Low Power and High Density Optical Interconnects for Future Supercomputers

Petar Pepeljugoski[(1)], Jeffrey Kash[(1)], Fuad Doany[(1)], Daniel Kuchta[(1)], Laurent Schares[(1)], Clint Schow[(1)], Marc Taubenblatt[(1)], Bert Jan Offrein[(2)], Alan Benner[(3)]

[(1)]IBM T.J. Watson Research Center, 1101 Kitchawan Rd., Yorktown Heights, NY 10598

e-mail address: petarp@us.ibm.com

[(2)]IBM Research GmbH, Säumerstrasse 4, 8803 Rüschlikon, Switzerland

[(3)]IBM Systems and Technology Group, 2455 South Rd, Poughkeepsie, NY

Abstract: Increasing performance in supercomputers requires a concomitant increase in intra-system interconnect bandwidth. We review the status and prospects of technologies required to build low power, high density board and chip level interconnects.

©2010 Optical Society of America

OCIS Codes: (200.4650) Optical interconnects; (060.2330) Fiber optics communications

1. Introduction

There has been sustained exponential growth of the performance of supercomputers (~60% per year) over the last two decades (Figure 1). We observe this trend for both the number 1 and number 500 on the top 500 list, as well as the combined performance of all 500 supercomputers on the list [1]. While the scaling of the performance is primarily achieved by increasing the number of cores, it requires improvements in all aspects of the system. A balanced system design requires corresponding increase in the intra-system interconnection bandwidth. However, it is not practical for power consumption of the system to scale with performance. If we simply scale today's PF machines, we would need gigawatts of power for an Exascale machine, which makes building such machine impossible. To build an acceptable Exascale supercomputer (around 2020) the total power consumption needs to be less than approximately 50 MW [2, 3]. Based on aggressive assumptions for power consumption gains (around 50%) in supercomputers from the past several years such a system is possible. If we could stay on the trend of ten fold increase in performance for a two-fold increase in power consumption, then we could even expect to build a 20MW Exascale machine (Table 1). However, this will require some rather extreme power take downs, including optical links (e.g. from today's 30-40 mW/Gbps to sub-mW/Gbps).

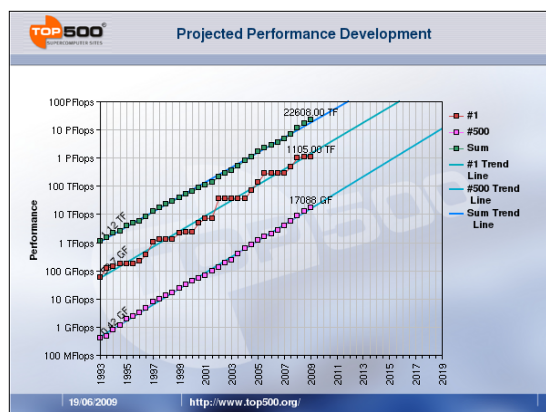

Figure 1. Performance trends for supercomputers.

Table 1. Projected trends for required power density.

Year	Peak Performance	Power Consumption	(Bidi) Optical Bandwidth	Optics Power Consumption
2008	1PF	2.5 MW	0.012PB/s ($1.2 \cdot 10^5$Gb/s)	0.012MW
2012	10PF	5 MW	1PB/s (10^7Gb/s)	0.5MW
2016	100PF	10 MW	20PB/sec ($2 \cdot 10^8$Gb/s)	2MW
2020	1000PF (1EF)	20 MW	400PB/sec ($4 \cdot 10^9$Gb/s)	8MW

In the past, electrical interconnects were able to keep pace with the requirements on intra-system interconnect bandwidth. However, increasing bitrates (required to reduce the power consumption and achieve high spatial density per bit), make it increasingly difficult for electrical interconnects to compete with optics, except perhaps for short

intra-rack interconnects. While optical interconnects have already displaced electrical interconnects in rack to rack intra-system interconnects [4], their use will continue to grow as volumes increase and costs decrease. Nonetheless, optical interconnect technologies will have to overcome significant challenges to reach the required power, cost, and density. These challenges will require innovations in packaging, test, and reliability. Here we review technologies for intra-system optical interconnects (rack to rack, board to board and chip to chip) currently in use or considered for use in future supercomputers.

2. **Optics in today's supercomputers and servers**

The first application of optical technology in supercomputers and servers server systems is for inter-rack communication distances and beyond. Typically, the optics is placed at the edge of the card, far from the host adapter. Since most links are typically <30m, with the longest links ~100m, solutions based on active optical cable technology have been deployed. Active optical cables provide optical interconnect technology with the optics hidden inside the cable connector housing. This offers the important advantage that no optical know-how is required to assembly the system; the cables are mounted at the edge of the system in the same manner as electrical cables, and the specifications are necessary only for the electrical interface. The large majority of active optical cables use the same components as those in multi-gigabit LAN links like 10 Gigabit Ethernet and Quad Data Rate Infiniband. Components include Vertical Cavity Surface Emitting Lasers (VCSELs), high speed multi-mode fibers and optical receivers utilizing large area photo-diodes. As an example, the IBM Roadrunner supercomputer [4] employs active optical cable assemblies, whose total interconnect length is over 92 km of multi-mode fiber.

The placement of the optical transmitter and receiver on the edge of the card in a format compatible with electrical interfaces has implications for both the power consumption and the density of optical interconnects. The presence of a long electrical link to connect the host adapter or CPU to the optical transmitter or receiver limits the data rate on the electrical links, due to signal attenuation. In many instances, depending on the length of the electrical link, it also requires the use of equalization on the transmitting or receiving side (or both), and/or intermediate clock and data recovery circuits. All of this increases the total power consumption of the electrical-optical-electrical (EOE) link. Furthermore, the requirement for the packaging form factor to be compatible with existing electrical interfaces places a burden on the size of the optical interconnects. The optical transmitter and receiver are capable to be packaged in a much smaller form factor, enabling much higher density than the electrical interconnects. For example, an active cable would achieve 0.11-0.22 Gb/s/mm^2, while placing the optics inside the board and having a single multi-channel multi-layer connector (4x12 or 6x12) MTP could achieve anywhere between 1.7-10.3 Gb/s/mm^2.

While active optical cable technology is attractive for today, it is not suitable for the future supercomputers. Due to power and density considerations, as well as the overall size of the system, the current concept of placing the optics on the edge can not be sustained for very long and the optics will need to be placed very close to the host adapter or CPU. This will require much tighter integration of the optics with the rest of the system.

3. Optical interconnect roadmap: optical technologies

There have been numerous demonstrations of optical technologies suitable to be integrated on the board and placed close to the host adapter [5-7]. As a first step, electrical and optical components will be integrated in a hybrid fashion for board-level and off-board links. Several challenges need to be solved in order to enable electro-optical processor packages. The bandwidth requirements in computing systems demand the application of high-density optical transceivers with a large channel count. Optical signal transport may be provided by polymer optical waveguides integrated into the card, as it will be cost-inefficient to route and connect a large number of fibers. Polymer waveguide technology has been shown to be compatible with established fiber-based multimode parallel optical links, with bandwidth in excess of 40 GHz·m and propagation loss around 0.05 dB/cm at 850 nm. To enable simple and low-cost assembly of optical connectors and transceivers, we have developed passive alignment schemes based on standard electronic assembly processes applied in today's electrical printed circuit board manufacturing [5]. Transceiver technology integrated with board-level polymer waveguide optical links has been demonstrated [6]. Figure 2 shows a transceiver packaging concept based on 3D integration of the opto-electronic elements with the CMOS driver circuits. The optical signals are transmitted through 30 cm long polymer waveguides integrated on an otherwise conventional printed circuit board [6].

OThX2.pdf

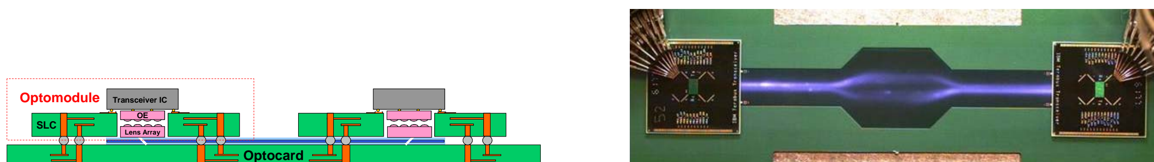

Figure 2. 160 Gb/s full duplex board-level optical link, including transceiver packages

This technology may support two orders of magnitude improvement in supercomputer performance beyond the current 1 PF system [4]. In support of this improvement in system performance, it has been shown that multimode links using VCSELs can achieve data rates up to 30-40 Gb/s [8]. High count parallel optical transmitters and receivers compatible with multimode links have been demonstrated [7]. Up to 48 channels were demonstrated operating at 20 Gb/s for the transmitter and 15 Gb/s for the receiver. This means that, for example, a 48 channel transmitter module can achieve 1920 Gb/s.

While innovations are possible, we believe that for Exaflop systems, a transition to wavelength division multiplexing will be required. This transition will include change in the technology used for the optical transmitters and receivers from 850-nm to 1550-nm, using CMOS silicon photonic technology. 3D integration technologies may be used at the processor level, to create separate planes for the logic, the memory and the silicon photonics. In this concept, logic and memory have electrical vias to the silicon photonics plane, where the signals are first converted in the optical domain using electro-optical modulators and then multiplexed into a single WDM data stream. We anticipate that silicon photonics can help to substantially improve the interconnect power and density efficiencies. The enormous intra-chip bandwidth requirements and the limited area available demand the development of ultra-small electro-optical and optical components [9].

Conclusions

Optical interconnects will play increasingly important role in future supercomputers and servers. The intra-system optical interconnects (rack-to-rack, board-to-board and chip-to-chip) of the future will feature higher levels of electro-optical integration, compared to existing commercial technologies. We expect the first board-level optical interconnects to be based on hybrid packaging concepts, providing solutions for the supercomputers up to ~100-200 PF. For true Exaflop supercomputers and for processor-level integration, we envision a transition from multimode based optics to a silicon-compatible WDM technology and ultra-small ultralow power optical components.

References

1. http: //www.top500.org
2. Alan Benner, “Cost effective optics: enabling the exascale roadmaps”, IEEE Hot Interconnects 17, August 2009, New York, NY.
3. A. F. Benner et al., "Exploitation of optical interconnects in future server architectures," IBM J. Res. & Dev., vol. 49, pp. 755-775, 2005.
4. http://www.lanl.gov/roadrunner/.
5. R. Dangel et al., “Polymer-waveguide-based board-level optical interconnect technology for datacom applications,” IEEE Trans. Adv. Pack., vol. 31, pp. 759-767, 2008.
6. F. E. Doany.et al., “160 Gb/s bidirectional polymer-waveguide board-level optical interconnects using CMOS-based transceivers,” IEEE Trans. Adv. Pack., vol. 32, pp. 345-359, 2009.
7. C. L. Schow, F. E. Doany, C. Tsang, N. Ruiz, D. Kuchta, C. Patel, R. Horton, J. Knickerbocker, and J. Kash, “300-Gb/s, 24-channel full-duplex, 850-nm, CMOS-based optical transceivers,” Optical Fiber Communication Conference – OFC 2008, paper OMK5, San Diego, CA, Feb 24-28, 2008
8. R. Johnson and D. Kuchta: “30 Gb/s Directly Modulated 850 nm Datacom VCSELs”, CLEO 2008, San Jose, CA. May, 2008.
9. Y. Vlasov et al., "High-throughput silicon nanophotonic wavelength-insensitive switch for on-chip optical networks," Nature Photonics, vol. 2, pp. 242-246, 2008.