Application-aware Configuration of All-optical Interconnects in Hyper-FleX-LION

(*Invited Paper*)

Hao Yang[†‡] and Zuqing Zhu[†]

[†]School of Information Science and Technology, University of Science and Technology of China, Hefei, China

[‡]Department of Information Engineering, Southwest University of Science and Technology, Mianyang, China

[†]Email: {zqzhu}@ieee.org

***Abstract*—Due to the advantages of optical circuit switching (OCS), all-optical interconnects (AOIs) for data center networks (DCNs) have attracted intensive interests recently. Hyper-FleX-LION is a highly-flexible AOI architecture that operates with the OCS based on wavelength-division multiplexing (WDM). In this paper, we present our recent research activities on Hyper-FleX-LION. First, to prove the superiority of Hyper-FleX-LION, we enumerate various communication patterns of distributed machine learning (DML) in DCNs, and analyze the acceleration effect achieved by Hyper-FleX-LION over existing interconnect architectures, such as the hybrid optical/electrical interconnect based on optical cross-connect (HOE-w/OXC). Results show that Hyper-FleX-LION can better accelerate the tasks of DML. Then, we classify network applications in DCNs as bandwidth-intensive and data-intensive ones, and analyze the operational costs and task completion time that Hyper-FleX-LION brings to them and how to optimize the topology design and traffic routing of Hyper-FleX-LION adaptively. The Analysis results confirm the importance of designing an application-aware configuration scheme for Hyper-FleX-LION.**

***Index Terms*—Data center networks, All-optical interconnects, Application-aware configuration, Distributed machine learning.**

I. Introduction

Recently, the fast emergence of network applications has promoted the research and development (R&D) on data center networks (DCNs) [1, 2]. Hence, the architectures of network interconnects in DCNs are undergoing revolutionary changes to match the capacity/latency/cost of interconnects among computing and storage platforms and the quality-of-service (QoS) demands of network applications. The traditional interconnects that are solely based on electrical packet switching (EPS) have a number of drawbacks, such as ever-increasing power consumption and limited port capacity, which make them difficult to cope with the time-varying traffic patterns and huge traffic volumes of most network applications today [3]. Optical circuit switching (OCS) is a promising replacement for EPS, since it can achieve higher energy efficiency and larger port capacity [4–8]. Therefore, the all-optical interconnects (AOIs) that leverage OCS have been designed and implemented to assist or even replace EPS-based interconnects in DCNs for supporting various network applications better [9–13].

Yoo *et al.* [15, 16] designed and fabricated FleX-LION, which is an integratable OCS device, and with it, large-scale reconfigurable AOIs (namely, Hyper-FleX-LION) can be architected to carry different traffic patterns efficiently [13]. Here, the topology of an AOI that interconnects N racks can be referred to as N-Hyper-FleX-LION, and Fig.1 shows that the architecture and operation principle of a 4-Hyper-FleX-LION. There is an arrayed waveguide grating router (AWGR) sitting in the middle of the 4-Hyper-FleX-LION, and the transmitting and receiving structures of each rack are located on its left and right, respectively. As for each rack, its top-of-rack (ToR) switch has 4 ports and each port equips a transceiver (TRX).

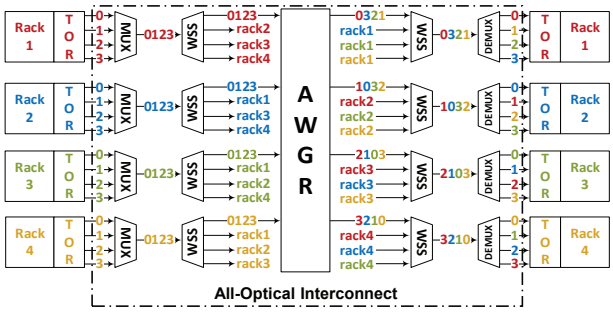

Fig. 1. AOI in Hyper-FleX-LION interconnecting 4 racks, MUX/DEMUX: wavelength multiplexer/demultiplexer, WSS: wavelength selective switch, AWGR: arrayed waveguide grating router (adapted from [14]).

As illustrated in Fig. 1, the wavelengths used by the TRXs are labeled by numbers, while the color of each number denotes the source rack of a TRX. In the transmitting structure, all the outputs of a ToR switch are wavelength-multiplexed before entering a wavelength selective switch (WSS). One of the WSS' outputs is connected to the AWGR, and the others go directly to the inputs of the WSS' in the receiving structures of other racks. In the receiving structure of each rack, the optical signals are first groomed by its WSS and then distributed to the TRXs of the ToR switch by a wavelength de-multiplexer. Hence, by utilizing the wavelength switching capability of the AWGR and adjusting the switching states of the WSS', we can get various topologies to interconnect the racks (*e.g.*, the configuration in Fig. 1 leads to a full-mesh topology among the 4 racks). In other words, if we regard each wavelength channel from the TRXs on a ToR switch as the origin of a lightpath, the lightpath can be set up adaptively to use any of the racks in the Hyper-FleX-LION as its destination (*e.g.*,

with software-defined networking (SDN) [17, 18]).

Note that, compared with EPS, OCS is actually less adaptive because of the large switching granularity and long reconfiguration latency [19]. Even though these issues can be partially resolved by introducing virtualization techniques [20–25] to slice virtual networks and grooming the traffic of similar network applications with them, the benefits of AOIs cannot be fully explored if their topologies are not managed intelligently or the traffic through them are not scheduled adaptively [10]. Therefore, we need to study the problem of how to effectively operate the AOIs in Hyper-FleX-LION to efficiently serve complicated network applications in DCNs.

In this paper, we describe our recent research activities on the aforementioned problem. Section II discusses our investigation that confirms the acceleration effect on distributed machine learning (DML) achieved by Hyper-FleX-LION over an existing hybrid optical/electrical interconnect based on optical cross-connect (HOE-w/OXC). In Section III, we classify network applications in DCNs as bandwidth- and data-intensive ones, and analyze the operational costs and task completion time that Hyper-FleX-LION brings to them and how to optimize the topology design and traffic routing of Hyper-FleX-LION. Finally, Section IV summarizes the paper.

II. Acceleration of Distributed Machine Learning

As one of the current mainstream network applications, DML not only occupies high communication bandwidth, but also has a variety of traffic patterns with different characteristics, which brings great challenges to network interconnects in DCNs [26]. Previously, researchers added OXC to EPS-based interconnects to realize HOE-w/OXC, and they have proved that this architecture can effectively accelerate DML tasks by reasonably reconfiguring the OXC's connectivity to assist the traffic routing in the EPS-based interconnect [27]. However, limited by the one-to-one connectivity of OXC, HOE-w/OXC still cannot accelerate DML tasks to the maximum extent.

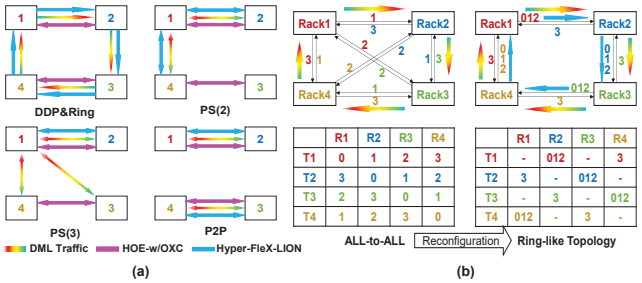

Fig. 2. Acceleration of DML tasks with various traffic patterns using a 4-Hyper-FleX-LION, (a) Traffic patterns of DML architectures, (b) Reconfiguration of Hyper-FleX-LION. DDP: Distributed Data Parallel, Ring: Ring-AllReduce, PS(N): Parameter Server with 1 master node and N worker nodes, P2P: Peer-to-Peer (adapted from [14]).

Fig. 2(a) shows the traffic patterns of different DML tasks, where the colorful arrows denote the traffic, and the purple arrows represent the acceleration bandwidth that can be provided by HOE-w/OXC to the DML tasks [14]. Here, we define the *acceleration bandwidth* as the bandwidth that can be provisioned to DML in addition to that of the basic EPS-based interconnect. We can find that the acceleration bandwidth provided by the OXC cannot adapt to the traffic patterns of all the DML tasks. For example, OXC can only provision acceleration bandwidth between *Racks* 1 and 2 and between *Racks* 3 and 4 for the DML in *DDP&Ring*. Hence, the other traffic flows that cannot be accelerated will become bottlenecks and affect the overall acceleration of the DML. The same problem also occurs for the DML in *PS*, and the more worker nodes that the DML in *PS* allocates, the more bottlenecks will be generated. On the other hand, the blue arrows represent the acceleration bandwidth that Hyper-FleX-LION can provide. The configuration for the DML in *Ring&DDP* is shown in Fig. 2(b). We configure the wavelength switching in the WSS' according to the table at the bottom-left of the figure, and thus the topology changes from a full-mesh to a ring, which matches to the traffic pattern of the DML in *Ring&DDP* exactly. Moreover, in addition to the wavelength that makes up the basic ring topology, there are two more wavelength channels per rack, which can be leveraged to provide more acceleration bandwidth. Therefore, Hyper-FleX-LION can achieve a better acceleration effect than HOE-w/OXC for the DML in *Ring&DDP*.

Similarly, Hyper-FleX-LION can also provide more acceleration bandwidth to the DML in *PS(2)* than HOE-w/OXC. Specifically, *Rack* 1 uses two of the ports on its ToR switch to carry the basic communications between the server and the two workers, and allocates the remaining two ports to provide acceleration bandwidth. However, the situation becomes different for the DML in *PS(3)*, when an additional worker is placed on *Rack* 3. As Hyper-FleX-LION already uses three ports on *Rack* 1 for the basic communications, it can only use the remaining port to provide acceleration bandwidth, which makes its acceleration effect the same as that of HOE-w/OXC. At last, as for the DML in *P2P*, because its traffic only occurs between 2 racks, the acceleration effects of Hyper-FleX-LION and HOE-w/OXC are the same again, as shown in Fig. 2(a).

III. Application-aware Topology Configuration

Our previous work in [14] has verified that Hyper-FleX-LION possesses the flexibility for dealing with various network applications in DCNs. However, how to better plan its network topology to maximize the advantages for specific network applications is still an unexplored problem. Hence, we need to roughly classify network applications as bandwidth-intensive and data-intensive ones, and study the topology configuration schemes for them, respectively.

Bandwidth-intensive applications mostly exist in tenant-oriented commercial DCs, and their main QoS demands are on bandwidth capacity. Specifically, such an application usually has a traffic matrix to describe the least bandwidth that it needs to occupy among the racks. Therefore, operators usually want to serve these applications with the minimum operational cost (*e.g.*, the least number of active ports). Fig. 3(a) explains how an improper topology design can make traffic routing of

a bandwidth-intensive application infeasible. Here, there are three demands among the ToR switches, as 1→2, 1→3, and 2→3 for the bandwidth of 15, 5, and 5 units, respectively, and the capacity of each port assumed to be 10 units. Then, if the topology is designed as the left one in Fig. 3(a) (*i.e.*, the established connections in the Hyper-FleX-LION are marked as black arrows), the demand of 1→2 cannot be satisfied. On the other hand, the right configuration in Fig. 3(a) can serve all the demands properly. Although the two configurations in Fig. 3(a) activate same number of ports (*i.e.*, 3), their supports to the bandwidth-intensive application are different.

Fig. 3(b) explains why proper traffic routing helps to explore the bandwidth in a specific topology design. This time, the demands are still for 1→2, 1→3, and 2→3, but their bandwidth requirements become 5, 11, and 5 units, respectively. Then, to support the bandwidth requirements in the left topology in Fig. 3(b), we need to activate three ports on *ToR* 1, while the right topology only requires two active ports. Hence, the traffic routing in the right topology helps to save one active port.

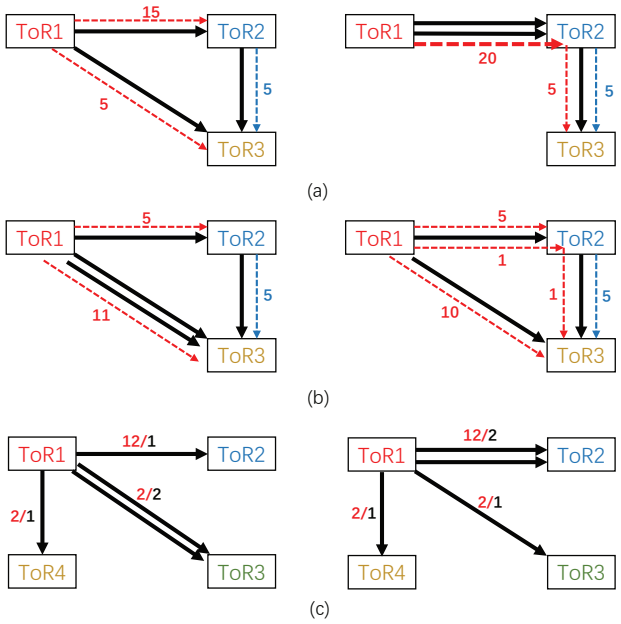

Fig. 3. Examples on correlation of topology design and traffic routing for (a) and (b) bandwidth-intensive applications, and (c) data-intensive application.

On the other hand, data-intensive applications need to transfer certain amounts of data among racks, and such an application is completed only when all of its data has been transferred. Therefore, operators need to plan the topology reasonably to complete all the data transfers as soon as possible. Fig. 3(c) shows a straightforward example to explain why topology management is also important for these applications. Here, we consider an application that needs to transfer data from *Rack* 1 to the other three racks, and the amounts of data that needs to be transferred are 1→2: 12 units, 1→3: 2 units, and 1→4: 2 units, respectively. Here, each black arrow denotes an optical connection with a bandwidth capacity of one unit, and the number on it shows the duration of the data transfer that uses it. Therefore, according to the left topology configuration in Fig. 3(c), the data transfer of 1→2 will become the bottleneck of the application, *i.e.*, it takes $\frac{12}{1} = 12$ time-units. But if we use the right topology configuration in Fig. 3(c), we can reduce the completion time of the data transfer to $\frac{12}{2} = 6$ time-units without activating more ports. To this end, we can see that a well-designed topology configuration mechanism together with adaptive traffic routing can make Hyper-FleX-LION accelerate both bandwidth- and data-intensive applications better with less operational cost.

IV. Summary

This paper summarized our recent research activities on AOIs in Hyper-FleX-LION. We first proved that Hyper-FleX-LION can provide a better acceleration effect than HOE-w/OXC when serving DML. Then, we introduced the acceleration effects brought by Hyper-FleX-LION on bandwidth- and data-intensive applications, and explained why we should consider the application-aware topology configuration of Hyper-FleX-LION for enhancing the QoS of applications.

Acknowledgments

This work was supported by NSFC project 61871357 and Fundamental Fund for Central Universities (WK3500000006).

References

[1] "Cisco Annual Internet Report (2018-2023)," *Online White Report*. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html.

[2] P. Lu *et al.*, "Highly-efficient data migration and backup for Big Data applications in elastic optical inter-datacenter networks," *IEEE Netw.*, vol. 29, pp. 36–42, Sept./Oct. 2015.

[3] W. Lu *et al.*, "AI-assisted knowledge-defined network orchestration for energy-efficient data center networks," *IEEE Commun. Mag.*, vol. 58, pp. 86–92, Jan. 2020.

[4] Z. Zhu, W. Lu, L. Zhang, and N. Ansari, "Dynamic service provisioning in elastic optical networks with hybrid single-/multi-path routing," *J. Lightw. Technol.*, vol. 31, pp. 15–22, Jan. 2013.

[5] M. Zhang *et al.*, "Bandwidth defragmentation in dynamic elastic optical networks with minimum traffic disruptions," in *Proc. of ICC 2013*, pp. 3894–3898, Jun. 2013.

[6] W. Shi, Z. Zhu, M. Zhang, and N. Ansari, "On the effect of bandwidth fragmentation on blocking probability in elastic optical networks," *IEEE Trans. Commun.*, vol. 61, pp. 2970–2978, Jul. 2013.

[7] L. Gong *et al.*, "Efficient resource allocation for all-optical multicasting over spectrum-sliced elastic optical networks," *J. Opt. Commun. Netw.*, vol. 5, pp. 836–847, Aug. 2013.

[8] Y. Yin *et al.*, "Spectral and spatial 2D fragmentation-aware routing and spectrum assignment algorithms in elastic optical networks," *J. Opt. Commun. Netw.*, vol. 5, pp. A100–A106, Oct. 2013.

[9] N. Farrington *et al.*, "Helios: a hybrid electrical/optical switch architecture for modular data centers," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 40, pp. 339–350, Oct. 2010.

[10] K. Chen *et al.*, "OSA: An optical switching architecture for data center networks with unprecedented flexibility," *IEEE/ACM Trans. Netw.*, vol. 22, pp. 498–511, Apr. 2013.

[11] H. Ballani *et al.*, "Sirius: A flat datacenter network with nanosecond optical switching," in *Proc. of ACM SIGCOMM 2020*, pp. 782–797, Jul. 2020.

[12] J. Benjamin *et al.*, "PULSE: Optical circuit switched data center architecture operating at nanosecond timescales," *J. Lightw. Technol.*, vol. 38, pp. 4906–4921, May 2020.

[13] G. Liu *et al.*, "Architecture and performance studies of 3D-Hyper-FleX-LION for reconfigurable All-to-All HPC networks," in *Proc. of SC 2020*, pp. 1–16, Nov. 2020.

[14] H. Yang, Z. Zhu, R. Proietti, and B. Yoo, "Which can accelerate distributed machine learning faster: Hybrid optical/electrical or optical reconfigurable DCN?" in *Proc. of OFC 2022*, pp. 1–3, Mar. 2022.
[15] X. Xiao *et al.*, "Multi-FSR silicon photonic Flex-LIONS module for bandwidth-reconfigurable all-to-all optical interconnects," *J. Lightw. Technol.*, vol. 38, pp. 3200–3208, Mar. 2020.
[16] ——, "Silicon photonic Flex-LIONS for bandwidth-reconfigurable optical interconnects," *IEEE J. Sel. Top. Quantum Electron.*, vol. 26, pp. 1–10, Nov. 2020.
[17] Z. Zhu *et al.*, "Demonstration of cooperative resource allocation in an OpenFlow-controlled multidomain and multinational SD-EON testbed," *J. Lightw. Technol.*, vol. 33, pp. 1508–1514, Apr. 2015.
[18] S. Li *et al.*, "Protocol oblivious forwarding (POF): Software-defined networking with enhanced programmability," *IEEE Netw.*, vol. 31, pp. 58–66, Mar./Apr. 2017.
[19] N. Bitar, S. Gringeri, and T. Xia, "Technologies and protocols for data center and cloud networking," *IEEE Commun. Mag.*, vol. 51, pp. 24–31, Sept. 2013.
[20] L. Gong and Z. Zhu, "Virtual optical network embedding (VONE) over elastic optical networks," *J. Lightw. Technol.*, vol. 32, pp. 450–460, Feb. 2014.
[21] W. Fang *et al.*, "Joint spectrum and IT resource allocation for efficient vNF service chaining in inter-datacenter elastic optical networks," *IEEE Commun. Lett.*, vol. 20, pp. 1539–1542, Aug. 2016.
[22] L. Gong, Y. Wen, Z. Zhu, and T. Lee, "Toward profit-seeking virtual network embedding algorithm via global resource capacity," in *Proc. of INFOCOM 2014*, pp. 1–9, Apr. 2014.
[23] Q. Sun, P. Lu, W. Lu, and Z. Zhu, "Forecast-assisted NFV service chain deployment based on affiliation-aware vNF placement," in *Proc. of GLOBECOM 2016*, pp. 1–6, Dec. 2016.
[24] L. Gong, H. Jiang, Y. Wang, and Z. Zhu, "Novel location-constrained virtual network embedding (LC-VNE) algorithms towards integrated node and link mapping," *IEEE/ACM Trans. Netw.*, vol. 24, pp. 3648–3661, Dec. 2016.
[25] J. Liu *et al.*, "On dynamic service function chain deployment and readjustment," *IEEE Trans. Netw. Serv. Manag.*, vol. 14, pp. 543–553, Sept. 2017.
[26] M. Abadi *et al.*, "Tensorflow: A system for large-scale machine learning," in *Proc. of OSDI 2016*, pp. 265–283, Nov. 2016.
[27] C. Wang *et al.*, "Acceleration and efficiency warranty for distributed machine learning jobs over data center network with optical circuit switching," in *Proc. of OFC 2021*, pp. 1–3, Jun. 2021.