

# Optical Interconnects for Present and Future High-Performance Computing Systems

Ashok V. Krishnamoorthy  
Xuezhe Zheng  
John E. Cunningham  
SUN Microsystems,  
Microelectronics Physical  
Sciences Center,  
San Diego, CA 92121,  
USA

Jon Lexau  
Ron Ho  
SUN Laboratories,  
Menlo Park, CA 94025,  
USA

Ola Torudbakken  
SUN Microsystems  
Oslo, Norway

## ***Abstract***

*Optical interconnects are useful for high-performance electronic computing systems when the number-of-channels, the bit-rate per channel, the channel density, and the communication distance of electrical links are simultaneously stressed. We review the near-term and longer term opportunities for optical communication at the chassis, chip-package and silicon micro-system levels.*

## **1. Introduction**

Compute performance must continue to scale to meet the growing proliferation of computing capability and the reduced price per performance expected from computing systems manufacturers. We have already witnessed over 5 orders of magnitude improvement in performance-per-dollar over the last three decades, well in excess of the improvements delivered by technology scaling alone. “Redshift” applications are those that are underserved by Moore’s law, and require continued scaling and improvements in efficiencies at the system level. Examples of such applications include content servers and high-performance computing.

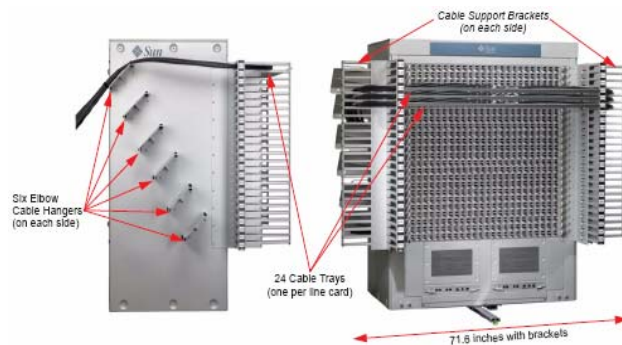
To date architects are combating these challenges by devising processor chips with multi-core, multithreaded solutions and continuing to improve the clock speed of the processor cores. But demand from high-end applications continues to grow at an even greater pace. Processors with up to 8 cores and 64 threads are now commercially available. But this

architectural scaling of processors alone is insufficient to meet all application requirements. System limitations in interconnect power, bandwidth, and density threaten to mask the benefits of improved chip performance, and hinder our ability to create large, power-efficient machines with optimal bisection bandwidth and adequate performance on critical performance benchmarks. For instance, HPC systems require sophisticated architectures for optimizing an increased number of processor-memory units, lower latency signaling between chips, and larger system bisection bandwidths for communication. Leading HPC systems are driving towards delivering petaflops of computational capacity and petabytes of storage capacity, and have bisection bandwidths between 1 and 10Tbps while future systems look for 10-100X improvements in bandwidth.

## **2. Opportunities for optical interconnects**

One of the key challenges is the choice of interconnect technology that can connect hundreds or thousands of processors without introducing unacceptable levels of latency. An example of a leading-edge HPC open-system architecture is the Constellation system [1]. This system can provide non-blocking connectivity with a port-to-port latency of 700ns between any two nodes through a 3,456-port, Infiniband (IB) switch, code-named Magnum. As shown in Figure 1, one of the key issues in deploying such a switch is the cable management. Scaling such system architectures necessitate vast numbers of interconnect modules and cables. Electrical double-data-rate IB cables were exclusively used (i.e. no

optics) in the first generation of this switched system architecture. However, as IB data-rates and bandwidth-density requirements increase, and as distance targets also increase, there is an imminent need for lighter and faster optical interconnect technologies. Based on optical interconnect modules and active optical cables currently under development, one will soon be able to design enough system bandwidth to enable optical interconnection of processor/memory units to high-density switches such as Magnum.



**Figure 1: Sun's Magnum switch side and front perspective. 18 switch cards are placed horizontally and connected via a passive orthogonal midplane to up to 24 vertically placed fabric cards (with integral fans), each providing 48 12x connectors delivering 144 DDR 4x Infiniband ports. A 12x cable and custom 12x connector supports three times the density of conventional 4x IB cables. The chassis includes integrated cable management designed to route up to 1,152 12x to either the floor or ceiling datacenter cable management infrastructure. Total throughput exceeds 110Tbps.**

Looking beyond, improvements will be needed not only to increase the effective bandwidth-density achieved by such modules and active optical cables, but also a need to bring the optical interconnect “closer in” to the processor modules. Innovative techniques to integrate fiber-optic connectors into processor packages have recently been demonstrated. Shown in Figure 2 is an example of a experimental package jointly developed with Reflex photonics that provides optical connectivity directly to a semiconductor package without requiring any changes to the chip. Such techniques can be expected to scale to multiple terabits/s of optical bandwidth to an industry-standard package.

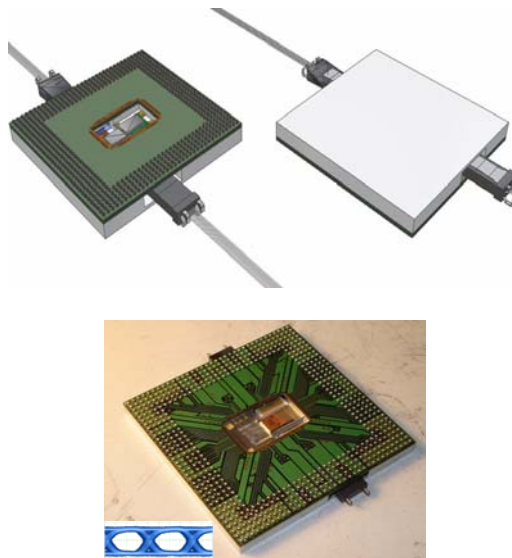
New system topologies containing multiple chips, each with multiple-cores and multi-threading would represent a major advancement if the inter-chip network and routing topologies can achieve the required capacity, functionality and latency. An examination of processor trends leads one to the conclusion that we will no longer think of processors as single chips but rather as micro-systems that integrate into a package several chips, with multiple cores, and with each core capable of launching multiple computing threads. This micro-system represents the focal point of communication and also of power consumption in the system, stressing bandwidth density and power density. Taken as an entity, such a micro-system will immensely benefit from advanced communication technologies. Within the multi-chip processor package, one may envision high bandwidth and low power electrical signaling technologies such as proximity communication (using low-energy capacitive coupling) for communication between execution cores and memory chips to achieve a flatter memory hierarchy [2]. Proximity communication provides high-bandwidth, high-density channels between chips, but is only applicable over small distances. To complement this, optical links can provide dense data communication between packages for longer spans (Figure 3).

As bandwidth demands increase further, techniques that increase the number of multiplexed optical data channels per fiber, such as WDM and multi-level encoding, will be key. By virtue of its abilities to achieve dense integration, promote electrical-optical symbiosis<sup>1</sup>, utilize wafer-scale silicon manufacturing, achieve WDM and other encoding techniques, silicon-based photonic interconnects appear to be a promising optical interconnect candidate.

As we look forward, the efficient delivery of power to the chip and the effective dissipation of the heat created on the circuits will be critical. The need for the most energy efficient optical transceivers will dominate over other design considerations. We can expect to see a hundred-fold reduction in energy to communicate an optical bit of information, thereby enabling optical interconnects to transition to the inter-

<sup>1</sup> There are many opportunities for electrical-optical symbiosis; for instance, we (and other groups) have shown that it is possible to significantly extend (by over 2x) the modulation bandwidth of optical modulators by using electronic pre-emphasis and equalization circuits. As a complementary example, it can be easier for the optical channels to transport 10-20Gbps modulated signals globally on-chip and between chips in a multi-chip package, which may otherwise require either electrical transmission lines or power-hungry electrical repeaters.

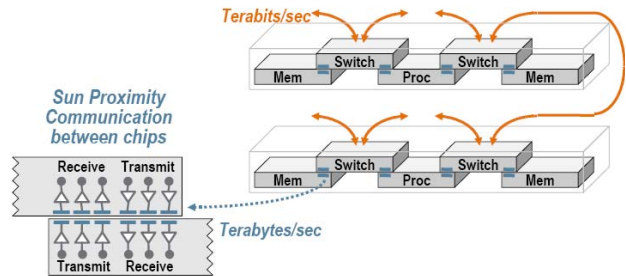
chip and intra-chip stage and provide systems level benefits exceeding that offered by scaled electrical technologies. To meet this aggressive challenge, a completely new class of disruptive components not previously explored with photonic-enhanced silicon VLSI circuits will be needed. Developing both a broad spectrum of components as well as a comprehensive toolbox of components and design tools will greatly help. Future research programs will explore changing the physical dimensionality of component designs, exploit newly discovered material science, and develop novel fabrication methods for complex, multi-dimensional optical devices. Further innovation will come from the receiver and transmitter electronics. Innovative low-voltage, low-energy circuit architectures and families can result in dramatically reduced energy per bit for end-to-end optical links. For instance low-area and low-power CMOS receivers using simple CMOS front-end circuits are projected to scale to electrical energies of 100fJ with photon energies of 1fJ [3].



**Figure 2: Experimental 45mm BGA package developed by Reflex Photonics for Sun. Fiber ribbon access is provided directly to the package. Inset: loopback eye diagram of one of four 2.5Gbps channels with a 90nm CMOS TxR chip [4]**

Prototype technology demonstrations using hybrid integration techniques have already shown that multi-terabit optical capacities can be achieved. We have also seen the advent of CMOS-compatible silicon photonics technologies that allow some of the functions of detection and modulation to be implemented in the

VLSI electronics, and hence presumably be more cost-effective and efficient. One can imagine 100Tbps of I/O from future multi-core, multi-chip optical packages. In addition to physical improvements, we can also expect architecture and system innovations to play a large role in heralding optical interconnects deeper into the system interconnect hierarchy.



**Figure 3: Future “micro-systems” will contain multiple function and process-optimized chips. One potential version would use low-power proximity links within the micro-system and optical links within and between micro-systems.**

**References:**

- [1] See <http://www.sun.com/servers/hpc/SunConstellationPreview.jsp>
- [2] R. Drost, et. al., “Challenges in building a flat-bandwidth memory hierarchy for a large-scale computer with proximity communication”, *Proc. IEEE 13<sup>th</sup> Annual Symp. on High-Performance Interconnect (HotI’05)*, pp. 13-22, 2005.
- [3] A. V. Krishnamoorthy & D. A. B. Miller, “Scaling OE-VLSI circuits into the 21<sup>st</sup> century: a technology roadmap”, *IEEE Jour. Select. Top. Quantum Elec.*, Vol. 2, No. 1, April 1996.
- [4] X. Zheng, et. al., “Optical transceiver chips based on co-integration of capacitively-coupled proximity interconnects and VCSELs”, *IEEE Photonics Tech. Letters*, Vol. 19, No. 7, April 2007.