality and an efficient use of resources. We devised photonic-based distributed arbitration schemes that can arbitrate a large number of resources with a low latency.

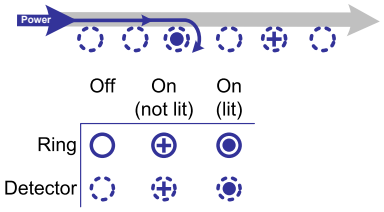

Fig. 4. Optical Arbitration

The key components of our network are silicon ring resonators and silicon waveguides. When placed next to a waveguide, a ring can be used to modulate or to detect light of its particular wavelength on that waveguide, or to divert (switch) the light from one waveguide to another. The modulation, detection, and diversion functions are controlled by applying an electrical signal to the ring, which brings it into or out of resonance with its specific wavelength. Functioning ring resonators have been demonstrated [3, 4].

An activated ring detector removes all the light in the process of detecting it, thus implementing a destructive read. When the detector is inactive, the light passes the ring unperturbed thus realizing a transparent switch. Thus, an activated detector can detect a light signal only if no upstream (towards the light source) detector is activated. The output of the detector is therefore a logical function of the state of all of the upstream detectors; this wired-or-like combinational operation is performed without any delay other than the time of flight of light in the waveguide.

In the simplest approach to optical arbitration, presented in Figure 4, a one-bit-wide pulse of monochromatic light travels down an arbitration waveguide. The presence of this light represents the availability of a resource: it is a token. Each node has a detector on this waveguide. Nodes that want to use the channel activate their detectors (solid- and cross- dotted rings); the other nodes do not activate theirs (empty dotted rings). At most one node can detect the token (solid dotted ring), because reading the token removes the light from the waveguide. As a result, a node detecting a token wins exclusive use of the channel. In our protocols, it uses the channel for some fixed period.

This simple arbitration scheme has a fixed priority, upstream nodes have higher priority than downstream nodes. Furthermore, if applied naively, it is unfair since upstream nodes could starve downstream nodes. We explore low-latency, efficient, and fair optical arbitration in [5].

5. Conclusion

Silicon nanophotonics has the potential to change several aspects of computer architecture, from off-chip communication, to on-chip interconnect, and even to global arbitration.

References

1. D. Vantrease, R. Schreiber, M. Monchiero, M. McLaren, N. P. Jouppi, M. Fiorentino, A. Davis, N. Binkert, R. G. Beausoleil, and J. Ahn, "Corona: System Implications of Emerging Nanophotonic Technology," in "Proceedings of the 35th International Symposium on Computer Architecture," (2008).
2. K. Fukuda, H. Yamashita, G. Ono, R. Nemoto, E. Suzuki, T. Takemoto, F. Yuki, and T. Saito, "A 12.3mw 12.5gb/s complete transceiver in 65nm cmos," in "Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2010 IEEE International," (2010), pp. 368 –369.
3. Q. Xu, B. Schmidt, S. Pradhan, and M. Lipson, "Micrometre-scale silicon electro-optic modulator," Nature **435**, 325 (2005).
4. L. Zhang, M. Song, T. Wu, L. Zou, R. G. Beausoleil, and A. E. Willner, "Embedded ring resonators for microphotonic applications," Optics Letters **33**, 1978–1980 (2008).
5. D. Vantrease, N. Binkert, R. S. Schreiber, and M. H. Lipasti, "Light Speed Arbitration and Flow Control for Nanophotonic Interconnects," in "Proceedings of the 42nd Annual IEEE/ACM International Symposium on Microarchitecture," (2009).