OTuH1.pdf

Optical Systems for Data Centers

Ron Ho, *Senior Member, IEEE,* Herb Schwetman, *Member, IEEE,* Michael O. McCracken, Pranay Koka, Jon Lexau, *Member, IEEE,* John E. Cunningham, Xuezhe Zheng, *Senior Member, IEEE,* Ashok V. Krishnamoorthy, *Member, IEEE*

***Abstract*—Building future data centers requires unprecedented internal bandwidth, at very high per-channel energy efficiency. Optics offers potential solutions for these challenges, especially if we rethink the construction of high-performance computers into compact systems-in-a-package. Here we discuss the drivers for optics in these large-scale systems.**

THE design, build-out, and maintenance of tomorrow's data centers face a host of challenges. Chief among them is the provisioning of adequate and efficient communication bandwidth between compute units. Insufficient communication bandwidth will throttle system performance, and inefficient communication circuits will stress thermal and power delivery limits. While photonic communication has found broad acceptance in long-haul applications, its use inside data centers and inside compute racks is still nascent. In this paper we discuss drivers for optics in data centers and then rethink the construction of high-performance systems in a way that may make the case for optics clearer.

I. Data center drivers for optics

Data centers integrate several rows of equipment racks, each about a meter thick and two meters tall. Racks are placed back-to-back, with a "hot aisle" between the backs of adjoining racks. This hot aisle, so-called because it accepts the exhausted hot air from the racked equipment, typically holds the cabling between racks. Today, data centers with 50K to 100K cores are the norm, leading to massive amounts of data center interconnect, both for core switches and end-of-row switches, with a concomitant impact on hot air flow.

Workloads and communication patterns in data centers are varied and depend on customer applications. However, the growing use of computations that rely on non-local, distributed communications means that high levels of uniform global traffic may soon dominate total system bandwidth (such as with MapReduce/Hadoop for large-scale analytics and data sorting). In addition, some estimates place internal data center traffic as 1 *million* times that of its external traffic: that is, for every byte transmitted over the internet, 1 MB is transmitted within data centers [1]. When combined with forward-looking trends of internet traffic–with a projected 45% cumulative annual growth rate–this leads to seemingly intractable bandwidth requirements within data center computers [1]. As copper cabling is heavy and has signal integrity issues when sending high data rates over distances spanning data centers, it will be hard-pressed to satisfy these needs. Optics running over lightweight, easily-replacable fiber has no significant distance limitations (at the scale of a data center), and may well play a useful role in serving such bandwidth requirements.

Another driver for data center optics lies in the energy efficiency of interconnect circuits. Today chips drive data into copper printed-circuit board (PCB) transmission lines that lead either to other chips, to a copper backplane, or to a cable plug. To get this data off the chip and into the transmission line, designers allocate solder balls between the packaged chip and the PCB. These solder balls are a precious commodity: they need to carry not only all of the data bandwidth, but also power supply current on and off the chip. As a result, data pins are typically highly over-subscribed, and must be overclocked as a result, leading to heavy serialization at the transmitter and deserialization at the receiver.

At high data rates, aggressive "SerDes" circuits for long-reach interconnect consume 5-10 mW per Gbps of bandwidth (this represents energy: a mW/Gbps is equal to a pJ/bit). When multiplied by the total bandwidth of a modern chip, this can easily scale up to a few tens of watts, or a significant fraction of the total power of the chip. Trying to ramp up total bandwidth further with data center consolidations and build-out will lead to ever-increasing power costs for that bandwidth. Not only will this stress power delivery and heat removal from the chips, but it will also increase costs, both for building provisioned HVAC systems in data centers and also for running it. Wholesale electricity costs in California averaged 4.5 cents per kilowatt-hour in 2010, so a 2.5 MW installation would cost a million dollars per year in electricity [2]. To the extent that optical communications can provide a similar bandwidth at energy costs well under 1mW per Gbps [3], they may be useful in reducing total system power, and hence total operating and ownership costs.

II. A CPU-scaling driver for optics

A third driver for optics lies in the scalability of individual processors. Historically, processor performance has grown through a combination of clock frequency scaling and architectural improvements. The former was driven by aggressive circuit and process improvements, and the latter by increases in silicon integration. However, clock frequency scaling has largely stopped, largely for power and complexity reasons: instead of the 10 GHz processors predicted by the 2001 ITRS semiconductor roadmap for 2010, we have 3

Ho and Lexau are with Oracle, Menlo Park, CA 94025 USA; Cunningham, Zheng, and Krishnamoorthy are with Oracle, San Diego, CA 92121 USA; and Schwetman, McCracken, and Koka are with Oracle, Austin, TX 78727 USA.

This work was supported in part by DARPA under Agreement HR0011-08-09-0001. The views, opinions, and/or findings contained in this article/presentation are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense. This document has been approved for public release, distribution unlimited.

©Optical Society of America

OTuH1.pdf

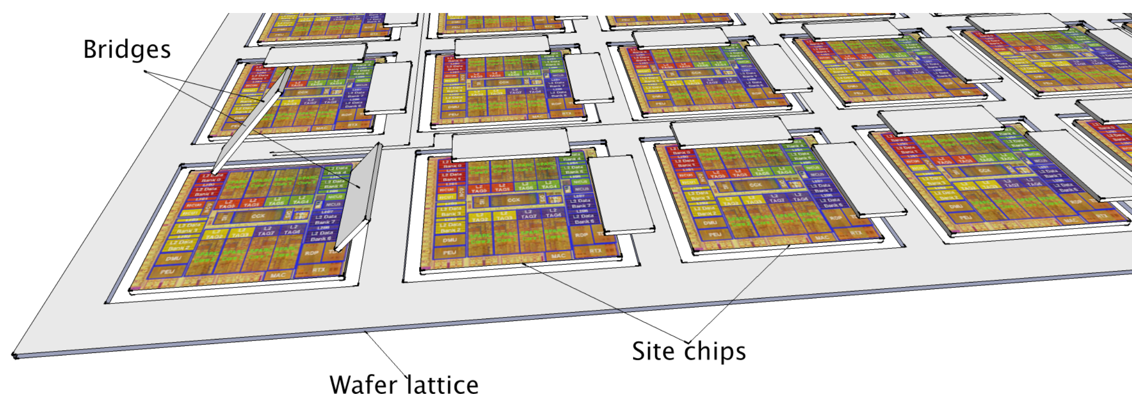

Fig. 1. Physical arrangement of multiple chips in a wafer lattice. Face-down bridges allow chip-to-waveguide connectivity. Here, two bridges are shown tilted above their final position. Taken from [4]

GHz CPUs instead [5]. Future increases in performance must come from pushing integration and leveraging architectural improvements: more cores per processor, larger caches, more functionality per core, and so on.

However, the engine for increasing integration–Moore's Law–will soon slow, perhaps dramatically [6], [7]. The reasons are many, but principal among them is limited growth in the financial investment in semiconductors. The semiconductor market is not infintely elastic, and the industry will soon no longer be able to afford the process machinery and factories to keep pushing silicon integration. The result is that designers will no longer be able to keep designing chips that are ever-larger logically, but that still fit in the same physical space.

A natural solution might be to create chips that are both logically larger and physically larger: if Moore's Law doesn't give us the ability to shrink such chips in size, why not just make really big chips? But limited chip yield from manufacturing defects means that such large chips are not economically viable. Any chip larger than 750 mm^2 or so becomes too expensive because most of the manufactured chips need to be discarded, and the overhead costs of the silicon processing must be borne by the few that survive. Optics can be useful in designing large-scale multi-chip systems, if they can help to make extremely large chips without a corresponding yield penalty. Next we describe how.

III. A MACROCHIP FOR HIGH-PERFORMANCE COMPUTING

We have proposed an optically-connected computer system comprised of an array of processors and memories arranged on a silicon lattice [3]; see Figure 1. The lattice contains passive optical waveguides, and these waveguides create a spanning network connecting all the processing and memory sites. Each site chip (processor or memory, or both) sits face-up in its location, and a small thinned bridge chip sits face-down, overlapping both the site chip and the silicon lattice. This overlap is small, enabling the processor/memory chips to still connect via traditional solder to a package from above. The bridge chip contains optical-electrical and electrical-optical circuits, and those circuits connect to the site chip through micro-solder and to the silicon lattice waveguides through a passive optical proximity communication link [3], [8],

A key part of this system is the microsolder. Although such solder is not well-suited to delivering large power supply currents, because of its short height it is ideal for carrying data from a processor or memory to a thinned bridge chip. Microsolder pitch has been demonstrated at 45 μm, and targets significantly tighter pitch in future systems [9]. At that scale, the microsolder enables sufficiently high areal density that the optical circuits to which they connect need not be serialized; this significantly relaxes their timing constraints and thus their complexity and power consumption. As a result, at this tight pitch, the dense microsolder makes the off-chip (site-to-bridge) interconnect cost the same as on-chip wires, in chip real estate, in latency, and in power.

The microsolder connects the electrical site chip to optical devices on the thin bridge. For transmitters these devices can be high-finesse ring modulators and thermal tuners [10], [11] that tie into shared waveguides using wavelength division multiplexing (WDM). For receivers these are de-mux filters followed by photodiodes with highly efficient CMOS receiver circuits on the electrical site chip [12]. In either case, waveguides on the bridge connect seamlessly to waveguides on the silicon lattice for site-to-site communication. Demonstrations of circuits and optical devices operating at 5 Gbps with 1.2 pJ/bit [13] and 10 Gbps with lower energy have shown the viability of the transceivers.

This structure, which we have called a "macrochip," shows the utility of optical communication for large arrays of processors and memories. Because the connections from each electrical site chip are sufficiently dense that they look like on-chip wires, and because the optical connections are low-latency, high-bandwidth, and able to support high-radix networks, such a system becomes a virtual single-chip processor of enormous complexity and integration. The arrayed chips look and operate as tightly integrated as if they had been fabricated on the same piece of silicon. With adequate power delivery and heat removal (such as described in [14]–[16]), this system allows designers to contemplate an entire server-

in-a-package. And not shown in Figure 1 are off-macrochip fiber connections so that macrochips can be stitched together with a secondary network. One can imagine an entire data center in a handful of racks.

IV. Networks: One of many system design issues

Many system design issues arise with this type of multi-chip aggregation, from packaging to circuits to reliability and verification to optical devices. While we acknowledge the overall difficulty of building macrochip-like systems, we will highlight just one of the design questions as an example here: What is the right network for the silicon lattice waveguides?

Networks for large-scale systems, either over multiple chips or within a single multi-core chip, have been extensively studied (for example, [17]). Sorting them by radix gives, at one extreme, very low-connectivity networks such as 2-D meshes, in which each site talks to just its neighbors in a *NEWS* structure (north-east-west-south connectivity). At the other extreme is a very high-radix network with full connectivity, such as a crossbar. The former network is simple to design but complicated to use: routing methods must avoid deadlock and deal with the large number of hops required to cross the system. By contrast, the latter network is simple to use but difficult to design and scale. Intermediate forms, such as concentrated meshes, flattened butterflies, and Clos networks, fall in the middle between these two extremes.

In most large-scale systems, designers have focused on lower-radix networks due to the cost of cabling. Building high-radix chips demands a richness of interconnect that is often unattainable, due to limited solder ball availability and restrictions on total power consumption. However, we anticipate having optical links with low overhead, so that the incremental costs of adding more optical links are negligible: waveguides are under 10 μm in pitch, and shared among 8 to 16 wavelengths; the transciever circuits are compact and likely smaller than a tiny 8kB SRAM array; and connections to the electrical chips use fine-pitch microsolder connectors.

In that case, the dictum to "waste what is cheap and conserve what is dear," would encourage the use of a high-radix, connection-rich network to the system that simplifies the programming and usage of the overall machine. Therefore, the macrochip employs a fully-connected, static point-to-point network, where each site has a dedicated optical data path to every other site. To minimize waveguides, the network uses WDM routing to share physical channels, with different routes using different colors of laser light. We envision processors at each array site to be multi-core processors, and memories to be multi-banked arrays with multiple ports to satisfy several incoming requests. As such, the point-to-point network is in reality a non-routed direct connection between two crossbars, one at the transmit site and one at the receive site.

Of course, several alternative networks can also serve, including optical dynamic packet switched networks arranged in a mesh, token rings, or circuit-switched networks. However, simulation studies based on common benchmarks support the advantages offered by simpler and more plentiful network connectivity [18]. However, because in such networks much of the provisioned bandwidth lies unused in many applications, we note a subtle design issue: circuits supporting optical interconnect typically have both static and dynamic power consumption. The latter is proportional to data activity, while the former represents a fixed overhead. Cited energy per bit for the full optical transceiver is typically for 100% activity; at 0% activity, the energy, in some circuits, may not be much lower. In those cases, the underutilized links in a fully-connected point-to-point network may be consuming more standby power than otherwise expected. Of course, as data center utilization continues to grow under global workload scaling, the importance of this issue decreases.

V. Looking ahead

Designers trying to leverage the advantages of optics to large-scale data centers have a formidable set of challenges ahead of them. These include circuit trade offs between link reliability and power, the integration and tuning of optical devices, and the mapping of large-scale applications to these systems. But optical interconnects that can provision otherwise unmanageable data center bandwidth needs, at efficiencies that make it feasible, and in order to scale up total system performance, make these challenges worth confronting.

References

[1] G. Astfalk, "Why optical data communications and why now?" *Applied Physics A*, vol. 95, no. 4, pp. 933–940, June 2009.

[2] http://www.eia.doe.gov.

[3] A. Krishnamoorthy *et al.*, "Computer systems based on silicon photonic interconnects," *Proceedings of the IEEE*, vol. 97, no. 7, pp. 1337–1361, July 2009.

[4] R. Ho *et al.*, "Optical interconnect for high-end computer systems," *IEEE Design and Test of Computers, in press*, 2010.

[5] http://www.itrs.net.

[6] G. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, no. 8, pp. 114–117, Apr. 1965.

[7] ——, "No exponential is forever," in *IEEE International Solid State Circuits Conference, Dig. of Tech. Papers*, Feb. 2003, pp. 20–23.

[8] J. Cunningham *et al.*, "Optical proximity communication in packaged Si photonics," in *IEEE International Conference on Group IV Photonics*, Sept. 2008, pp. 383–385.

[9] ——, "Aligning chips face-to-face for dense capacitive communications," in *Coupled Data Communication Techniques for High-Performance and Low-Power Computing*, R. Ho and R. Drost, Eds. New York, NY: Springer, 2010, ch. 1, pp. 157–178.

[10] X. Zheng *et al.*, "An ultra-low energy all CMOS modulator integrated with driver," *Optics Express*, no. 3, pp. 3059–3070, 2010.

[11] G. Li *et al.*, "Ultralow-power high-performance Si photonic transmitter," in *OSA Optical Fiber Communications Conference*, Mar. 2010.

[12] X. Zheng *et al.*, "A sub-picojoule-per-bit CMOS photonic receiver for densely integrated systems," *Optics Express*, no. 1, pp. 204–211, 2010.

[13] ——, "Ultra-low power silicon photonic transceivers for inter/intra-chip interconnects," in *Proceedings, SPIE Optics + Photonics*, Aug. 2010.

[14] I. Shubin *et al.*, "Novel packaging with rematable spring interconnect chips for MCMs," in *Proceedings, 59th Electronic Components and Technology Conference*, May 2009.

[15] J. Mitchell *et al.*, "Integrating novel packaging technologies for large-scale computer systems," in *ASME Pacific Rim Technical Conference and Exhibition on Packaging and Integration of Electronic and Photonic Systems, MEMS, and NEMS (Interpack)*, June 2009.

[16] I. Shubin *et al.*, "A package demonstration with solder free compliant flexible interconnects," in *Proceedings, 60th Electronic Components and Technology Conference*, June 2010, pp. 1429–1435.

[17] W. Dally and B. Towles, *Principles and Practices of Interconnection Networks*. San Francisco, CA: Morgan Kaufmann, 2003.

[18] P. Koka *et al.*, "Silicon-photonic network architectures for scalable, power-efficient multi-chip systems," in *Proceedings of the 37th International Symposium on Computer Architecture*, June 2010.