Toward higher-radix switches with co-packaged optics for improved network locality in data center and HPC networks [Invited]

PAVLOS MANIOTIS,* LAURENT SCHARES, DANIEL M. KUCHTA, BENGI KARACALI

IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA.
**Corresponding author: ppmaniotis@ibm.com*

Received XX Month XXXX; revised XX Month, XXXX; accepted XX Month XXXX; posted XX Month XXXX (Doc. ID XXXXX); published XX Month XXXX

We study the network locality improvements that can be achieved by using co-packaged optics in data center and high-performance computing networks. The increased escape bandwidth offered by co-packaged optics can enable switches with speeds of 51.2 Tb/s and beyond. From a network architecture perspective, the key advantages of introducing co-packaged optics at the switch points include: the implementation of large-scale topologies of >12K end points with 4x higher bisection bandwidth, and the reduction of the required number of switches by >40% compared to state-of-the-art approaches. From a network operation perspective, both improved network locality and faster operation can be achieved since the higher-radix switches can mitigate the impact of network contention. Placing applications under fewer leaf switches reduces the number of packets that cross the spine switches in a leaf-spine topology. The proposed scheme is evaluated via discrete-event simulations: we initially evaluate the network locality properties of the system by using virtual-machine traces from a production data center, and we subsequently quantify the performance improvements by simulating an all-to-all pattern for a variety of message sizes over a number of nodes. The results suggest that co-packaged optics form a promising solution for keeping up with bandwidth scaling in future networks. The virtual-machine analysis shows that large-scale applications can be placed under up to 50% fewer 1st-level switches, while the network analysis shows speedups of up to 7.1, which translates to execution time reductions of up to 26% and 42.7% for applications with communication ratios of 0.3 and 0.5, respectively.

http://dx.doi.org/10.1364/JOCN.99.099999

1. INTRODUCTION

The recent machine learning explosion has led to an unprecedented increase in both the number of model parameters [1] and the size of the required data sets that are used for training complex artificial-intelligence models [1], [2]. As a result, highly parallel execution has become the common practice, often expanding beyond the boundaries of a single compute rack. At the same time, many parallel applications require their tasks to act in synchrony, mandating the need for high-speed networks to interconnect the nodes in computing systems. To fulfil these demands, a 40x increase in the switch Input/Output band-width (I/O BW) took place over the past decade, with today's state-of-the-art switches delivering speeds of up to 25.6 Tb/s [3]. This increase has enabled a wide range of switch configurations where, depending on the application needs, different combinations of switch radices and data rates/port are being offered. However, despite these impressive advancements, modern applications with high communication requirements [4] keep pushing the limits of network bandwidth, latency and energy efficiency [5], [6].

To cope with these demands, BW scaling remains a continuously active topic for both the industry and academic communities. As we move toward the 51.2-Tb/s switch generation and beyond, overcoming BW density and thermal cooling limits will require more energy-efficient and dense I/O solutions. One of the potential solutions is to integrate the optical transceivers directly onto the top surface of the 1st-level package, a.k.a., co-packaged optics [7]-[17]. Co-packaged optics come with two key advantages: (a) they can substantially increase the package escape BW by offering an extra dimension for wiring additional chip pins, and (b) they can minimize the power for driving the optics. In today's switch implementations, pluggable optics are typically located 10-24 inches from the switch Application-Specific Integrated Circuit (ASIC), i.e., at the edge of the motherboard. In contrast to this, co-packaged optics are placed on top of a first- or second-level package within a switch module, in close proximity of the switch ASIC. This has the potential to unlock significant energy savings since much shorter wires with significantly lower channel loss are required. On top of that, this approach has the potential to overcome the limitations associated with the Ball Grid Arrays (BGA) and Land Grid Arrays (LGA), for which the pin count is typically at least 4x lower than the corresponding ASIC pin count [5], [18], [19]. The industry has started some initial efforts to standardize co-packaged optics. In addition to the co-packaging specifications in [12], the Optical Internetworking Forum (OIF) has

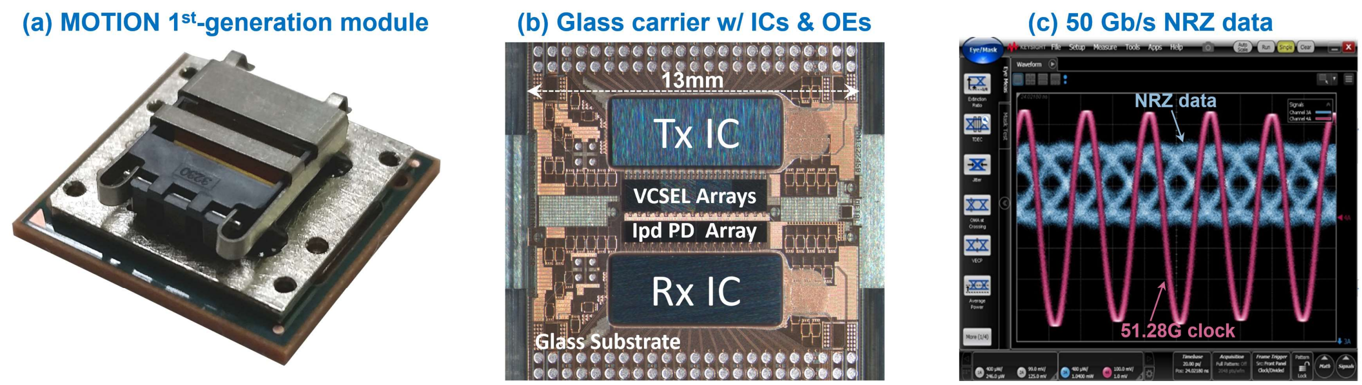

Fig. 1 [10]. 1st-generation MOTION module and indicative eye diagram for one channel (16 total). Bit error rate (BER) tested to $<10^{-12}$ pre-FEC.

begun work on a Co-Packaging Framework Implementation Agreement [20]. However, both of these efforts are aimed at longer distance links, 500-2000m, for large data centers and do not appear to be striving towards low energy consumption.

Beyond co-packaged optics, a variety of network scaling solutions that rely on all-optical switches have been proposed. A few of the examples include: (a) a flat datacenter network architecture based on a scalable optical switch system with optical flow control for solving packet contentions at the buffer-less optical switches [21], (b) a bandwidth steering concept that aims to mitigate system fragmentation by using optical switches at the lower layers of a hierarchical network [22], (c) an optical packet switch design for disaggregated data centers that combines the broadcast-and-select scheme with arrayed waveguide grating routers [23], and (d) all-optical networks that are based on demand-oblivious reconfiguration, e.g., Rotornet [24], and Sirius [25], which however incur additional complexity for the congestion control mechanism. Although this paper does not focus on optical switching solutions, in the context of a potential optical-switching era, co-packaged optics can form a compact and efficient solution for interfacing the various system components, e.g., processors, accelerators, or network interface cards, with the optical switching elements.

In this paper, we extend our previous work presented at the European Conference on Optical Communication (ECOC) conference 2021 [11], and we report on the recent activities within the framework of the MOTION research project (Multi-wavelength Optical Transceivers Integrated On Node) [7]. In section 2 we provide a brief overview of the MOTION hardware development activities for the 1st-generation module and we discuss the target specifications for the 2nd-generation module. In section 3, we present a comparison between a baseline network and a network that makes use of a co-packaged-optics-enabled switch module as the basic building block. For the baseline network we consider state-of-the-art switch technologies that are available in the market, while for the proposed network we assume a co-packaged-optics-enabled switch module with 128 ports and 51.2 Tb/s total I/O BW (i.e., 400 Gb/s per port). The comparison shows that—for a network size of >12K end points—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%.

To the best of our knowledge, although there have been studies that focus on the cost implications of the switch radix/capacity in fat-tree networks [26], [27], not much focus has been given on quantifying the effects to the application performance. As such, we carry out a two-fold simulation analysis in section 4 of this paper: we initially evaluate the network locality properties of the system by using virtual-machine (VM) traces from a production data center (section 4.A), and subsequently we quantify the network performance improvements by simulating an all-to-all communication pattern for a variety of message sizes over number of nodes (section 4.B). The first part of the analysis reveals that large-scale applications can be placed under up to 50% fewer 1st-level switches, therefore significantly improving the network locality properties of the system. In the second part, the network performance analysis suggests communication speedups of up to 7.1, which translates to execution time reductions of up to 26% and 42.7% for applications with communication ratios of 0.3 and 0.5, respectively. Finally, section 5 concludes the paper.

2. MOTION CO-PACKAGED OPTICS MODULE

The MOTION project is a collaboration between IBM Research and II-VI Incorporated with the aim to develop a chip-scale co-packaged optics module, which can be directly attached to the top of an organic 1st-level package. The module consists of (a) Vertical Surface Emitting Laser (VCSEL) arrays, (b) surface-illuminated photodiodes, and (c) CMOS driver/receiver Integrated Circuits (ICs) flip-chip attached on a glass carrier. Fig. 1 (left) shows the 1st-generation module along with an indicative eye diagram (right) for one of the sixteen channels operating at a non-return-to-zero (NRZ) modulation format and 50 Gb/s. Four key differences between the MOTION technology and the technologies being developed in [12], [20] are: (1) the use of multimode fiber with a distance range of 50-100m, (2) a separable optical connector interface (as opposed to fixed fiber pigtails), (3) a package that can be directly soldered to the first-level package avoiding the mechanical complications and expense of an electrical socket, and (4) the use of internal lasers which avoid the extra package and special fibers needed to support external lasers. Another key difference is that the MOTION

Table 1: MOTION specifications [8]

Parameter	1st generation	2nd generation
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	

*The NRZ format does not require Forward Error Correcting codes.

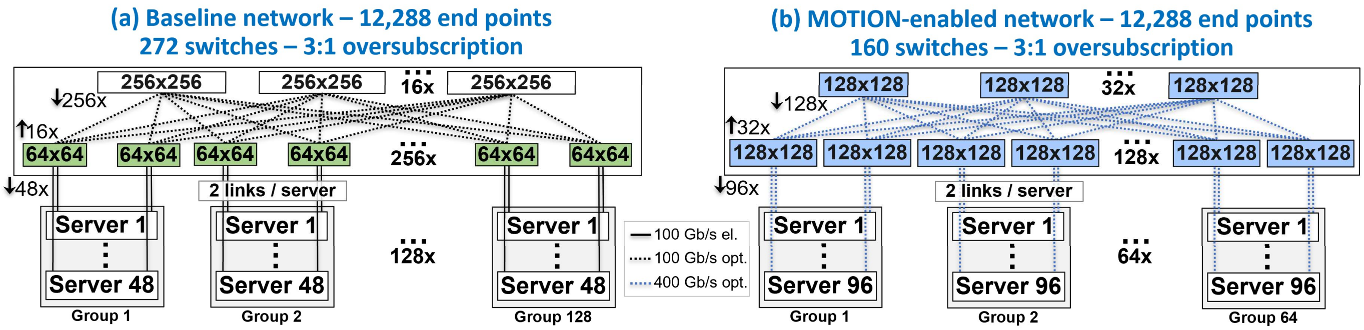

Fig. 2. (a) Baseline network, (b) MOTION-enabled network with 4x bandwidth per server and 2x servers per 1st-level switch.

project has low energy consumption as a primary objective. The energy consumption of the 1st generation module is ~5x lower than the current OIF specifications. The MOTION module is built upon a 13x13-mm^2 glass carrier that holds the flip-chip mounted VCSEL and surface-illuminated photodiodes arrays as well as driver and receiver ICs (see Figure 1 middle). The VCSELs and photodiodes operate in the 930-950-nm range, and the light path is through the glass carrier. These 1st-generation optical components and the ICs are frequency-agile and can operate at any data rate up to 56 Gb/s/ch. The total I/O BW for the 16-channel transceiver is 0.9 Tb/s, and the BW density is 5.3 Gb/s/mm^2. Table 1 summarizes the specifications for both the first and second generations of the MOTION co-packaged optics module. The final target for the 2nd-generation module is to incorporate 32 optical channels in an area of 13 x 13 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. We refer to [7]-[9], for more details about the physical-layer design and packaging as well as for figures including exploded views of the components inside the MOTION module.

3. BASELINE AND MOTION-ENABLED NETWORKS

Fig. 2 (a) shows the baseline data center network that relies on state-of-the-art switches and follows the design principles of modern data center systems [28]. The network has an oversubscription ratio of 3:1 (which is in line with today's state-of-the-art), it uses 25.6-Tb/s switches at the second (spine) switch layer and 6.4-Tb/s switches at the first (leaf) switch layer. All links operate at 100 Gb/s throughout the network. The use of 25.6-Tb/s spine switches (today's state of the art [3]) simplifies the network topology since it reduces both the number of switch layers and the number of required switches, which is advantageous from an administrative/management perspective as well. Due to the higher distances between the switch layers, i.e., higher reach is required, it is the standard approach to use Active Optical Cables (AOCs) for these connections. On the other hand, we follow the common approach of using Direct Attach Copper (DAC) cables between the 1st-level switches and the servers, which is typically driven by cost-saving reasons. However, one important consequence of increasing the data rates in modern systems is that the effective reach of these DAC links has shrunk dramatically. Therefore, this constrains the equipment placement in proximity to lower-radix switches, hence the choice of 6.4-Tb/s leaf switches. In total, the system incorporates 272 switch ASICs, which can interconnect up to 12,288 end points. For network redundancy, the servers connect to two end points from two different switches, which results in a total number of 6,144 servers. The top switch layer has sixteen 256-port spine switches, while the bottom layer has 256 64-port leaf switches. Every two 1st-level switches connect to a *group* of 48 servers, resulting in 128 groups in total. Each group of servers can physically expand over multiple racks, e.g., two racks of 24 servers each. The intra-group and inter-group communication costs are 1 and 3 hops, respectively. The system's bisection bandwidth is 409.6 Tb/s.

Fig. 2 (b) shows a promising next-generation network that uses the same switch radix at the leaf and spine layers and only optics for connectivity, i.e., servers are connected to the 1st-level switches optically. The MOTION analysis in this paper is based on this network, and we assume that it is composed of 51.2-Tb/s switches that make use of co-packaged optics for I/O. As we presented in the scalability analysis of [9], co-packaged optics technology (or possibly electrical I/O off the top of the carrier, or a mix of top-side and bottom side I/O) appear to become necessary for the 102.4-Tb/s switch generation. However, the technology holds promise for top-side-only I/O in the 51.2-Tb/s switch generation as well. A 51.2-Tb/s implementation with only optics for I/O is appealing, since it could enable energy [29] and potential cost [30] gains by eliminating the need for driving the pluggable optics modules through long electrical channels. We also note that our modelling analysis can be readily scaled to follow-on switch generations; it is beyond the scope of this paper to predict when switches with co-packaged optics will be used in large-scale production systems. Again, an oversubscription ratio of 3:1 is applied, and we assume a switch radix of 128 ports with a data rate of 400 Gb/s per port. As has been presented in our previous scalability analysis presented in [10], a 51.2-Tb/s switch module with only optics for I/O could be built on a 90x90-mm^2 carrier, and it would require sixteen 2nd-generation MOTION modules. This solution would maximize the energy gains since it would eliminate the need for driving any 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 of the package. Such a solution would require a smaller carrier size of 70x70 mm^2, and it would make use of only eight 2nd-generation modules, i.e., half modules compared to the fully co-packaged approach [10].

Unlike as in the baseline case, the longer reach of optics, combined with the higher-radix switches, enables a 2x increase in the number of servers that connect to every 1st-level switch. This results in a total number of 64 groups with 96 servers each, which is combined with a 4x increase in the server BW. This can be 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, this approach requires 112 fewer switches, resulting in a switch count reduction of 41%. This translates to both reduced cost and less administration and management overhead. As for the bisection BW, the higher data rates result in a 4x increase, which corresponds to a total bisection BW of 1,638.4 Tb/s. Regarding the power consumption, co-packaged optics can in principle lead to reductions of 25-50% over pluggable optics, provided that a power efficiency target of at least 10-15 pJ/bit can be met [31]. In addition, current estimations forecast that an up to 50% system-level reduction in the cost per capacity compared to pluggable optics is feasible [31]. As such, co-packaged optics appear as a promising solution for keeping up with BW scaling in future data center and HPC networks.

Although the network architectures of this analysis have been inspired by the data center area, the advantages of using co-packaged optics apply to the HPC area as well; fat-tree topologies are widely used in HPC networks since they have: (a) low-complexity routing schemes, (b) good scale-out characteristics, and (c) efficient partitioning properties.

4. PERFORMANCE ANALYSIS

In this section, we assess the performance of the proposed system by directly comparing the two architectures of Fig. 2 through discrete-event simulations. Our analysis is categorized into two parts. For the first part, we have extended CloudSim Plus [32], and we use it in conjunction with a VM trace file from a production data center in order to evaluate the network locality properties of both systems. CloudSim Plus is an open source simulator for modeling, simulation, and experimentation of cloud computing infrastructures and services. Regarding the VM trace file, we extracted it from the publicly-available AzureTracesForPacking2020 dataset [33], [34], and we used it as an input to CloudSim Plus to simulate the dynamic placement and departure of the VMs. More details about the characteristics of the trace file are provided in section 4.A.

For the second part of our analysis, we use the Venus discrete-event network simulator [35] in order to quantify the network performance improvements for the proposed system. Venus is an OMNeT++-based [36] simulator and has been developed within IBM Research over the past decade. To quantify the improvements, we simulate an *all-to-all* communication pattern that emulates the function of collective operations like the Message Passing Interface (MPI) all-to-all, MPI all-gather, and MPI all-reduce. Such operations are typical in a variety of parallel applications. The simulated all-to-all pattern follows the optimized exchange pattern implementation presented in [37]. 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 systems of Fig. 2 have an oversubscription ratio of 3:1, there is always some form of contention for the shared resources due to the messages that need to 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 assume that it remains constant throughout the communication operation, which is in agreement with modern deep learning models in which the message size is dictated by the number of parameters in the model [38].

A. Network Locality Analysis with Virtual-Machine Traces

For both the baseline and the proposed systems we simulate 48-core servers with 384 GiB of memory, while for the network BW we use 100 Gb/s for the baseline case and 400 Gb/s for MOTION case. The original AzureTracesForPacking2020 dataset covers a 14-day period, while we focuse on the requests that arrived within the first seven days. The VM end times can extend beyond the 14th day and they are capped at 90 days to anonymize time [34]. Our analysis focuses on large-scale applications and the network-locality properties of the tested systems, thus we selected only VM group requests that require at least 48 servers to host them, which is equal to the size of a group of servers that connect to the same 1st-level switches for the baseline case. Since the dataset does not explicitly identify group requests that correspond to large-scale applications, we assume that a set of VMs belongs to a large-scale VM group request if all the following conditions are met: (a) the VMs share the same tenant and vmType ids, i.e., they have been requested by the same user account and they request the same amount of resources,

Table 2: VM types appearing in the trace file 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

(b) they have the same priority id, i.e., they all have been requested as either high-priority or low-priority, and (c) they start and end at the same times with a max difference of 1 sec, which implies that they were possibly requested with the intention to execute a parallel workload/application. Regarding the vmTypes, the dataset associates them with multiple resource request ratios depending on different type of servers, information which is also anonymized [34]. For the purposes of this analysis and to stress both systems, we use the ratios that maximize the requested number of cores. For the above filtering criteria, the trace file consists of 631 group requests that correspond to a total number of >62.5K VMs. Table 2 presents the VM types of the trace file sorted by popularity, while the [min / avg / max / stddev] values of the request lifetimes are [3s / 1.49d / 89.9d / 8.17d] and respective interarrival times of the requests are [0s / 14.1m / 1.2d / 1.27h].

Fig. 3 (a) shows the distribution of the requests according to their size, which is expressed in number of VMs: 50.7% of the requests belong to the [48, 50] category, while [51, 100] is the 2nd most popular category with 20.4% of the requests. The biggest request in number of VMs consists of 400 16-core VMs (i.e., 6,400 cores), while the biggest request in number of cores consists of 312 32-core VMs (i.e., 9,984 cores).

Fig. 3 (b) considers 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 minimum number of server groups, and (b) there are always available resources for the placement. For the baseline case, 4.9% of the requests fit in 1 group, 79.7% require 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 fit 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 server group comes with two key advantages: (a) communication cost is always one hop max, and (b) network contention associated with crossing the network spine is eliminated. It is common practice for many cloud providers to offer various options to control the placement of VM groups, which can help reducing the network latency and/or minimizing the chances of service disruption [39]-[42]. Such offerings can contribute to both enhanced application performance and improved user experience overall.

To study the impact of contention for system resources, we simulate two VM placement algorithms. Initially, we simulate first-fit [43], [44], which is a well-known algorithm for its simplicity; servers are scanned in serial order until a suitable server is found to host the VM. If no such server exists, the request is denied. Since we simulate group requests, we consider that a request is denied when no resources exist for all the VMs of a group cumulatively. Fig. 3 (c) shows the corresponding simulation results. Since first-fit is network agnostic, the placement of the VMs is realized in the same way independently of the network topology; for both systems, 20.1% of the requests (not shown in Fig. 3 (c)) are denied due to either lack of sufficient resources or due to system fragmentation that turns the resources unreachable (more details are given below in the description of Fig. 3 (e) and Fig. 3 (f)). For the baseline case, 0.5% of the requests are placed in 1 group, 40.1% in 2 groups,

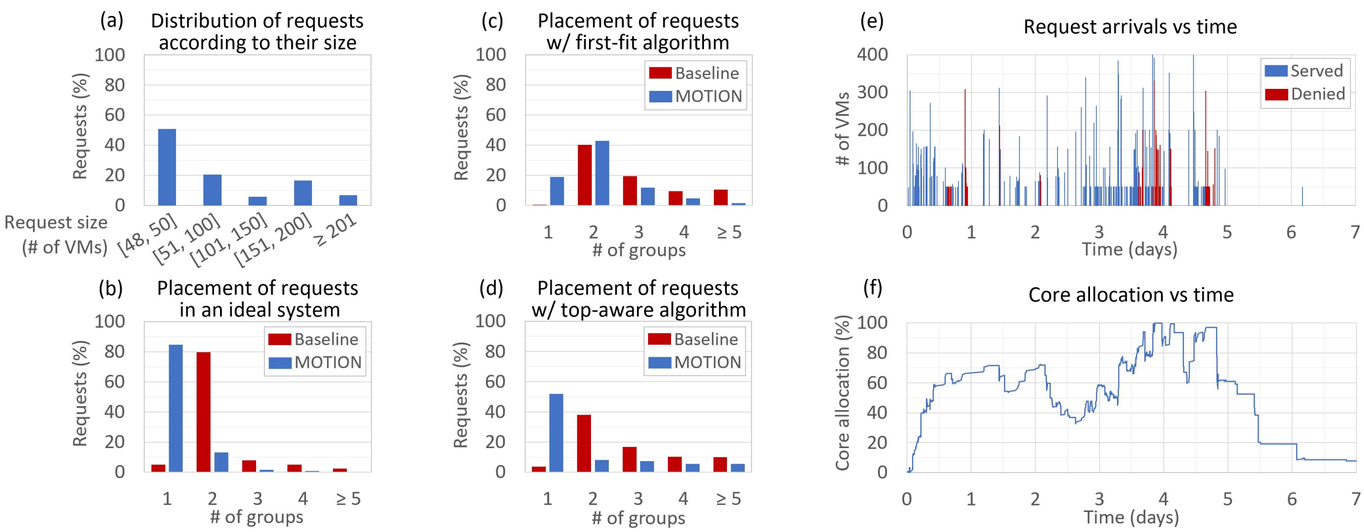

Fig. 3. (a) Distribution of requests according to their size, (b) Placement of requests in an ideal system, i.e., assuming that (i) every request can be placed in the minimum number of groups and (ii) there are always available resources, (c)-(d) Placement with first-fit and topology-aware algorithms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

while 19.3%, 9.5% and 10.5% in 3, 4 and 5-10 groups, respectively. For the MOTION case on the other hand, 19% of the requests are placed in 1 group, while 42.8%, 11.7%, 4.8% and 1.6% are placed in 2, 3, 4 and 5-7 groups, respectively.

As a next step, we simulate a topology-aware algorithm that follows part of the principles of the scheduling policy presented in [43], which targets to reduce network sharing and fragmentation in HPC systems. According to the simulated algorithm, small requests that can fit in a server group populate the system from the first group onwards, whereas big requests that require multiple server 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 be particularly beneficial for the MOTION case since 84.6% of the requests can ideally fit in one server group (which is 4.9% for the baseline case). Fig. 3 (d) shows the respective simulation results. As with the first-fit case, slightly over 20% of the requests are denied (not shown in Fig. 3 (d)), i.e., despite the differences between the algorithms and the way they place the VMs, the capacity of the systems remains the same, which results in slight variations only in the number of denied requests. For the baseline case, no significant differences are observed vs first-fit, although the number of requests placed in one server group increases from 0.5% to 3.8%. On the other hand, the MOTION system presents remarkable improvements: 52% of the requests are placed in 1 group (vs 19% for first-fit), while 8.1%, 7.4%, 5.4% and 5.6% of the requests are placed in 2, 3, 4 and 5-10 groups, respectively. It should be noted that the majority of the denied requests, i.e., 17.3% of all requests, can ideally be placed in 1 group.

Finally, Fig. 3 (e) and Fig. 3 (f) focus on the MOTION case and the topology-aware algorithm: they show the request arrivals along with the respective system core allocation vs time, respectively. For example, at the end of the 4th day, requests are denied because the system core allocation reaches 100%, i.e., there are no available cores to host additional requests; at the end of the 1st day on the other hand, requests are denied due to system fragmentation, i.e., most servers have 16 free cores available, prohibiting the system from being able to host additional 32-core VMs. We observed similar behavior for the rest configurations as well.

B. Network Performance Analysis for an All-to-All Communication Pattern.

For a total number of N nodes, an all-to-all communication is a deterministic sequence of N phases, where in every successive phase each individual node communicates with a different destination node. As a result, the total number of messages exchanged between the nodes equals N^2 (each node sends a message to itself as well). In our analysis, we assume that each individual node corresponds to a distinct physical server of the systems of Fig. 2, while we also assume that all N nodes are located in the minimum number of server groups. This corresponds to the case of the topology-aware algorithm presented in section 4.A, which is designed to place the VMs of a request in the minimum possible number of server groups. Therefore, the total number of messages that need to go through 1 hop can be expressed as $M_1 = N \times S$, where S equals the server group size of the system, i.e., 48 or 96 for the baseline and MOTION cases, respectively. As such, the number of messages that need to undergo 3 hops can be expressed as $M_3 = N^2 - N \times S$. Since both systems have a 3:1 oversubscription ratio, it is evident that the higher the M_3, the slower the all-to-all communication will be completed since there are more messages that contend for the oversubscribed resources. We plotted the percentage of 3-hop messages vs the number of nodes in Figure 4, while for the N value we iterate over the [96, 768] range with a step 96. We do not examine the case of 48 nodes, i.e., the server group size of the baseline case, because both systems would behave identically since there are only 1-hop messages for both of them. It is expected that both the bigger server group size and the BW increase will be beneficial to the MOTION system for lower N values, since it has substantially smaller percentage of 3-hop messages in that range. On the other hand, for higher N values it is expected that the BW increase will dominate over the increase in the server group size, which can be explained by the smaller difference in the percentage of 3-hop messages between the two systems.

Regarding the simulation parameters, for both baseline and MOTION cases we initially use a data rate of 100 Gb/s. In this way, we can examine the potential advantages of the bigger server group size independently of the BW increase in the system. For the MOTION system, we also use a link rate of 400 Gb/s, which corresponds to the 51.2-Tb/s switch discussed in section 3. To distinguish between the two

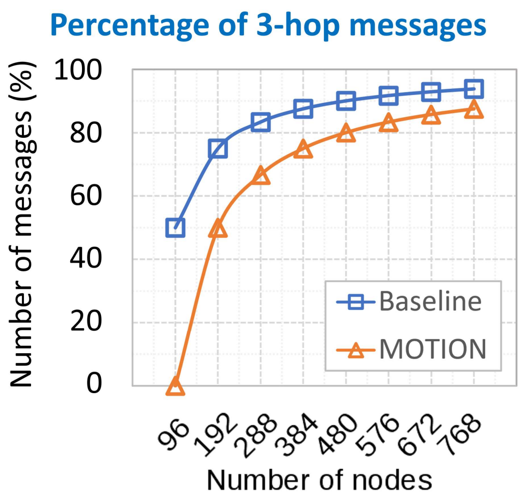

Fig. 4. Percentage of 3-hop messages vs number of nodes. 3-hop messages cross the spine, while 1-hop messages only cross leaf switches. Fewer 3-hop messages correspond to less contention for the oversubscribed resources of the studied systems; thus, fewer 3-hop messages is better.

MOTION configurations, we refer to them as MOTION-100 and MOTION-400, respectively. Regarding the message size, we iterate over the 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. We assume that each NIC and switch adds a 100-ns delay to the traversing data (besides any potential queueing delays). The switches in our model use shared input buffers (128 KiB per port) with Virtual Output Queueing (VOQ), which is known for addressing the well-known problem of Head-of-Line (HOL) blocking [45], [46]. We also simulate the random routing algorithm where, at each routing stage, the next hop is randomly selected over the set of all shortest paths that lead to the destination (contiguous packets from the same message follow the same path). Table 3summarizes the most significant parameters of our analysis.

Fig. 5 compares the three system configurations, i.e., baseline, MOTION-100, and MOTION-400, and it shows the speedup that can be achieved by introducing the topological and BW changes to the system. Fig. 5 (a) compares the baseline and MOTION-100 cases and shows the

Table 3: Simulation parameters

Traffic Generators	
Message size	$[2^8, 2^{15}]$ bytes
Rate	100, 400 Gb/s
Network Interface Cards	
Packet size	$[2^8, 2^{12}]$ bytes
Delay	100 ns
Bandwidth	100, 400 Gb/s
Switches	
Architecture	Shared input buffers with VOQs
Flow control	Credit-based
Buffer size	128 KiB per port
Delay	100 ns
Bandwidth per port	100, 400 Gb/s
Routing algorithm	Random

improvements we can get by doubling the number of servers per 1st-level switch. As expected, the improvements are greater for lower N values. For N=96 nodes, the MOTION-100 case has 1-hop messages only, i.e., it is not affected by the 3:1 oversubscription ratio. For all message sizes we have speedups of >1.7, with the greatest value being 1.78 for message size 2^{14} bytes (i.e., 16 KiB). For N=192, the improvements are lower and the respective speedups range between 1.44 and 1.63. For N=288 and beyond, we observe that the improvements decrease significantly, and the respective speedup values converge toward 1. For N=768 for example, we have minor improvements only with the speedup ranging between 1.06 and 1.08. In general, although there are slight fluctuations in the speedup as the message size increases, we do not observe any particular proportionality between them.

Proceeding to Fig. 5 (b), the comparison between the MOTION-100 and MOTION-400 cases shows the additional improvements we can get by introducing the 4x higher BW to the system. Increasing the BW (without making any changes to the MOTION topology) leads to lower improvements when we combine smaller message sizes with smaller number of nodes. This can be explained by the fact that as the message size and number of messages (which is proportional to the number of nodes) increase, there is a lower relative impact on the performance associated with the all-to-all initialization overhead. From a single-node perspective, the all-to-all communication involves the continuous transmission of N messages, where each successive message is

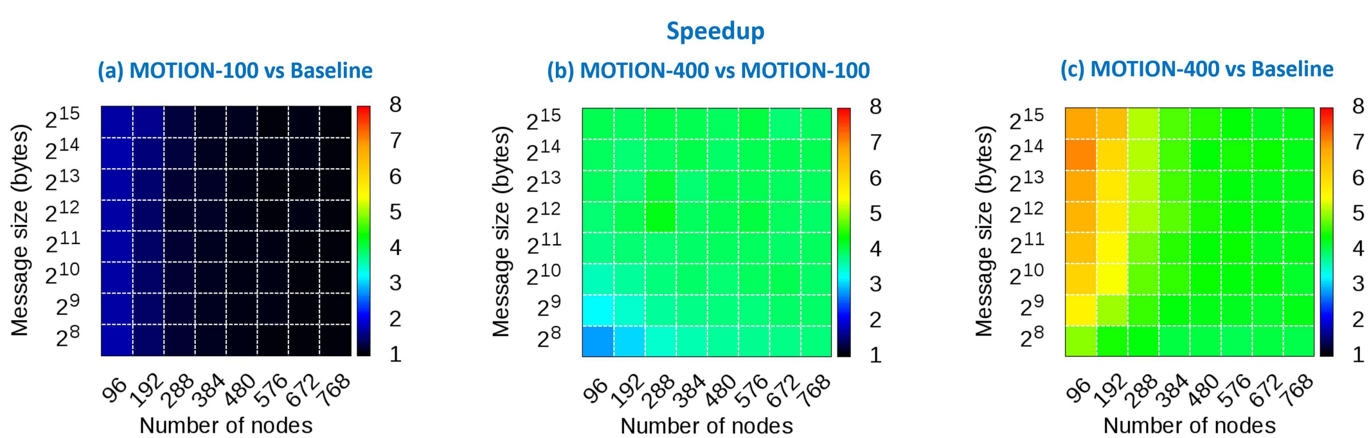

Fig. 5. (a) Speedup for the MOTION-100 case over the baseline case, (b) Speedup for the MOTION-400 case over the MOTION-100 case, (c) Speedup for the MOTION-400 case over the baseline case.

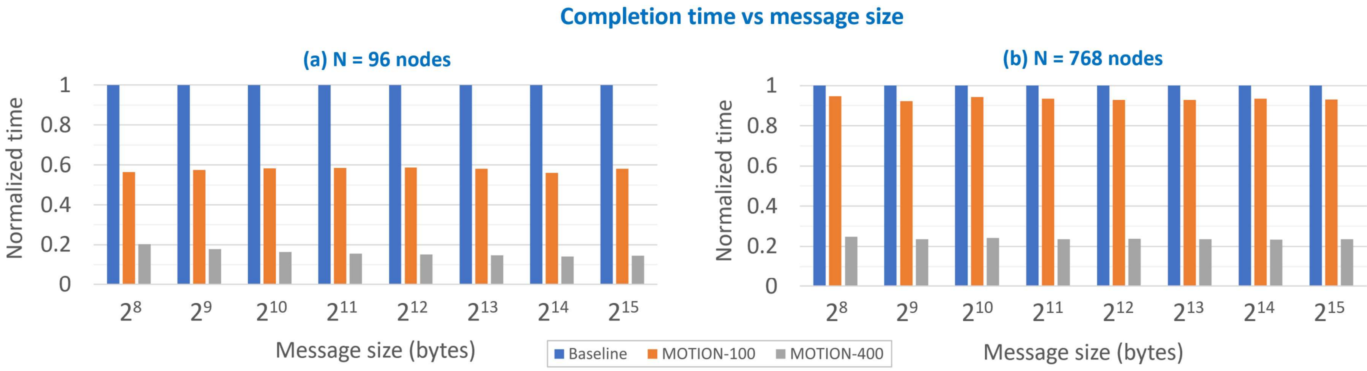

Fig. 6. Total completion time for the all-to-all communication vs message size for all three system configurations, i.e., baseline, MOTION-100, and MOTION-400. Results are for N=96 (left) and N=768 (right) nodes and they have been normalized to the baseline case.

transmitted toward a different destination node. However, there is always an unavoidable initialization overhead that must be paid before the messages start getting delivered to the destination nodes. This overhead originates from (a) the NIC delay, which is paid twice at the transmitter and the receiver points, (b) the switch delay, which is paid once for the 1-hop messages and three times for the 3-hop messages, and (c) the time-of-flight between the different components of the network. This overhead is more pronounced when we combine smaller message sizes with smaller number of messages, i.e., it is relatively much larger when compared to the duration of the messages, hence the lower speedup. Overall, the comparison between the two cases reveals speedups that start from 2.8 and reach up to slightly over 4 for certain cases.

Fig. 5 (c) presents the comparison between the baseline and MOTION-400 cases and shows the total improvements we can get by introducing both the topological changes and the higher BW to the system. It is apparent that the heatmap of Fig. 5 (c) is the result of combining the improvements presented in Fig. 5 (a) and Fig. 5 (b). As a result, the combined improvements appear greater in the region where we combine bigger message sizes with smaller number of nodes. The greatest speedup equals 7.1 and corresponds to the combination of N=96 and message size of 2^{14} bytes (i.e., 16 KiB). As we move toward the other areas of the heatmap, the speedup decreases gradually, but it remains always above 4, which corresponds to the minimum improvements we can get by the 4x increase in the BW.

Beyond the speedup comparisons, Fig. 6 focuses on the cases where N equals 96 (left) and 768 (right) and shows the completion time measurements vs message size for all the three system configurations, i.e., baseline, MOTION-100, and MOTION-400. To facilitate the comparison, we normalized the results to the longest simulated time, which always corresponds to the baseline case. Focusing first on the MOTION-100 case, we see that the topological changes introduce significant improvements when N=96, i.e., when the 3:1 oversubscription ratio is completely bypassed. As seen, the improvements are relatively constant with respect to the message size, and they range between 41.5% and 43.9%. On the other hand, the improvements are far lower for the case of N=768 nodes, i.e., when for both systems the majority of the messages needs to undergo three hops, which results in little room for improvement overall. Again, the improvements are relatively constant, and they range between 5.4% and 7.8%. As for the MOTION-400 case, the comparison versus the baseline system reveals significant improvements for both N=96 and N=768 cases, which are even greater for the case of N=96 nodes. The completion time reduction for N=96 nodes ranges between 79.8% and 85.8%, while the respective reduction for N=768 ranges between 75.1% and 76.6%.

While the results of Fig. 6 focus on the completion time of the all-to-all communication itself, Fig. 7 shows the performance improvements we can get for a potential application as a whole, i.e., an application that incorporates both computation and communication phases. For the calculations we use a range of communication ratios that correspond to the time spent for communication in the baseline case. Additionally, we also assume that the communication ratios correspond to communication time that cannot be overlapped by the computation phases, hence this part of the application can benefit from the improvements in the network. Fig. 7 (top) corresponds to a message size of 2^8 bytes, while Fig. 7 (bottom) corresponds to a message size of 2^{15} bytes. For both message sizes, we focus on three different N values of 96, 192, and 288 nodes. As expected, the benefits are proportional to the communication ratio increase and they are greater for the MOTION-400 case. For a ratio of 0.3 for example and message size of 2^8 bytes, the comparison between the MOTION-400 and baseline cases shows improvements of 23-24%, while the respective improvements for the message size of 2^{15} bytes range between 24-26%. On the other hand, for a more aggressive ratio of 0.5, the improvements range between 38.4-42.7%.

The communication ratios in our analysis are representative of a number of real-world HPC applications. According to a study conducted by the Argonne National Laboratory on the MPI usage on a production supercomputer between 2016-2017 [47], around half of the executed production jobs spent at least 30% of their time in MPI, while 15% of the jobs spent 60% or more. Overall, 34% of the system's core hours were spent in MPI. In more detail, around 30% of the jobs spent at least 30% of their time on collectives, while 10% of the jobs spent 60% or more. At the same time, half of the jobs did not spend any time on point-to-point communication, while around 10% of the jobs spent 30% or more. As for the message size, the majority of the operations were all-reduce with a message size of ≤256 bytes, while slightly more than 20% of the all-reduce operations used a message size of ≥32 KiB. Therefore, optimizing MPI and/or improving the underlying network infrastructure can substantially help in optimizing the performance of the applications, which can lead to notable savings on the core-hour budget as well.

5. CONCLUSION

We investigated the advantages of using co-packaged optics in data center and HPC networks. Co-packaged optics can enable switch implementations with speeds of 51.2 Tb/s and beyond, which can lead to network configurations with 4x higher bisection bandwidth and 41% fewer switch modules. Simulations with virtual-machine traces suggest that co-packaged optics can significantly improve network locality, i.e.,

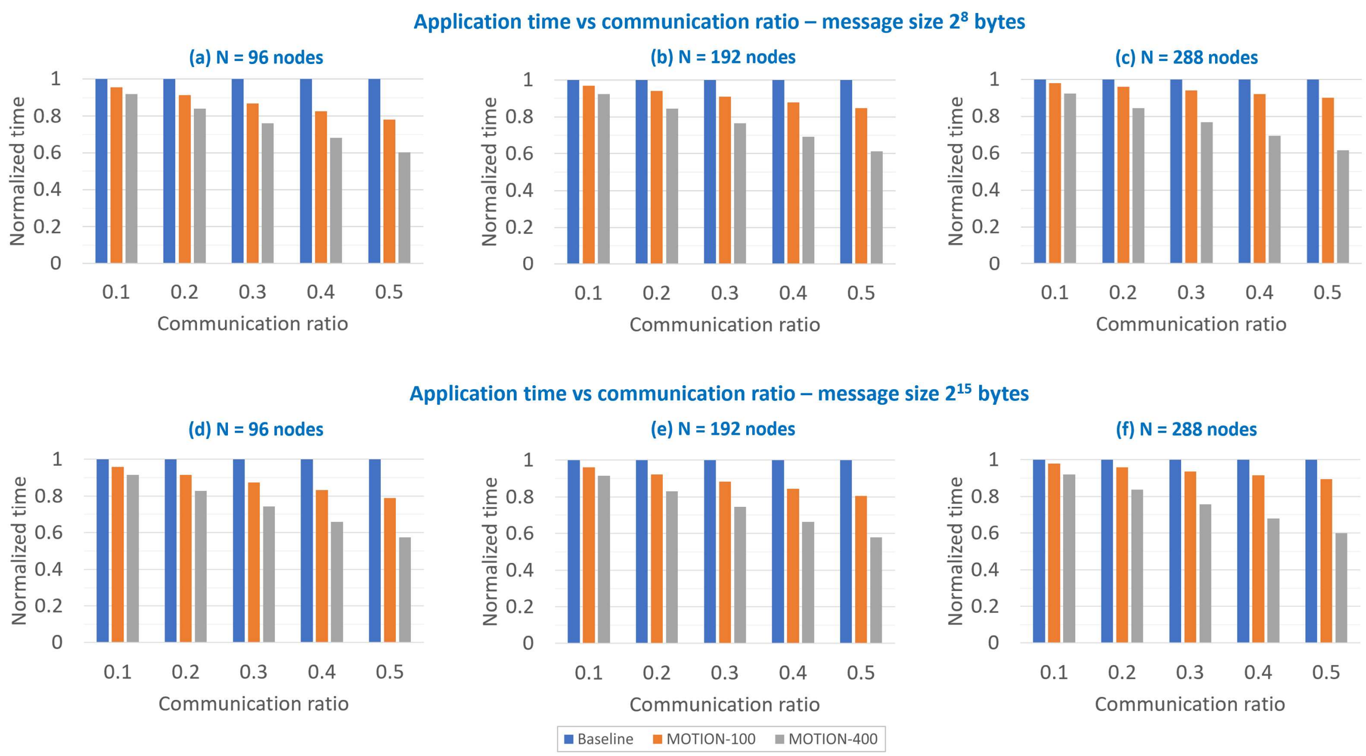

Fig. 7. Application time vs communication ratio for two different message sizes, i.e., 2^8 bytes (top) and 2^{15} bytes (bottom), for N=96, 192, and 288 nodes. Communication ratio corresponds to the time spent for communication in the baseline case. Results have been normalized to the baseline case.

large-scale applications can be placed under up to 50% fewer 1st-level switches. In addition, network simulations for an all-to-all pattern suggest communication speedups of up to 7.1, which translate to execution time reductions of up to 26% and 42.7% for communication ratios of 0.3 and 0.5, respectively, that are typical for a wide range of high-performance computing applications. Co-packaged optics offer a promising solution for keeping up with the bandwidth scaling in future data center and HPC networks.

Funding Information. 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. C. Tsai, C. Lin, P. Liu and J. Wu, "Communication Scheduling Optimization for Distributed Deep Learning Systems," 2018 IEEE 24th International Conference on Parallel and Distributed Systems (ICPADS), Singapore, Singapore, 2018, pp. 739-746, doi: 10.1109/PADSW.2018.8644593.
5. J. Shalf, "HPC Interconnects at the End of Moore's Law," OFC2019, San Diego, CA, USA, 2019, pp. 1-3.
6. G. Georgakoudis, N. Jain, T. Ono, K. Inoue, S. Miwa and A. Bhatele, "Evaluating the Impact of Energy Efficient Networks on HPC Workloads," 2019 IEEE 26th International Conference on High Performance Computing, Data, and Analytics (HiPC), Hyderabad, India, 2019, pp. 301-310, doi: 10.1109/HiPC.2019.00044.
7. D. M. Kuchta, et al., "Multi-Wavelength Optical Transceivers Integrated on Node (MOTION)," OFC2019, San Diego, CA, USA, 2019, pp. 1-3.
8. D. M. Kuchta, "High Speed VCSELs and Co-Packaging for Short Reach Communication within Cloud and High-Performance Computing," 2019 24th Microoptics Conference (MOC), Toyama, Japan, 2019, pp. 298-299, doi: 10.23919/MOC46630.2019.8982885.
9. 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
10. P. Maniotis, et.al., "Co-packaged optics for HPC and data center networks," Proc. SPIE 11692, Optical Interconnects XXI, 1169205 (5 March 2021)
11. P. Maniotis, L. Schares, D. M. Kuchta, B. Karacali, “Improving Data Center Network Locality with Co-packaged Optics,” 2021 European Conference on Optical Communications (ECOC), We1F.1, 13-16 September 2021
12. https://www.facebook.com/CoPackagedOpticsCollaboration/
13. 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
14. 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
15. https://rockleyphotonics.com/wp-content/uploads/2020/03/Rockley-Photonics-OptoASIC-Flyer.pdf
16. https://acacia-inc.com/wp-content/uploads/2019/06/Optinet-China-2019_Acacia_Fenghai-Liu_UpLoad_v1.pdf
17. https://www.sdxcentral.com/articles/news/broadcom-unveils-first-co-packaged-optical-switch/2021/01/
18. S. Chun, W. D. Becker, J. Casey, S. Ostrander, D. Dreps, J. A. Hejase, R. M. Nett, B. Beaman, J. R. Eagle, "IBM POWER9 package technology and

design," in IBM Journal of Research and Development, vol. 62, no. 4/5, pp. 12:1-12:10, 1 July-Sept. 2018, doi: 10.1147/JRD.2018.2847178.
19. W.D. Becker, H. Harrer, A. Huber, W.L. Brodsky, R. Krabbenhoft, M.A. Cracraft, D. Kaller, G. Edlund and T. Strach, "Electronic packaging of the IBM z13 processor drawer," in *IBM Journal of Research and Development*, vol. 59, no. 4/5, pp. 13:1-13:12, July-Sept. 2015
20. https://www.oiforum.com/technical-work/hot-topics/co-packaging/
21. F. Yan, W. Miao, O. Raz and N. Calabretta, "Opsquare: A flat DCN architecture based on flow-controlled optical packet switches," in Journal of Optical Communications and Networking, vol. 9, no. 4, pp. 291-303, April 2017, doi: 10.1364/JOCN.9.000291.
22. G. Michelogiannakis, et al., "Bandwidth steering in HPC using silicon nanophotonics," In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC '19). Association for Computing Machinery, New York, NY, USA, 2019, Article 41, 1–25.
23. N. Terzenidis, et al., "High-Port and Low-Latency Optical Switches for Disaggregated Data Centers: The Hipoλaos Switch Architecture [Invited]," J. Opt. Commun. Netw. 10, B102-B116 (2018)
24. William M. Mellette, et al. 2017. RotorNet: A Scalable, Low-complexity, Optical Datacenter Network. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication* (*SIGCOMM '17*). Association for Computing Machinery, New York, NY, USA, 267–280.
25. Hitesh Ballani, et al. 2020. Sirius: A Flat Datacenter Network with Nanosecond Optical Switching. In *Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication* (*SIGCOMM '20*). Association for Computing Machinery, New York, NY, USA, 782–797.
26. C. Minkenberg, R. P. Luijten and G. Rodriguez, "On the optimum switch radix in fat tree networks," 2011 IEEE 12th International Conference on High Performance Switching and Routing, 2011, pp. 44-51
27. C. Minkenberg, "How to slice and dice your switch capacity," 45th European Conference on Optical Communication (ECOC 2019), 2019, pp. 1-4
28. https://www.arista.com/en/solutions/software-defined-cloud-networking-solutions#Spine-Leaf-Network
29. C. Minkenberg et al., "Reimagining datacenter topologies with integrated silicon photonics," in IEEE/OSA Journal of Optical Communications and Networking, vol. 10, no. 7, pp. 126-139, July 2018, doi: 10.1364/JOCN.10.00B126.
30. N. Kucharewski, C. Minkenberg and G. Rodriguez, "Network Architecture in the Era of Integrated Optics," 2018 Optical Fiber Communications Conference and Exposition (OFC), San Diego, CA, 2018, pp. 1-3.
31. C. Minkenberg, R. Krishnaswamy, A. Zilkie, and D. Nelson, (2021), Co-packaged datacenter optics: Opportunities and challenges. IET Optoelectron, 15: 77-91. https://doi.org/10.1049/ote2.12020.
32. 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
33. 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.
34. https://github.com/Azure/AzurePublicDataset/blob/master/AzureTracesForPacking2020.md
35. R. Birke, G. Rodriguez, and C. Minkenberg, "Towards massively parallel simulations of massively parallel high-performance computing systems," In Proceedings of the 5th International ICST Conference on Simulation Tools and Techniques (SIMUTOOLS '12). ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2012, Brussels, BEL, 291–298.
36. A. Varga, "A Practical Introduction to the OMNeT++ Simulation Framework," In: Virdis A., Kirsche M. (eds) Recent Advances in Network Simulation. EAI/Springer Innovations in Communication and Computing. 2019, Springer, Cham.
37. B. Prisacari, G. Rodriguez, C. Minkenberg, and T. Hoefler. "Bandwidth-optimal all-to-all exchanges in fat tree networks," In Proceedings of the 27th international ACM conference on International conference on supercomputing (ICS '13). Association for Computing Machinery, New York, NY, USA, 139–148, 2013.
38. F. Iandola. (2016). Distributed deep neural network training: A measurement study. 10.13140/RG.2.2.27894.04165.
39. https://cloud.ibm.com/docs/virtual-servers?topic=virtual-servers-placement-groups
40. https://azure.microsoft.com/en-us/blog/introducing-proximity-placement-groups/
41. https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/placement-groups.html
42. https://cloud.google.com/compute/docs/instances/define-instance-placement
43. A. Jokanovic, et al., "Quiet Neighborhoods: Key to Protect Job Performance Predictability," 2015 IEEE International Parallel and Distributed Processing Symposium, 2015, pp. 449-459
44. https://en.wikipedia.org/wiki/Bin_packing_problem#First-Fit_(FF)
45. R.J. McMillen, "Packet switched multiple queue NxM switch node and processing method," US Patent no. 4623996A, filed Oct. 18, 1984, published Nov. 18, 1986.
46. M. Karol, M. Hluchyj and S. Morgan, "Input Versus Output Queueing on a Space-Division Packet Switch," in IEEE Transactions on Communications, vol. 35, no. 12, pp. 1347-1356, December 1987
47. https://sc18.supercomputing.org/proceedings/tech_paper/tech_paper_files/pap537s5.pdf