Improving Data Center Network Locality w/ Co-packaged Optics

Pavlos Maniotis*, Laurent Schares, Daniel M. Kuchta, Bengi Karacali

IBM T. J. Watson Research Center, Yorktown Heights, New York, USA, *ppmaniotis@ibm.com

Abstract: Co-packaged optics can enable switches with unprecedented speeds of 51.2 Tb/s and beyond. This translates to networks with 4x higher bisection bandwidth, >40% fewer switches, and substantially improved network locality, i.e., large-scale applications can be placed under up to 50% fewer 1st-level switches.

Introduction

The machine learning explosion led to an unprecedented increase in both the number of model parameters[1] and the size of the required data sets[1],[2]. As such, parallel execution has become the common practise, often expanding beyond a single rack. However, many parallel applications require their tasks to act in synchrony, mandating in this way the need for high-speed interconnects among the processors. To fulfil these demands, a 40x increase in the switch Input/Output bandwidth (I/O BW) took place over the past decade (25.6-Tb/s ASICs are today's state of the art[3]). However, data-hungry applications keep pushing the limits of BW, latency, and energy-efficiency[4].

As a result, BW scaling remains an always open topic. To overcome BW density and thermal cooling limits, more energy-efficient and dense solutions are required. In this context, one of the potential solutions is to integrate the optics onto the 1st-level package, a.k.a., co-packaged optics (CO)[5]-[13]. CO combine two key advantages: they can substantially increase the package escape BW, and they can minimize the power for driving optics if the economics can be worked out. Today, pluggable optics are typically located several inches from the switch ASIC. In contrast, CO are placed next to the ASIC, which can unlock significant energy gains since much shorter wires with significantly lower loss are required.

This paper reports on the recent activities within the MOTION research project (Multi-wavelength Optical Transceivers Integrated On Node)[5]-[7]. MOTION aims to develop a CO module for chip-packaging applications. First, we discuss on the hardware developments; the module consists of Vertical Surface Emitting Laser (VCSEL) arrays, surface-illuminated photodiodes, and driver/receiver Integrated Circuits (ICs) flip-chip attached on a glass carrier. Next, we study the benefits of using CO in data center (DC)/HPC networks. The study suggests that—for a network size of >12K end points—the higher-BW and higher-radix switches enabled by CO can offer a 4x bisection BW increase, which is combined with a switch count reduction of 41%. In addition, simulations with virtual-machine traces suggest that CO can enable greatly improved network locality, i.e., large-scale applications can be placed under up to 50% fewer 1st-level switches.

Considering that the HPC cloud area is continuously expanding (spending is projected to grow at a 17.6% compound annual growth rate until 2024[14]), CO form a promising solution for keeping up with BW scaling in DC/HPC networks.

MOTION co-packaged optics module

Fig. 1 shows the 1st-generation module along with an indicative eye diagram for 1 of the 16 channels at an NRZ modulation format. The module includes VCSEL arrays and surface-illuminated photodiodes that operate at 56 Gb/s, as well as driver and receiver ICs, that are flip-chip attached onto a 13x13-mm^2 glass carrier. The VCSELs & photodiodes operate in the 930-950-nm range, and the light path is through the glass carrier. The total I/O BW is 0.9 Tb/s, and the BW density is 5.3 Gb/s/mm^2. The target for the 2nd-generation module is to incorporate 32 channels in the same area, where each of them will operate at 112 Gb/s with a PAM4 modulation format. The result will be a total I/O BW of 3.58 Tb/s and a BW density of 21.2 Gb/s/mm^2. More details on the design and packaging can be found in[5],[6].

Baseline & MOTION-enabled networks

Fig. 2(a) shows the baseline network of our analysis. The network has an oversubscription ratio of 3:1[15] and uses 6.4-Tb/s and 25.6-Tb/s switches

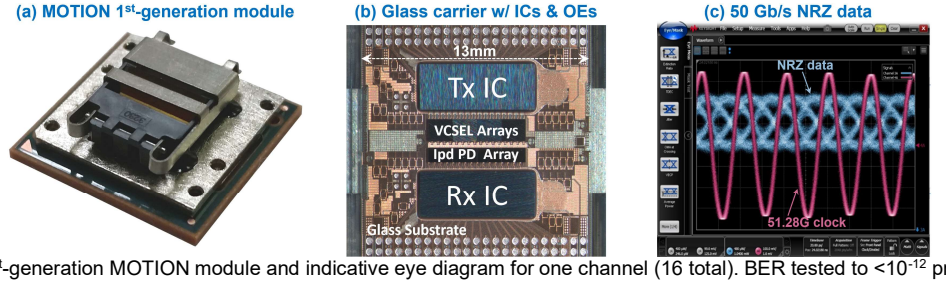

Fig. 1[7]: 1st-generation MOTION module and indicative eye diagram for one channel (16 total). BER tested to <10^{-12} pre-FEC.

978-1-6654-3868-1/21/$31.00 ©2021 IEEE

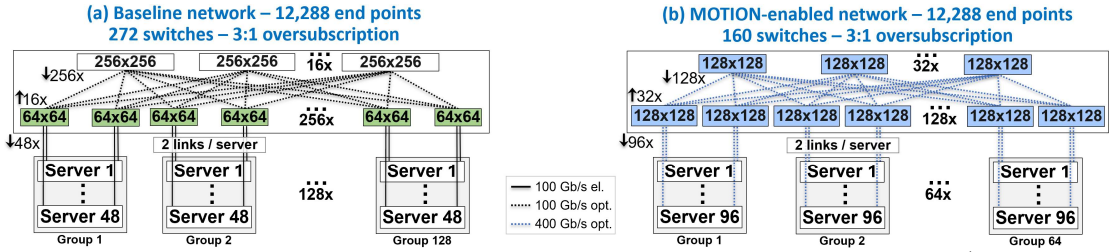

Fig. 2: (a) Baseline network, (b) MOTION-enabled network with 4x bandwidth per server and 2x servers per 1st-level switch.

switches (100 Gb/s/port) at the 1st and 2nd switch layers, respectively. The servers connect to the network by using direct attach copper cables, while the 2 switch layers are interconnected with active optical cables. The system consists of 272 switch ASICs and 12,288 end points. For network redundancy, the servers connect to 2 end points from 2 distinct switches, resulting in 6,144 servers in total. The top switch layer has sixteen 256-port spine switches that connect to 256 1st-level switches. Every two 1st-level switches connect to a *group* of 48 servers, resulting in 128 groups in total (a group of servers can physically expand over multiple racks, e.g., 2 racks of 24 servers each). The intra-group and inter-group communication costs are 1 and 3 hops, respectively. The system's bisection BW is 409.6 Tb/s.

Fig. 2(b) shows the proposed architecture, which assumes a 128-port 51.2-Tb/s switch (400 Gb/s/port), and, again, an oversubscription ratio of 3:1. According to our previous analysis[7], a 51.2-Tb/s switch with only optics for I/O could be built on a 90x90-mm^2 carrier with 16 CO modules. This solution would maximize the energy gains since it would eliminate the need for driving pluggable modules. A more conservative approach could combine 25.6-Tb/s electrical I/O from the bottom of the package with 25.6 Tb/s optical I/O from the top. Such a solution could be built on a 70x70-mm^2 carrier with 8 CO modules[7].

Opposed to the baseline case, the proposed network uses a single type of switch and only optics for connectivity, i.e., servers are connected to the 1st-level switches optically. The higher reach of optics, combined with the higher switch radixes enabled by CO, enables a 2x increase in the number of servers that connect to every 1st-level switch. This results to a total number of 64 groups of 96 servers each, which is also combined with a 4x increase in the server BW. This can be greatly beneficial for large-scale applications since communication between every 96 servers (e.g., >4.5K cores for 48-core servers) requires only 1 hop and can be realized at 4x higher data rates. Moreover, the proposed approach requires 112 fewer switches, resulting in a switch count reduction of 41%. This translates to both reduced cost and power consumption and less management overhead. Finally, the higher data rates result in a 4x higher bisection BW of 1,638.4 Tb/s.

Simulation analysis

To assess the performance of the systems under test we extended CloudSim Plus[16], an open source simulator for cloud infrastructures and services. For both systems we considered 48-core servers with 384 GiB of memory, while for the network BW we considered 100 Gb/s for the baseline case and 400 Gb/s for MOTION case.

For our study we extracted a virtual-machine (VM) trace from the publicly available *AzureTracesForPacking2020* dataset[17],[18]. The dataset covers a 14-day period, while we considered requests that arrived within the first 7 days. The VM end times can extend beyond the 14th day and they are capped at 90 days to anonymize time[18]. Since we focus on large-scale applications and the network-locality properties of the tested systems, we selected VM group requests that require at least 48 servers to host them (i.e., equal to the size of a group of servers that connect to the same 1st-level switches for the baseline case). We consider that a set of VMs corresponds to a large-scale application if all the following conditions are met: (a) the VMs share the same *tenant* and *vmType* ids, (b) they have the same *priority*, and (c) they start and end at the same times with a max difference of 1 sec. Regarding the *vmTypes,* the dataset associates them with multiple resource request ratios depending on different type of servers, also anonymized[18]. To stress both systems, we used ratios that maximize the requested number of cores. For the above criteria, the trace consists of 631 group requests that correspond to >62.5K VMs. Tab. 1 shows the VM types sorted by popularity, while the [min / avg / max / stddev] values of the request lifetimes and interarrival times equal [3s / 1.49d / 89.9d / 8.17d] & [0s / 14.1m / 1.2d / 1.27h].

Fig. 3(a) shows the distribution of the requests according to their size expressed in number of

Tab. 1: VM types appearing in the trace sorted by popularity

# of VMs (%)	Cores	Memory (GiB)	Network BW (Gb/s)	
			Baseline	MOTION
49.8	32	64	20	80
41.7	16	32	1.25	5
5.8	32	64	1.25	5
1.1	32	224	20	80
0.7	32	256	40	160
0.7	32	112	20	80
0.2	16	128	20	80

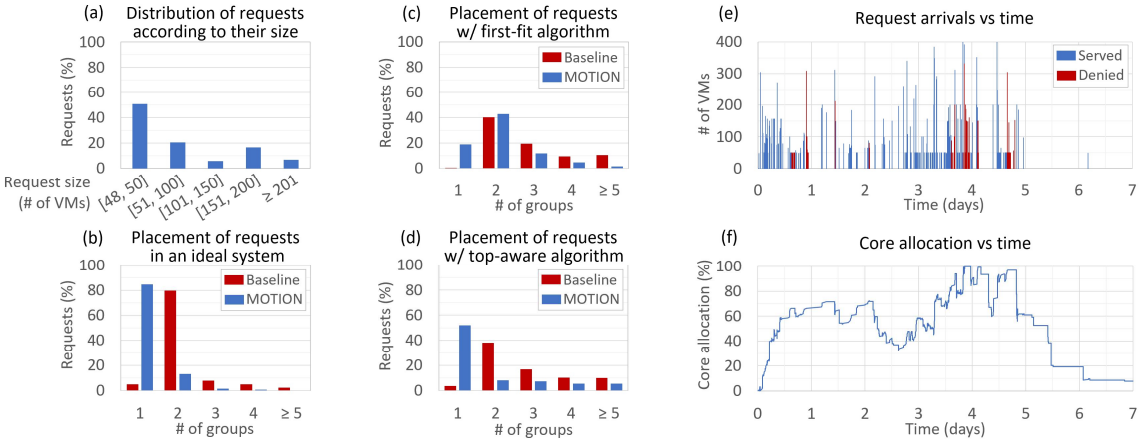

Fig. 3: (a) Distribution of requests according to their size, (b) Placement of requests in an ideal system, i.e., assuming that (i) every request can be placed in the min number of groups and (ii) there are always available resources, (c)-(d) Placement w/ first-fit and top-aware algorithms showing the improved network locality for the MOTION system, (e)-(f) Request arrivals and core allocation vs time for MOTION system w/ top-aware algorithm: 100% core allocation or system fragmentation can lead to request denials.

VMs. 50.7% of the requests belongs to the *[48,50]* category, while *[51,100]* is the 2nd most popular category (20.4%). The biggest request consists of 400 16-core VMs, while the biggest request in terms of cores consists of 312 32-core VMs.

Fig. 3 (b) assumes an ideal system and shows how the requests can be placed by assuming no contention for the resources, i.e., assuming that (a) every request can be placed in the min number of groups, and (b) there are always available resources. For the baseline case, 4.9% of the requests fits in 1 group, 79.7% requires 2 groups, while 7.9%, 5.1% and 2.4% require 3, 4 and 5-7 groups, respectively. For MOTION case, 84.6% of the requests fits in 1 group, while 13%, 1.6%, and 0.8% require 2, 3 and 4 groups, respectively. Placing the VMs of a request in a single group comes with 2 key advantages: (a) communication cost is 1 hop max, and (b) network contention associated with crossing the spine is eliminated.

To study the impact of contention for resources, we simulated two placement algorithms. Initially, we considered first-fit[19],[20], which is well-known for its simplicity; servers are scanned in serial order until a suitable server is found. If no such server exists, the request is denied. Since we simulate group requests, a request is denied if no resources exist for all VMs cumulatively. Fig. 3(c) shows the simulation results. For both systems, 20.1% of the requests (not shown in Fig. 3(c)) is denied due to lack of resources or system fragmentation that turns the resources unreachable. For the baseline case, 0.5% of the requests is placed in 1 group, 40.1% in 2 groups, while 19.3%, 9.5% and 10.5% in 3, 4 and 5-10 groups, respectively. For MOTION case, 19% of the requests is placed in 1 group, while 42.8%, 11.7%, 4.8% and 1.6% is placed in 2, 3, 4 and 5-7 groups.

Next, we considered a topology-aware algorithm *(top-aware)* that follows part of the principles of the scheduling policy presented in[19], which targets to reduce network sharing/fragmentation in HPC systems. Small requests that can fit in a group populate the system from the first group onwards, whereas big requests that require multiple groups populate the system from the last group backwards. Whenever possible, the algorithm ensures that small requests are placed in a single group. This is expected to favour the MOTION case since 84.6% of all requests can ideally fit in 1 group (4.9% for the baseline case). Fig. 3(d) shows the respective simulation results. As with first-fit, slightly above 20% of the requests is denied (not shown in Fig. 3(d)). For the baseline case, no significant differences are observed vs first-fit, although the number of requests placed in 1 group increases from 0.5% to 3.8%. On the other hand, the MOTION system presents remarkable improvements: 52% (vs 19%) of the requests is placed in 1 group, while 8.1%, 7.4%, 5.4% and 5.6% of the requests is placed in 2, 3, 4 and 5-10 groups, respectively. It should be noted that most denied requests, i.e., 17.3% of all requests, can ideally be placed in 1 group.

Finally, Fig. 3(e) and (f) focus on the MOTION case w/ top-aware and show the request arrivals along with the respective core allocation vs time. E.g., at the end of the 4th day, requests are denied because core allocation reaches 100%, while at the end of the 1st day, requests are denied due to system fragmentation, i.e., most servers have 16 free cores, prohibiting the system from accepting additional 32-core VMs. Similar behaviour was observed for the rest configurations as well.

Conclusions

We investigated the advantages of using CO in DC/HPC networks. CO can enable networks with 4x higher bisection BW and 41% fewer switches. Simulations with VM traces suggest that CO can significantly improve network locality, i.e., large-scale applications can be placed under up to 50% fewer 1st-level switches.

Acknowledgements

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] M. Shoeybi, M. Patwary, R. Puri, P. LeGresley, J. Casper, B. Catanzaro, “Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism”, ArXiv, abs/1909.08053, Mar. 2020

[2] Y. Weng, A. Kumar, M. B. Saleem and B. Zhang, "Big Data and Deep Learning for Terabyte-Scale Renewable Datasets", 2018 Power Systems Computation Conference (PSCC), 2018, pp. 1-7

[3] https://www.nextplatform.com/2019/12/12/broadcom-launches-another-tomahawk-into-the-datacenter/

[4] J. Shalf, "HPC Interconnects at the End of Moore's Law," OFC2019, San Diego, CA, USA, 2019, pp. 1-3.

[5] D. M. Kuchta, et al., "Multi-Wavelength Optical Transceivers Integrated on Node (MOTION)," OFC2019, San Diego, CA, USA, 2019, pp. 1-3.

[6] P. Maniotis, et al., "Toward lower-diameter large-scale HPC and data center networks with co-packaged optics," in IEEE/OSA JOCN, vol. 13, no. 1, pp. A67-A77, January 2021

[7] P. Maniotis, et.al., "Co-packaged optics for HPC and data center networks," Proc. SPIE 11692, Optical Interconnects XXI, 1169205 (5 March 2021)

[8] https://www.facebook.com/CoPackagedOpticsCollaboration/

[9] M. Wade et al., "TeraPHY: A Chiplet Technology for Low-Power, High-Bandwidth In-Package Optical I/O," in IEEE Micro, vol. 40, no. 2, pp. 63-71, 1 March-April 2020

[10] S. Fathololoumi et al., "1.6 Tbps Silicon Photonics Integrated Circuit and 800 Gbps Photonic Engine for Switch Co-Packaging Demonstration," in Journal of Lightwave Technology, vol. 39, no. 4, pp. 1155-1161, Feb. 15, 2021

[11] https://rockleyphotonics.com/wp-content/uploads/2020/03/Rockley-Photonics-OptoASIC-Flyer.pdf

[12] https://acacia-inc.com/wp-content/uploads/2019/06/Optinet-China-2019_Acacia_Fenghai-Liu_UpLoad_v1.pdf

[13] https://www.sdxcentral.com/articles/news/broadcom-unveils-first-co-packaged-optical-switch/2021/01/

[14] https://hyperionresearch.com/wp-content/uploads/2021/01/Hyperion-Research-Special-Analysis-Clouds-and-HPC-December-2020.pdf

[15] https://www.arista.com/en/solutions/software-defined-cloud-networking-solutions#Spine-Leaf-Network

[16] M. C. Silva Filho, R. L. Oliveira, C. C. Monteiro, P. R. M. Inácio and M. M. Freire, "CloudSim Plus: A cloud computing simulation framework pursuing software engineering principles for improved modularity, extensibility and correctness," 2017 IFIP/IEEE Symposium on Integrated Network and Service Management (IM), 2017, pp. 400-406

[17] O. Hadary, et al., "Protean: VM Allocation Service at Scale", in Proceedings of the 14th USENIX Symposium on Operating Systems Design and Implementation (OSDI 2020). USENIX Association, November 2020.

[18] https://github.com/Azure/AzurePublicDataset/blob/master/AzureTracesForPacking2020.md

[19] A. Jokanovic, et al., "Quiet Neighborhoods: Key to Protect Job Performance Predictability," 2015 IEEE International Parallel and Distributed Processing Symposium, 2015, pp. 449-459

[20] https://en.wikipedia.org/wiki/Bin_packing_problem#First-Fit_(FF)