How Data Center Networks Can Improve Through Co-packaged Optics

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

Abstract: Building higher-radix switches with co-packaged optics can help improving network locality in data center networks. Simulations show completion time reductions of up to 40% for an all-to-all communication pattern and a network stack latency of 1.25 µs. © 2022 The Author(s)

1. Introduction

The recent explosion of applications for Artificial Intelligence (AI) as well as the increased interest in foundation models [1], [2] has led to an unprecedented increase in both the size of the required data sets and the number of model parameters. Highly parallel execution has become the standard approach for AI training, which is often distributed over many servers located in different racks. While data center switches have provided an 80x increase in Input/Output bandwidth (I/O BW) over the past dozen years (today's state-of-the-art switches deliver speeds of up to 51.2 Tb/s [3]), modern applications with high communication requirements keep pushing the limits of network bandwidth, latency, and energy efficiency [4]. To address these demands, energy-efficient and dense I/O solutions remain a highly active topic for both industry and academia. Amongst the potential solutions is the integration of optical transceivers directly within chip-scale modules, a.k.a., co-packaged optics [5]-[9]. This solution can increase the package escape BW by offering an extra dimension for wiring chip pins. It can also minimize the power for driving the optics, since placing the optics in proximity of the switch ASIC (Application-Specific Integrated Circuit) eliminates the need for lossy wires to drive pluggable modules that are placed at the edge of the motherboard.

In this paper, we extend our previous work [6] and report on the recent network modelling and simulation activities within the framework of the MOTION research project (Multi-wavelength Optical Transceivers Integrated On Node) [7]. In section 2 we present a comparison between a baseline network and a network that makes use of MOTION-enabled switches. For the baseline network we consider off-the-shelf state-of-the-art switches, while for the proposed network we assume switches with 128 ports running at 400 Gb/s, i.e., 51.2 Tb/s I/O BW in total. For a data center network of >12K end points, the comparison shows that the higher-bandwidth and higher-radix switches enabled by co-packaged optics can offer a 4x bisection BW increase, and a switch count reduction of 41%. At the same time, the network locality properties of the system are significantly improved since large-scale applications can be placed under up to 50% fewer 1st-level switches; this is shown in the trace-based analysis of our previous work [6] that uses virtual-machine (VM) traces from a production data center. In addition, more than half of the large-scale applications can be placed under a single 1st-level switch, compared to fewer than 5% for the baseline case. Placing the VMs of an application under a single switch comes with two key advantages: (a) communication cost is one hop max, and (b) network contention associated with crossing the network spine is eliminated. In this work, we extend these findings, and we compare the overhead of spreading the VMs of the applications under multiple 1st-level switches to placing them under a single switch. Our simulation results suggest that the network stack (i.e., network virtualization method and/or networking technology), matters: the faster the network stack, the more important is to place the VMs under a single switch; the overhead for placement under two or more switches can reach up to 68% higher completion times for an all-to-all communication pattern and a network stack latency of 1.25 µs, i.e., the completion time reduction for single-switch vs multi-switch placements can reach up to 40%.

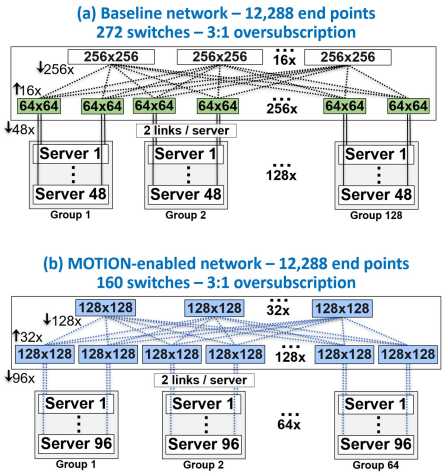
2. Baseline and MOTION-enabled networks

Table 1 summarizes the specifications for both the first and second generations of the MOTION co-packaged optics module. The target for the 2nd-generation is to incorporate 32 optical channels in an area of 13x13 mm^2, each of them operating at a data rate of 112 Gb/s with a Pulse Amplitude Modulation 4-level (PAM4) format. This corresponds to a total I/O BW of 3.58 Tb/s with a BW density of 21.2 $Gb/s/mm^2$. The transmitter portion of the module uses 2:1 laser sparing for improved reliability, meaning that each channel has one primary laser and one spare. The primary to spare laser switching time is <100 ns [7]. This can significantly reduce the downtimes since human intervention is not required for primary laser failures. We refer to [7] and [8] for details about the physical-layer design and packaging as well as for figures of the MOTION module.

Fig. 1 compares the baseline and the MOTION-enabled networks. Fig. 1 (a) shows the baseline network that uses state-of-the-art switches and has an oversubscription ratio of 3:1. The system uses 25.6-Tb/s switches at the second

Table 1: MOTION specifications

Parameter	1st gen.	2nd gen.
Data rate per channel	56 Gb/s	112 Gb/s
Format	NRZ	PAM4
# of channels (full duplex)	16	32
Total BW per module	0.9 Tb/s	3.58 Tb/s
Energy consumption	< 4 pJ/bit	< 2 pJ/bit
BW density	5.3 Gb/s/mm^2	21.2 Gb/s/mm^2
Dimensions	13 x 13 x 4 mm	
Temperature	0-70°C	
Electrical interface	Extra Short Reach (XSR/6 dB budget)	
Optical margin	2 dB, >30m OM4, OM5	



MOTION-enabled vs Baseline network

- **Same number of hops**
- **41% fewer switches**
  - Reduced cost / power consumption
  - Less administration / management overhead
- **2x more servers per 1st-level switch**
  - Improved network locality
  - E.g., 1 hop max for >4.5K cores (assuming 48-core servers)
- **4x higher bisection bandwidth**
  - 400 vs 100 Gb/s link rate → 1,638.4 vs 409.6 Tb/s
- **Fast laser sparing of MOTION module**
  - Laser switching in <100 ns
  - Reduced downtimes due to failed transmitters

Fig. 1: Comparison between the baseline and MOTION-enabled networks.

(spine) switch layer and 6.4 Tb/s switches at the first (leaf) switch layer. The link rate is 100 Gb/s. Due to the higher distances between the switch layers, it is the standard approach to use Active Optical Cables (AOCs). For cost-saving reasons, Direct Attach Copper (DAC) cables are used between the 1st-level switches and the servers, which, however, constrains the placement of the equipment in proximity to the switches, hence the use of 6.4-Tb/s leaf switches. The system has 272 switches and the servers connect to two end points from two different switches for network redundancy. Every two leaf switches connect to a group of 48 servers, resulting in 128 groups in total (each group can expand over multiple physical racks)

Fig. 1 (b) shows the MOTION-enabled network that uses the same radix at both the leaf and spine switches and only optics for connectivity. Again, an oversubscription ratio of 3:1 is applied. The system uses 51.2-Tb/s switches that incorporate co-packaged optics for I/O. While co-packaged optics (or possibly electrical I/O off the top of the carrier) may become necessary for future switch generations, the technology already holds promise for top-side-only I/O in the 51.2-Tb/s generation due to the energy/cost gains [5], [9]. A 51.2-Tb/s switch could be built with 16 2nd-generation MOTION modules. Compared to the baseline case, the MOTION-enabled system has 2x more servers per 1st-level switch, i.e., 64 groups of 96 servers each, which is combined with a 4x increase in the server BW. This is greatly beneficial for larger-scale applications since communication between every 96 servers (e.g., >4.5K cores assuming 48-core servers) requires a maximum of 1 hop and can be realized at 4x higher data rates. Moreover, the switch count is reduced by 41%, which translates to both reduced cost and less network administration/management overhead. The bisection BW is increased by a factor of 4x due to the higher data rates. Regarding the power consumption, co-packaged optics can lead to reductions of up to 50% over pluggable optics [9], while at the same time an up to 50% system-level reduction in the cost per capacity appears to be feasible [9].

3. Simulation results

We use the Venus discrete-event network simulator [10] to quantify the performance improvements of building networks with enhanced network locality properties. In our analysis we simulate an all-to-all communication pattern for 24 VMs. This pattern emulates the function of collective operations like the Message Passing Interface (MPI) all-to-all, MPI all-gather, and MPI all-reduce, which are widespread amongst parallel applications like AI training. The simulated pattern follows the optimized exchange pattern presented in [11]. In this implementation, exactly half of the messages cross the spine of the network in every successive phase, which allows for contention-free communication in cases where at least half of the full bisection BW is available. However, since both simulated systems have an oversubscription ratio of 3:1, there is always some form of contention due to the messages that cross the spine. In addition, the simulated implementation does not consider any explicit synchronization or separation of the successive all-to-all phases.

Regarding the message size, we iterate over the [2^8, 2^{15}] range in bytes, i.e., 256 bytes to 32 KiB. The packet size of the systems varies between 2^8-2^{12} bytes (or 256 bytes to 4 KiB), which means that messages of more than 2^{12} bytes are carried by multiple contiguous packets. For both Network Interface Cards (NICs) and switches we use a credit-based flow control mechanism, i.e., we simulate lossless networks. The switches have an input-buffer organization (128 KiB per port) with Virtual Output Queueing (VOQ), and they add a 1.5 µs delay to the traversing data (besides any potential queueing delays). We simulate an ECMP-like (Equal-Cost Multi-Path) routing algorithm where, at each routing stage, the next hop is randomly selected over the set of all shortest paths that lead to the

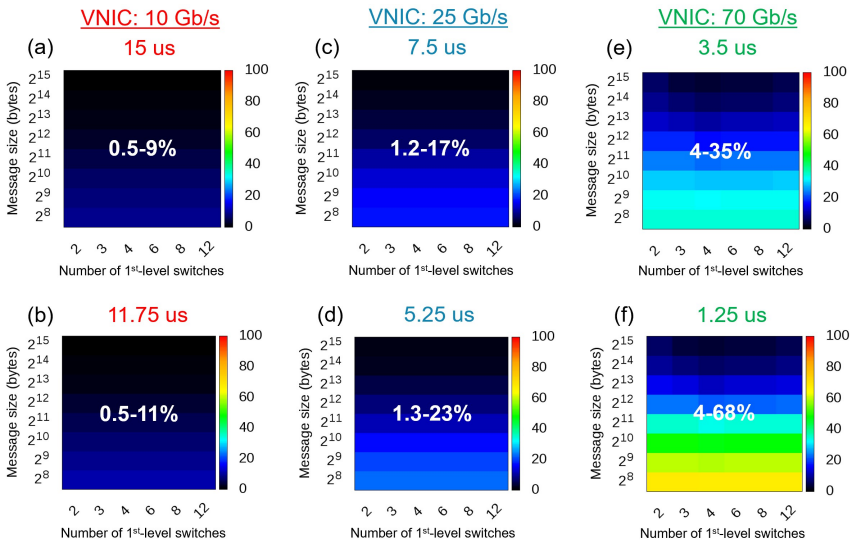


Fig. 2: All-to-all completion time percentage increase of spreading the VMs of an application under multiple 1st-level switches vs placing them under a single switch. Results correspond to different VNIC speeds and network stack latencies. The faster the network stack, the higher the overhead. Smaller message sizes are affected most due the relatively high overhead from the extra hops when compared to the message duration.

destination. According to this approach, contiguous packets from the same message follow the same path. Regarding the Virtual NICs (VNICs) of the simulated VMs, we consider three throttling speeds of 10, 25 and 70 Gb/s, which correspond to a range of bandwidths that are representative of typical cloud VM profiles. For each throttling speed, we consider two network stack latencies that correspond to different network device virtualization methods and/or networking technologies. Advanced network stacks typically provide higher throughput, i.e., higher packet per second (PPS) performance, and lower latencies.

Fig. 2 shows the simulation results for all the configurations. Each heatmap corresponds to a different network configuration and shows the completion time percentage increase of spreading the VMs under multiple 1st-level switches versus placing them under a single switch. The results suggest that the faster the network stack, the more important it is to place the VMs of an application in proximity. Smaller message sizes are affected most due the relatively high overhead from the extra hops when compared to the message duration. For slower network stack latencies of 11.75-15 µs, the overhead of spreading the VMs can reach up to 11%. For faster stacks of 5.25-7.5 µs the overhead can reach up to 23%, while for stacks of 1.25-3.5 µs the overhead can reach up to 68%. Therefore, building networks with improved network locality properties can enable substantial speedups, which are even more pronounced for faster network configurations.

4. Conclusion

We studied the advantages of building data center networks with improved network locality by means of higher-radix switches that use co-packaged optics. The simulation results show completion time reductions of up to 40% for an all-to-all communication pattern and a network stack latency of 1.25 µs.

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] https://research.ibm.com/blog/what-are-foundation-models
[2] R. Bommasani, et al., “On the opportunities and risks of foundation models,” arXiv preprint arXiv:2108.07258. Aug 16, 2021
[3] https://www.broadcom.com/products/ethernet-connectivity/switching/strataxgs/bcm78900-series
[4] P. Lu, et al., “A Survey of High-Performance Interconnection Networks in HPC Systems,” Electronics 2022, 11, 1369
[5] K. Muth, et al., “Key Technology Enablers for Co-Packaged Optics,” ECOC, Tu1F.1, Sep 2022
[6] P. Maniotis, et al., “Improving Data Center Network Locality with Co-packaged Optics,” ECOC, We1F.1, Sep 2021
[7] D. Kuchta, et al., “An 800 Gb/s, 16 Channel, VCSEL-Based, Co-packaged Transceiver With Fast Laser Sparing”, ECOC, Tu1F.2, Sep 2022
[8] P. Maniotis, et al., "Toward lower-diameter large-scale HPC and DC networks with co-packaged optics," JOCN, vol. 13, no. 1, January 2021
[9] C. Minkenberg, et al., “Co-packaged datacenter optics: Opportunities and challenges,” IET Optoelectronics, 15: 77-91, 2021
[10] R. Birke, et al., “Towards massively parallel simulations of massively parallel HPC systems,” SIMUTOOLS, Brussels, 291–298, 2012
[11] B. Prisacari, et al., “Bandwidth-optimal all-to-all exchanges in fat tree networks,” ICS '13, New York, NY, USA, 139–148, 2013