Scaling HPC Networks with Co-packaged Optics

P. Maniotis, L. Schares, B. G. Lee, M. A. Taubenblatt, D. M. Kuchta
IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA. ppmaniotis@ibm.com

Abstract: We propose an HPC network architecture with co-packaged optics enabling 128-port 51.2-Tb/s switches. Simulations for a >34,000-GPU system show up to 11.2x throughput improvement over a Summit-like supercomputer, opening the way to direct-network-attached GPUs.
OCIS codes: (200.4650) Optical interconnects; (060.4258) Networks, network optimization

1. Introduction

High Performance Computing (HPC) and high-end AI systems have a continued need for more network bandwidth (BW), lower latency and better energy efficiency [1]. Many modern workloads in HPC and distributed learning are primarily running on GPUs (or other accelerators) and use data sets that are far larger than the on-GPU memory (High-Bandwidth-Memory, HBM) [2]; they need to pull data from main memory or through oversubscribed networks, which is inefficient. In addition, moving large data sets from storage into the main memory is often constrained by BW and is known to severely limit AI training [3], [4]. To keep up with workload demands, continued network switch scaling towards the 51.2-Tb/s and beyond is a major activity in the data center and HPC industry. However, to overcome BW density and thermal cooling limits, highly energy-efficient and dense I/O solutions are required. A promising solution for continued BW scaling is the integration of optics onto the first level package, a.k.a. co-packaged optics [5], [6], since it can: (a) minimize the power of the electrical links to/from the optics and (b) substantially increase the total escape BW of the chip packages by offering an additional dimension for wiring additional chip pins, alleviating the limitations associated with the Ball Grid Arrays (BGAs) in which the pin count is typically at least 4x lower than the corresponding ASIC pin count [7], [8].

In this paper, we study the advantages of using co-packaged optics in HPC fat-tree networks where the switch radix is mainly limited by a combination of power, chip area, and package BGA limitations. Co-packaged optics can alleviate these issues by offering higher switch radixes that can lead to higher bisection BW and flatter topologies. In this work, we present a comparative simulation analysis between the Summit network architecture (world's #1 system on the June-2019 Top500 list [9]), and a network architecture where co-packaged optics is used in the network. This analysis has been realized within the framework of the MOTION research project [10], a collaboration between IBM and Finisar to develop a VCSEL-based chip-scale optical module that can be directly attached to the top of an organic first level package. The simulation results show that the co-packaged optics in future HPC networks can lead to significant advantages, offering up to 11.2x better throughput for representative synthetic benchmarks emulating network-bound workloads running on GPUs or other accelerators.

2. Summit-like and MOTION Network Architectures

Fig. 1 (a) illustrates how co-packaged optics can be used in a Summit-like compute node. This paper focuses on integrating co-packaged optics on switches, but the I/O requirements of CPUs and GPUs are not far behind. The example in Fig. 1(b) illustrates how a 51.2-Tb/s switch module could be built on an organic 90x90-mm^2 carrier populated with two switch ASICs and 3.2-Tb/s optical modules, using only optics for all I/O. We formalized this illustration through a spreadsheet analysis whose results are plotted in Fig. 1(c), also including electrical I/O through package pins and using the following assumptions: two 30x20-mm^2 ASICs with 256 SerDes at 112-Gb/s signaling, three different carrier sizes with pin pitch of 1.06 mm, 1:1 signal-to-ground ratio, 25% of package pins used for high-speed I/O, <4-pJ/bit optics ([10]) to yield manageable thermal densities, and optical modules of variable size,

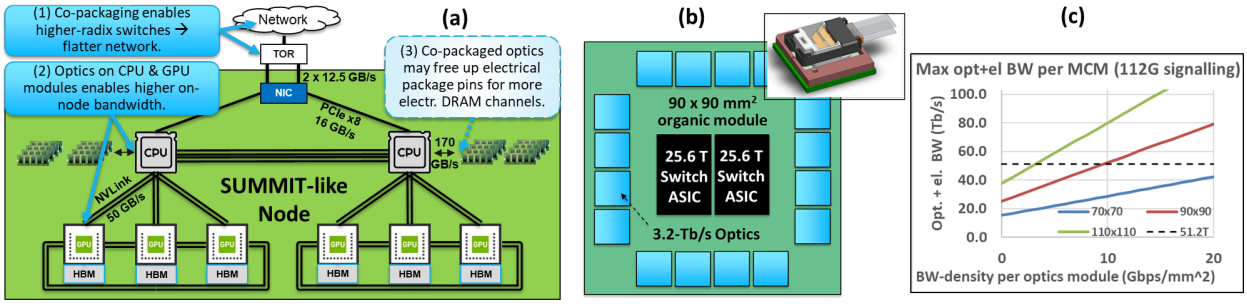

Fig. 1: (a) Possible insertion points of co-packaged optics on switches, CPUs or GPUs, (b) example of co-packaged optics enabling 51.2-Tb/s switch on 90 x 90-mm^2 carrier with 13x13 mm^2-modules (insert: MOTION phase-1 module), (c) max. off-module BW for 3 carrier sizes (70x70, 90x90, 110x110 mm^2) assuming a 40% fill factor and 112-Gb/s signaling.

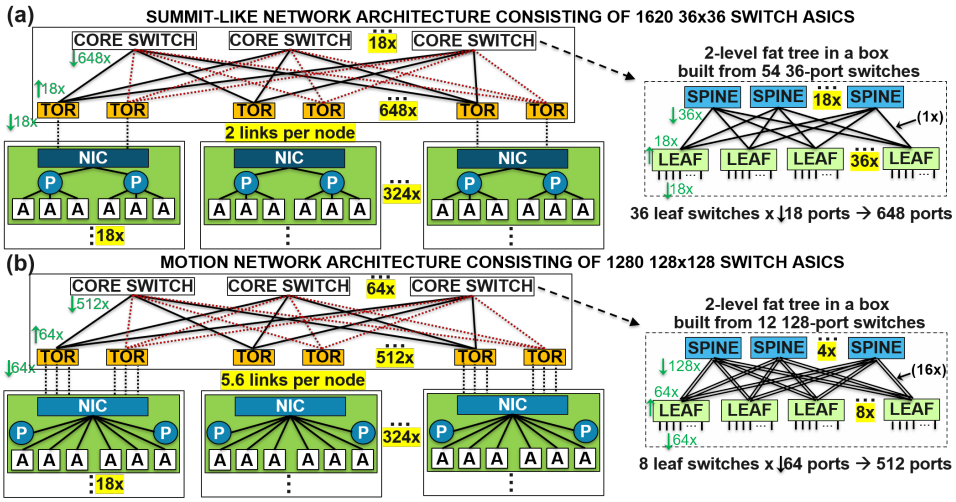

Fig. 2: (a) Summit-like network architecture, (b) MOTION network with co-packaged optics.

i.e. with their BW-density being a variable. The plot in Fig. 1(c) assumes that 40% of the free carrier area (i.e. w/o ASICs) is occupied by optics; this fill factor corresponds to 8/16/25 co-packaged optics modules for the 70x70 / 90x90 / 110x110-mm^2 carriers, respectively. For the example of the 90x90-mm^2 carrier, we would need 16 3.2-Tb/s 13x13-mm^2 optical modules for the 51.2-Tb/s BW if all I/O is optical (BW-density of 18.9 Gbps/mm^2). If the electrical I/O is included, we would get by either with fewer optical modules or with 16 modules having a reduced BW-density of ~10 Gbps/mm^2.

Fig. 2(a) presents a high-level illustration of the Summit-like network architecture that consists of 1,620 36-port switches and 5,832 compute drawers, or nodes. To simplify our analysis, although Summit consists of 4,608 compute nodes organized in 256 compute racks (plus 40 storage and 4 infrastructure racks), we assume a fully configured network of 324 racks of compute nodes only. As can be seen, the network incorporates 18 *core switches,* each being a 2-level fat-tree network in a box that consists of 54 ASICs (18 *spine* & 36 *leaf switches*). Each *core switch* has 648 ports that connect to the Top-of-Rack (*ToR*) switches. Each of the 324 racks incorporates 2 *ToRs* and 18 nodes, while each node is equipped with 1 NIC, 2 POWER9 CPUs and 6 NVIDIA GPUs, resulting in a total number of 5,832 NICs, 11,664 CPUs and 34,992 GPUs. Each NIC interconnects the 2 CPUs to the *ToRs*, while each CPU is connected to 3 GPUs. Given that each link operates at 100 Gb/s (4x25-Gb/s lanes), we get a total bisection BW of 1,166.4 Tb/s.

Fig. 2(b) presents the MOTION network architecture that has been designed by considering a 128-port, 400-Gbps/port, co-packaged-optics-enabled 51.2-Tb/s switch module as the basic building block. Targeting a fully configured fat-tree network, we designed a network architecture that consists of 64 *core switches,* each being a 2-level fat-tree network with 12 switch modules (4 *spine* & 8 *leaf modules)*. In order to allow for bottleneck-free connectivity, the connections between the 2 levels follow the channel bonding technique where 16 parallel physical links are grouped together to form logical connections. As a result, each MOTION *core switch box* has 512 ports that are connected to 512 *ToRs,* resulting in a total number of 32,768 available endpoints. For comparison fairness, the MOTION architecture uses the same number of compute nodes as the Summit-like system, which corresponds to 5.6 available network ports per node, or 11.2x higher BW per node. We note that in a realistic scenario we would assign 6 network ports to each node and end up with a system of 5,461 compute nodes, or 6.4% fewer than Summit. For almost the same number of nodes, the MOTION architecture enables a 11.2x higher bisection BW of 13,107.2 Tb/s with 21% fewer switch modules (1,280). While the MOTION architecture requires more sophisticated NICs that can provide 6 network, 6 GPU and 2 CPU ports, the combination of such NICs with higher-radix co-packaged-enabled switches allows for 3x more connections per node, or one connection per GPU versus 0.33 for the Summit-like case. Assuming the same 100-Gbps/port rate as in Summit, the MOTION architecture provides a 2.8x higher bisection BW of 3,276.8 Tb/s. The MOTION architecture offers a promising solution for enabling the direct network-attachment of GPUs to accommodate the increasing data movement demands of distributed workloads.

3. Simulation results

To investigate the performance of the proposed architecture we used the Venus discrete event simulator [11], which has been developed on top of the queue-based OMNeT++ simulation framework. For both Summit-like and MOTION cases, each network endpoint generates 100 Gb/s of traffic in the format of 1 KB packets, following the Bernoulli interarrival distribution. For MOTION, we also consider a link rate of 400 Gb/s that corresponds to the 51.2-Tb/s switch example of section 2. Both NICs and switches use the InfiniBand protocol and add a 100 ns delay to the traversing data (besides any queueing delays). For the buffer sizes, we assumed 128 KB per switch port, which translates to a maximum of 128 packets that can be queued per port. Regarding routing, we chose to simulate

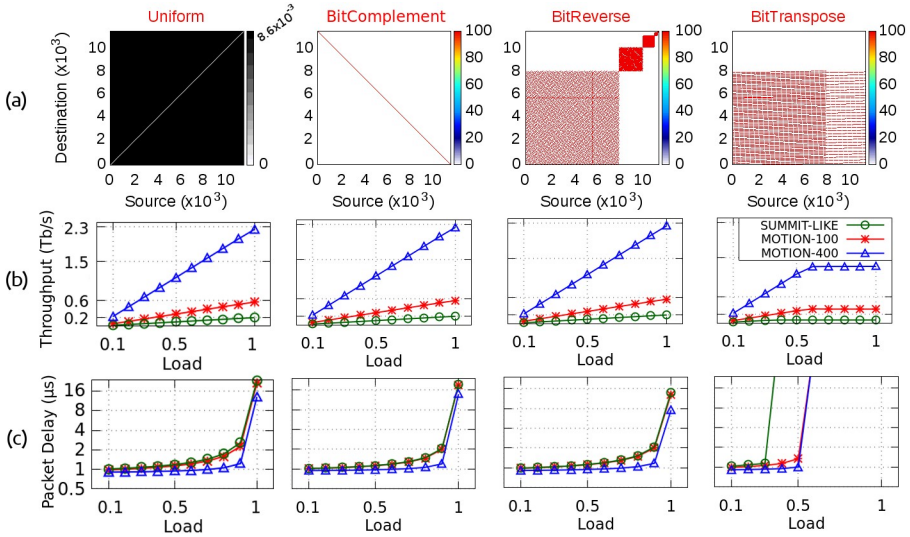

Fig. 3: (a) Traffic patterns, (b): Absolute throughput per node, (c): Mean packet delay.

the *Random* algorithm where, at each routing stage, the next hop is randomly selected over the set of all the shortest paths that lead to the destination (no adaptive routing was considered, which is the case for the real Summit system). This approach offers both load balancing and redundancy as it has been presented in practice in [12].

For our analysis, we considered four widely used synthetic traffic patterns [13], shown as heatmaps in Fig. 3(a) (for 11,664 network endpoints), combinations of which emulate the communication of many important HPC workloads. For *BitComplement*, *BitReverse* and *BitTranspose,* each source sends 100% of its traffic to a certain destination (certain destinations may receive traffic from multiple sources while others from none, e.g. *BitTranspose*). For the *Uniform* case, each source endpoint uniformly distributes all its traffic to the rest endpoints. Fig. 3(b) presents the throughput vs load comparison between the Summit-like and MOTION networks, where we plotted the absolute throughput per node that was measured throughout the complete simulation time. For all traffic patterns except *BitTranspose*, we observe that all architectures present a linear throughput increase and manage to deliver the highest possible throughput. As can be seen, the higher number of connections per node in the MOTION architecture leads to 2.8x and 11.2x higher throughput for the MOTION-100 and MOTION-400 cases over the Summit-like system, respectively. For *BitTranspose,* which stresses both architectures (there are fewer destination than source nodes), all cases present linear throughput increase until reaching their saturation points, which are at 75 Gb/s for Summit-like, 322 Gb/s for MOTION-100 and 1,295 Gb/s for MOTION-400. Finally, Fig. 3(c) presents the mean-packet-delay vs load curves, which agree with the corresponding throughput results. All architectures offer bounded packet delay in the 0.7-2-µs range before reaching their saturation points. As expected, the 400-Gb/s MOTION scenario offers the best mean packet delay for most cases.

4. Conclusion

We proposed an HPC network architecture that makes use of co-packaged optics at the switch modules. The higher-radix switches combined with higher link rates enable the implementation of a topology with 3x more endpoints. The simulation results show up to 11.2x higher throughput for representative benchmarks, keeping up with the increasing demands of distributed workloads and opening the way to direct-network-attached accelerators.

Acknowledgment

The information, data, or work presented herein was funded in part by the Advanced Research Projects Agency-Energy (ARPA-E), U.S. Department of Energy, under Award Number DE-AR0000846. The views and opinions of authors expressed herein do not necessarily state or reflect those of the U.S. Government or any agency thereof.

References

[1] K. Tu, “Identifying Network Data Transfer Bottlenecks in HPC Systems”, *Supercomputing Conference*, Dallas, TX, 11-16 Nov. 2018
[2] https://asc.llnl.gov/CORAL-benchmarks/#umt2013
[3] M. Cho et al., “Large model support for deep learning in caffe and chainer”, SysML, 2018.
[4] C. Pinto et al., “Hoard: A distributed data caching system to accelerate deep learning training on the cloud”, arXiv:1812.00669, 2018.
[5] CPO JDF: Co-Packaged Optical Module Discussion Document, https://www.facebook.com/CoPackagedOpticsCollaboration/
[6] C. Minkenberg et al., “Network architecture in the Era of integrated optics.” IEEE/OSA JOCN 11.1 (2019): A72-A83.
[7] S. Chun et al.,“IBM POWER9 package technology and design.” IBM J. R&D, 62.4/5 (2018).
[8] W. D. Becker et al. “Electronic packaging of the IBM z13 processor drawer.” IBM J. R&D, 59.4/5 (2015).
[9] https://fuse.wikichip.org/news/1351/ornls-200-petaflops-summit-supercomputer-has-arrived-to-become-worlds-fastest/
[10]D. Kuchta, *et. al.*, "Multi-wavelength Optical Transceivers Integrated on Node (MOTION)," *OFC 2019*, paper M4D.6.
[11]R. Birke, *et al.*, “Towards massively parallel simulations of massively parallel HPC systems,” *SIMUTOOLS*, Italy, Mar. 2012, pp. 291-298.
[12]A. Greenberg, *et al.*, “VL2: a scalable and flexible data center network,” *SIGCOMM 2009,* ACM, New York, NY, USA, 51-62.
[13]R. Rojas-Cessa, Interconnections for computer communications and packet networks, CRC Press, January 6, 2017.