Optical Interconnects in Future Servers

Jeffrey A. Kash[1], Alan F. Benner[2], Fuad E. Doany[1], Daniel M. Kuchta[1], Benjamin G. Lee[1], Petar K. Pepeljugoski[1], Laurent Schares[1], Clint L. Schow[1], Marc Taubenblatt[1]

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

[2]IBM Systems & Technology Group, Poughkeepsie, NY 12601 USA

jeffkash@us.ibm.com

Abstract: Optical interconnects are common in today's petascale supercomputers, and will become pervasive at the exascale during this decade. Technologies that can meet the challenging technological and economic requirements for the exascale will be reviewed.

OCIS codes: (200.4650) Optical interconnects; (230.3120) Integrated Optics Devices; (060.4510) Optical communications.

1. Optics in Petascale Supercomputers

In 2008, Roadrunner[1], the first petaflop supercomputer (i.e., $>10^{15}$ floating point operations per second) was constructed. The large number of microprocessor cores in such petascale machines must be interconnected with a high capacity communications network to permit efficient computation. Higher interconnect bandwidth will generally result in more efficient use of the microprocessors in real calculations. Prior to 2005, this network used electrical interconnects in essentially all supercomputers. Communication bitrates increased beyond ~2Gb/s starting in about 2005. For distances greater than about 20m, electrical interconnects were impractical and optics began to be used for these longer rack-to-rack interconnects (e.g., ASCI Purple[2]). For cost reasons, however, the shorter rack-to-rack interconnect in this machine was still electrical. By 2008, communication bitrates had increased to 5Gb/s (InfiniBand DDR), and the practical reach of electrical interconnects was shorter. At the same time, the cost of optics had decreased. As a result, for Roadrunner, all node-to-node communication was optical. However, in these examples, all the intra-node communication was electrical. Also, the conversion to optics was at the edge of the node/rack, requiring an electrical interconnect from the chip through several levels of packaging before conversion to photons. For machines with peak performance beyond 10PF, optics will play a far more important role. For example, in the planned POWER7-IH systems[3], all the board-to-board interconnects, will be optical. A major change in the packaging tightly integrates the optics onto the chip-level package. Optical transmitter and receiver modules are placed on the same ceramic Hub (switch) Module[4, 5] as the CMOS Hub chip, forming a first-level package with (284+284) GByte/s of electrical I/O bandwidth plus (280+280) GByte/s of optical I/O bandwidth. There are 28 TX "MicroPOD^{TM}" modules[6] and 28 RX modules on the POWER7 Hub Module in addition to the Hub chip. Each MicroPOD optical module has 12 VCSELs or photodiodes plus drive electronics in a footprint less than 0.65cm^2, with each channel using a 10Gb/s signaling rate and 8b10b coding.

2. Projections for future supercomputers

Because microprocessor clock speeds are not rapidly increasing, supercomputing performance improvements come from larger numbers of processing cores operating in parallel. For example[7], the #1 machine on the top500 list from June, 2002 was the NEC Earth-Simulator with 5120 single core processor chips, as compared to Roadrunner which has some 97,920 cores in 12,240 Cell Broadband Engine® processor chips plus 6,562 dual-core AMD Opteron® chips[8]. This increasing core and chip count drives increasing communication bandwidth at higher bitrates.

Trends for optical interconnects in future supercomputers can be extrapolated from past growth. Historically, computational power has grown approximately 10x every 4 years[7], and this growth is expected to continue. Affordable supercomputing requires that the cost and power consumption of supercomputers grow more slowly than computational power. A linear extrapolation of power and cost from the Roadrunner 1PF system[8] would result in the first ExaFlop/s (EF) machine (circa 2020) consuming ~2GW and costing ~$\$10^{11}$, neither of which is realistic. Past trends, would suggest that each 10x increase in performance has been accompanied by roughly a 2X power increase and a 1.5X cost increase at the system level, resulting in an EF machine consuming ~220MW and costing ~$500M.

An aggressive assumption would be that optics will be used even for on-chip communications in EF supercomputers[9, 10]. Even with a less aggressive use of optics, one should at least expect that optics will support most of the off-chip communications, including on-card links[11] between processors and processor-to-memory links[12]. This increased use of optics can only occur if the cost and power of an optical link can be substantially

reduced. Projections[13] suggest that a power below 1pJ/bit and a cost well below $0.10/Gb/s are required for an EF supercomputer around the year 2020.

3. Optical technologies for the exascale

Several optical technologies could be deployed as future servers approach the true EF performance. One direction is to simply improve the density and power consumption of VCSEL / multimode fiber parallel optical modules. A recent example[14] could represent a 20x improvement in bandwidth density over today's MicroPOD™ modules[6] if the were was pushed to 25Gb/s[15].

Unfortunately, such a MMF-based module does little to mitigate the fiber count explosion for an EF machine. Optical fiber cables are expensive and prone to errors during the manual connecting of this optical "wiring". Their bulk becomes also becomes an issue as the cable count increases. Replacing the fibers with polymer waveguides on the printed circuit board[11, 16] can eliminate much of the manual assembly, fiber count and cost, just as the introduction of printed circuit boards did for electrical wiring. The chip-like transceivers for these links have already demonstrated power as low as 5pJ/bit[17]. Advances in VCSELs, photodiodes and drive circuits[18] may let these links begin to approach true pJ/bit levels.

Ultimately, it will be necessary to reduce the fiber count. One way to do this is to put multiple cores in the fiber[19], although this approach means larger fiber diameters and also cannot likely be extended to the same number of channels per fiber as wavelength-division multiplexing (WDM). WDM has the potential to reduce the fiber count by an order of magnitude or more. In addition, if WDM is implemented by the use of silicon photonics, there is a significant opportunity[20] to reduce power consumption below VCSEL-based links. A VCSEL-based optical link in a server is really an EOE link: an optical link with a controlled-impedance electrical link on each side. Silicon photonics[9, 21, 22] can potentially be integrated directly on the logic chip. In that case, the electrical part of the link can become a low power on-chip electrical interconnect. Such integration is inevitable if we are to reach the pJ/bit goal for EF machine. Reduction of the link cost is another potential advantage of silicon photonics, which can take advantage of the existing IC manufacturing infrastructure, in addition to the reduced fiber count from WDM. Commercialized silicon photonics could meet the cost and power targets for a true EF supercomputer. But commercialization presents many unsolved challenges, including true monolithic integration[23-25], low cost optical packaging[26], temperature control[27] and the laser source[28].

4. References

[1] D. Grice, H. Brandt, C. Wright, P. McCarthy, A. Emerich, T. Schimke, C. Archer, J. Carey, P. Sanders, J. A. Fritzjunker, S. Lewis, and P. Germann, "Breaking the petaflops barrier," *IBM J. of Res. & Dev.,* vol. 53, pp. 1:1 - 1:16, 2009.

[2] A. F. Benner, M. Ignatowski, J. A. Kash, D. M. Kuchta, and M. B. Ritter, "Exploitation of optical interconnects in future server architectures," *IBM J. Res. & Dev. ,* vol. 49, pp. 755-775, 2005.

[3] A. F. Benner, D. M. Kuchta, P. K. Pepeljugoski, R. A. Budd, G. Hougham, B. V. Fasano, K. Marston, H. Bagheri, H. Xu, D. Meadowcroft, M. H. Fields, L. McColloch, M. Robinson, F. W. Miller, R. Granger, D. Childer, and E. Childers, "Optics for High-Performance Servers and Supercomputers," *Proceedings of OFC, paper OTuH1,* 2010.

[4] S. Clark, B. Arimilli, B. Drerup, J. Lewis, J. Irish, D. Krolak, K. Imming, J. McDonald, A. Koenig, D. Dreps, D. Siljenberg, S. Baumgartner, G. Wiedmeier, J. Strom, D. O'Connor, A. Maki, D. Sejpal, M. Ritter, D. Friend, and C. Geer, "The IBM Hub Module in 45nm CMOS SOI: A Terabyte Interconnect Switch for High-Performance Computer Systems," in *Proceedings of Hot Chips 22*, 2010.

[5] B. Arimilli, R. Arimilli, V. Chung, S. Clark, W. Denzel, B. Drerup, T. Hoefler, J. Joyner, J. Lewis, J. Li, N. Ni, and R. Rajamony, "The PERCS High-Performance Interconnect," *18th IEEE Symposium on High Performance Interconnects,* pp. 75-82, 2010.

[6] http://www.avagotech.com/docs/sm-avago-supercomputing.pdf.

[7] http://www.top500.org/.

[8] http://www-03.ibm.com/press/us/en/pressrelease/24405.wss.

[9] R. G. Beausoleil, "A Nanophotonic Interconnect for High-Performance Many-Core Computation," *Proceedings of SPIE Photonics West,* 2008.

[10] J. A. Kash, "IntraChip Optical Networks for a Future Supercomputer-on-a-Chip," in *Photonics in Switching*, San Francisco, CA, 2007, pp. 55-56.

[11] F. E. Doany, C. L. Schow, C. W. Baks, D. M. Kuchta, P. Pepeljugoski, L. Schares, R. Budd, F. Libsch, R. Dangel, F. Horst, B. J. Offrein, and J. A. Kash, "160 Gb/s Bidirectional Polymer Waveguide Board-Level Optical Interconnects using CMOS-Based Transceivers," *IEEE Trans. Adv. Pkg.,* vol. 32, pp. 345-359, 2009.

[12] Y. Katayama and A. Okazaki, "Optical Interconnect Opportunities for Future Server Memory Systems " in *Proceedings of IEEE 13th International Symposium on High Performance Computer Architecture*, 2007, pp. 46-50.

[13] J. A. Kash, A. F. Benner, F. E. Doany, D. M. Kuchta, B. G. Lee, P. K. Pepeljugoski, L. Schares, C. L. Schow, and M. Taubenblatt, "Optical Interconnects in Exascale Supercomputers," *Proceedings of the IEEE Photonics Society Annual Meeting, Paper WR1,* pp. 483-484, 2010.

[14] F. E. Doany, C. L. Schow, B. G. Lee, A. V. Rylyakov, C. Jahnes, Y. Kwark, C. Baks, D. M. Kuchta, and J. A. Kash, "Dense 24 TX + 24 RX Fiber-Coupled Optical Module Based on a Holey CMOS Transceiver IC," *Proceedings of ECTC,* 2010.

[15] R. H. Johnson and D. M. Kuchta, "30 Gb/s Directly Modulated 850 nm Datacom VCSELs," in *Conference on Lasers and ElectroOptics (CLEO) postdeadline paper CPDB2*, 2008.

[16] R. Dangel, C. Berger, R. Beyeler, L. Dellmann, M. Gmur, R. Hamelin, F. Horst, T. Lamprecht, T. Morf, S. Oggioni, M. Spreafico, and B. J. Offrein, "Polymer-Waveguide-Based Board-Level Optical Interconnect Technology for Datacom Applications," *IEEE Transactions on Advanced Packaging,* vol. 41, pp. 759-767, 2008.

[17] C. L. Schow, F. E. Doany, C. Chen, A. V. Rylyakov, C. W. Baks, D. M. Kuchta, R. A. John, and J. A. Kash, "Low-power 16 x 10 Gb/s Bi-Directional Single Chip CMOS Optical Transceivers operating at < 5 mW/Gb/s/link " *IEEE J. of Solid State Circuits,* vol. 44, pp. 301-313, 2009.

[18] C. P. Lai, C. L. Schow, A. V. Rylyakov, B. G. Lee, F. E. Doany, R. A. John, and J. A. Kash, "20-Gb/s Power-Efficient CMOS-Driven Multimode Links," in *Proceedings of OFC*, 2011.

[19] B. G. Lee, D. M. Kuchta, F. E. Doany, C. L. Schow, C. Baks, R. John, P. Pepeljugoski, T. F. Taunay, B. Zhu, M. F. Yan, G. E. Oulundsen, D. S. Vaidya, W. Luo, and N. Li, "Multimode Transceiver for Interfacing to Multicore Graded-Index Fiber Capable of Carrying 120-Gb/s over 100-m Lengths," in *Proceedings of the Annual Meeting of the Photonics Society, talk ThB6*, 2010, pp. 564-565.

[20] D. A. B. Miller, "Device Requirements for Optical Interconnects to Silicon Chips," *Proc. IEEE,* vol. 97, pp. 1166-1185, 2009.

[21] T. Barwicz, H. Byun, F. Gan, C. W. Holzwarth, M. A. Popovic, P. T. Rakich, M. R. Watts, E. P. Ippen, F. X. Kärtner, H. I. Smith, J. S. Orcutt, R. J. Ram, V. Stojanovic, O. O. Olubuyide, J. L. Hoyt, S. Spector, M. Geis, M. Grein, T. Lyszczarz, and J. U. Yoon, "Silicon photonics for compact, energy-efficient interconnect," *Journal of Optical Networking,* vol. 6, pp. 63-73, 2007.

[22] http://www.research.ibm.com/photonics.

[23] R. Soref, "The Past, Present, and Future of Silicon Photonics," *IEEE J. Sel. Topics Quant. Electron.,* vol. 12, pp. 1678-1687, 2006.

[24] P. Dumon, W. Bogaerts, R. Baets, J.-M. Fedeli, and L. Fulbert, "Towards foundry approach for silicon photonics: silicon photonics platform ePIXfab," *Electronics Letters,* vol. 45, pp. 581-582, 2009.

[25] C. Gunn, "CMOS Photonics for High-Speed Interconnects," *IEEE Micro,* vol. 26, pp. 58-66, 2006.

[26] B. G. Lee, F. E. Doany, S. Assefa, W. M. J. Green, M. Yang, C. L. Schow, C. V. Jahnes, S. Zhang, J. Singer, V. I. Kopp, J. A. Kash, and Y. A. Vlasov, "20-µm-Pitch Eight-Channel Monolithic Fiber Array Coupling 160 Gb/s/Channel to Silicon Nanophotonic Chip," *Proceedings of OFC, paper PDPA4,* 2010.

[27] H. F. Hamann, A. Weger, J. A. Lacey, Z. Hu, P. Bose, E. Cohen, and J. Wakil, " Hotspot-Limited Microprocessors: Direct Temperature and Power Distribution Measurements," *IEEE Journal of Solid-State Circuits,* vol. 42, pp. 56-65, 2007.

[28] H. Park, A. W. Fang, S. Kodama, and J. E. Bowers, "Hybrid silicon evanescent laser fabricated with a silicon waveguide and III-V offset quantum wells," *Optics Express,* vol. 13, pp. 9460-9464, 2005.