HoeSDCN: A Hybrid Optical/Electrical Switching DCN with Dynamic Bandwidth Allocation

Xiongfei Ren
State Key Laboratory of Information Photonics and Optical Communications
Beijing University of Posts and Telecommunications
Beijing, Ching
xfren@bupt.edu.cn

Xuwei Xue
State Key Laboratory of Information Photonics and Optical Communications
Beijing University of Posts and Telecommunications
Beijing, Ching
x.xue@bupt.edu.cn

Yuanzhi Guo
State Key Laboratory of Information Photonics and Optical Communications
Beijing University of Posts and Telecommunications
Beijing, Ching
guoyuanzhi@bupt.edu.cn

Yisong Zhao
State Key Laboratory of Information Photonics and Optical Communications
Beijing University of Posts and Telecommunications
Beijing, Ching
zhaoyisong@bupt.edu.cn

Changsheng Yang
State Key Laboratory of Information Photonics and Optical Communications
Beijing University of Posts and Telecommunications
Beijing, Ching
ycs25@bupt.edu.cn

Bingli Guo
State Key Laboratory of Information Photonics and Optical Communications
Beijing University of Posts and Telecommunications
Beijing, Ching
guobingli@bupt.edu.cn

***Abstract*—A Hybrid Optical/Electrical Switching DCN named HoeSDCN is proposed with the capability of dynamic bandwidth scheduling. Numerical investigations validate that HoeSDCN decreases 22.23% packet loss and 32.8% latency, compared with the static bandwidth allocation scheme.**

Keywords—optical switching, optical data center network, arrayed waveguide grating

I. INTRODUCTION

With the mass deployment of cloud computing, streaming media, and 5G services and applications, the rapid increase in traffic in data center networks (DCNs) put tremendous pressure on the electrical switch nodes and network topology. However, due to the inability to increase the pin density on the Ball Grid Array (BGA) packaging technique, current electrical switches are expected to hit the bandwidth bottleneck two generations from now [1]. Moreover, the electrical switches-based network topologies built in a multi-tier tree-like style feature high traffic completion time resulting from the extensive path diversity. Compared to electrical switches, optical switches benefitting from the data rate and format transparency, can provide much higher bandwidth to overcome the bandwidth bottleneck of electrical switches effectively. Furthermore, the higher bandwidth of the optical switches allows flatting the network topology to decrease the traffic transmission latency. Academic and industry communities have extensively investigated various optical switches-based DCN structures, such as Sirius, ReSAW, and LIONS [2-4], which verifies the potential of the optical switches for practical deployment.

Despite the promises held by the optical switching technique, there are still several challenges that need to be addressed to practically deploy optical switches in DCNs. First, to fully utilize the nanoseconds-level hardware switching time, a fast control mechanism is required to configure the switch in nanoseconds time scale to fast forward the packets[5]. Second, as no effective buffer exists in the optical domain, the conflicted packets at the optical switch would be dropped and this results in high packet loss. Third, in optically switched network, new physical connections are created every time the switch configuration changes. This implies that the receiver has to continuously adjust the local clock to properly sample the incoming packets and recover the data. The longer this process takes, the lower the network throughput will be. Due to the aforementioned challenges, the optical switching techniques especial for optical packet switch still need further investigations. Therefore, the hybrid optical/electrical switching system is with more potential for practical deployment so far, where optical switches are deployed to carry the elephant traffic flow while the electrical switches are used to forward the mice flow.

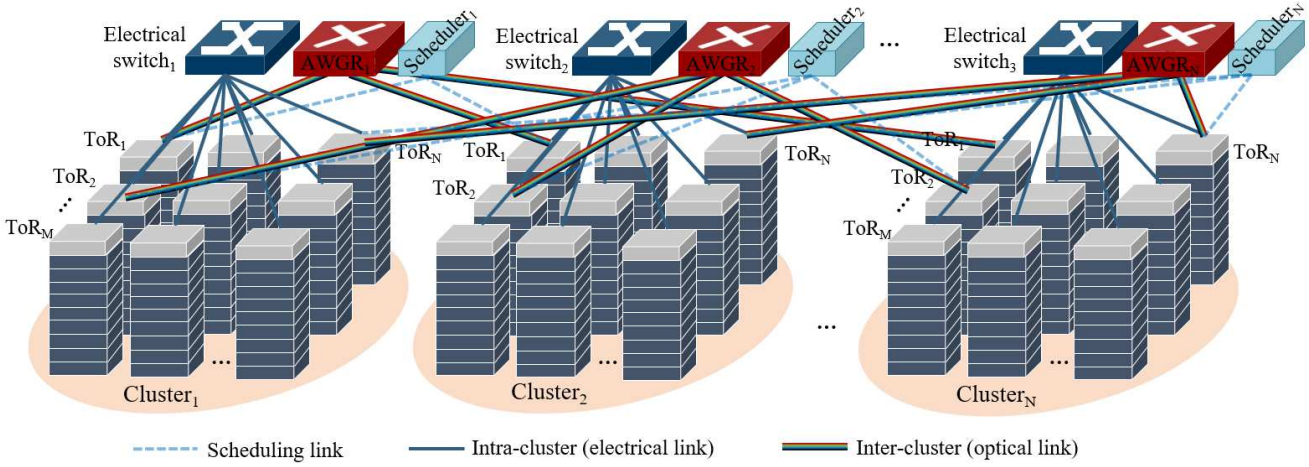

Fig. 1: The structure of the proposed HoeSDCN.

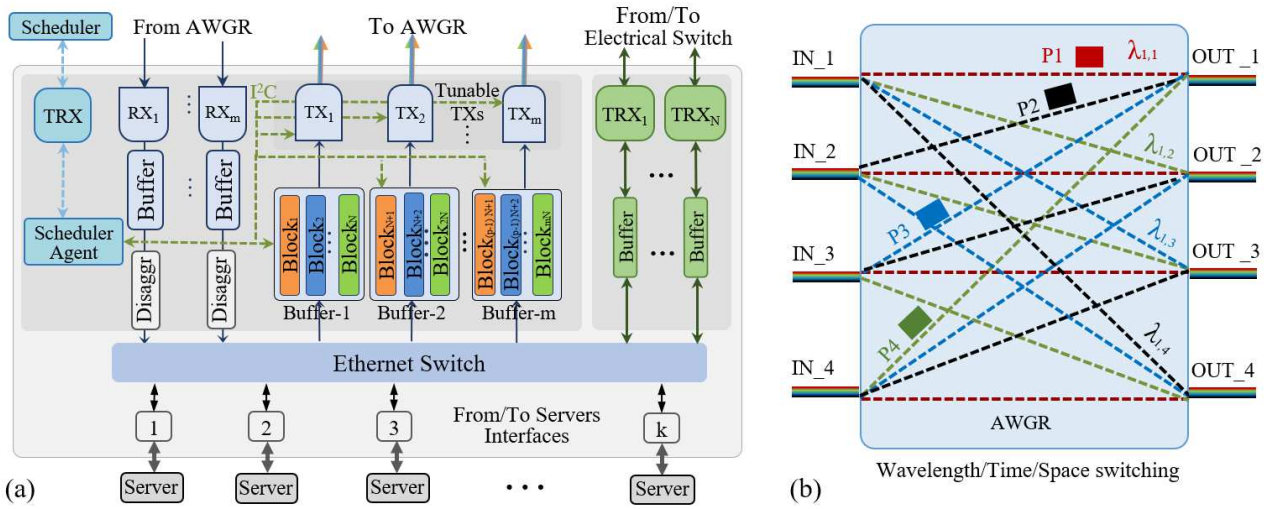

Fig. 2: (a) Schematic of tunable ToR (b) Switching traffic at the AWGR in the wavelength/time/space domain.

In this paper, we propose a hybrid optical/electrical switching DCN named HoeSDCN, where the arrayed waveguide grating router (AWGR), as a passive optical component, is used to interconnect the inter-cluster racks. Combing with the tunable laser at each top of rack (ToRs), the AWGR can optically switch inter-cluster traffic at nanoseconds magnitude. Under the arrangement of Scheduler, the optical bandwidth can be flexibly provisioned by allocating the various magnitude in the wavelength, time and space domains. The electrical switches are used to switch the intra-cluster traffic where many applications produce short data packets.

II. Principle of HoeSDCN

A. Structure of HoeSDCN

The structural diagram of HoeSDCN is shown in Fig. 1(a). The k servers in each rack are interconnected through the ToR, and N racks are grouped in each cluster. HoeSDCN divides the network into N clusters. The server-generated traffic (Ethernet frames) is classified into three types (intra-rack, intra-cluster, and inter-cluster) based on the frame's destination. The Ethernet frames are first processed at the Ethernet switch of each ToR, where the frame of intra-rack traffic will be directly forwarded to servers in the same rack. The ToRs in different cluster are connected to the AWGR through wavelength division multiplexing bidirectional optical links to realize inter-cluster communications. The intra-cluster is interconnected via optical switches. The i^{th} AWGR interconnects the i^{th} ToR of each cluster, with $i = 1, ..., N$, so the inter-cluster communications between any racks only needs two hops. More than 70% elephant traffic is forwarded between the cluster, the AWGR optical interconnections are then required to provide higher bandwidth than the electrical switches for processing intra-cluster mice traffic. The corresponding scheduler for each AWGR interconnections is connected to each ToR for packet contention resolution and bandwidth assignment by elastically allocating wavelength and timeslots.

B. Schematic of ToR

Fig.1(b) illustrates the schematic of the Field Programmable Gate Array (FPGA) implemented ToR. The ToR deploys m buffer groups that connect to the m tunable transmitters (TXs) equipping N wavelengths to connect the AWGR. One group consists of N buffer blocks and each block stores the packets with the same destination. At each timeslot, the tunable TX is controlled to generate one wavelength to forward the traffic stored in the corresponding buffer block. At the end of each timeslot, the Scheduler Agent will send the monitored traffic statics of each buffer block, including the source and destination of the packets and the block occupation ratio, to the Scheduler through the optical link. The scheduling algorithm deployed at the scheduler will select the traffic at the most-occupied buffer block to be forwarded under the premise of no packet contention. Before the start of the next timeslot, the granted scheduling commands will be sent back to the corresponding ToRs.

C. Flexible Scheduling

Based on the received traffic information from ToRs, the Scheduler first sorts the buffer blocks to a list according to their occupation ratio following the principle from high to low. The traffic in the most-occupied buffer blocks at each ToR has the highest priority to be selected and forwarded for the next time slot. Because the Scheduler knows all the packet-forwarding requests in this subnetwork, the traffic in the next buffer block in this sorted list will be forward when packet contention happens for the currently selected blocks. The scheduling algorithm thus prevents packet contention and guarantees the fast traffic release for buffer block to avoid buffer overflow causing packet loss. Compared with the static scheduling scheme, the flexible scheduling scheme provides dynamic bandwidth by allocating switching the traffic in the time, wavelength and space domain at the AWGR as shown in Fig. 2(b). Therefore, more traffic can be forwarded in this proposed solution which greatly improves bandwidth utilization.

In the typical "Request-Grant" scheduling scheme, the traffic scheduling and forwarding steps are executed at the same timeslot. This scheme could cause extra traffic storing delays because the execution of the forwarding step has to wait for the completion of the scheduling step. To avoid this typical scheduling style caused extra traffic waiting delay, a "Dislocation Pipeline" (DisP) strategy is designed in this HoeSDCN structure that advances the scheduling step to the previous timeslot and then schedules the packets in real-time. The DisP strategy has two steps. First, the Scheduler Agent

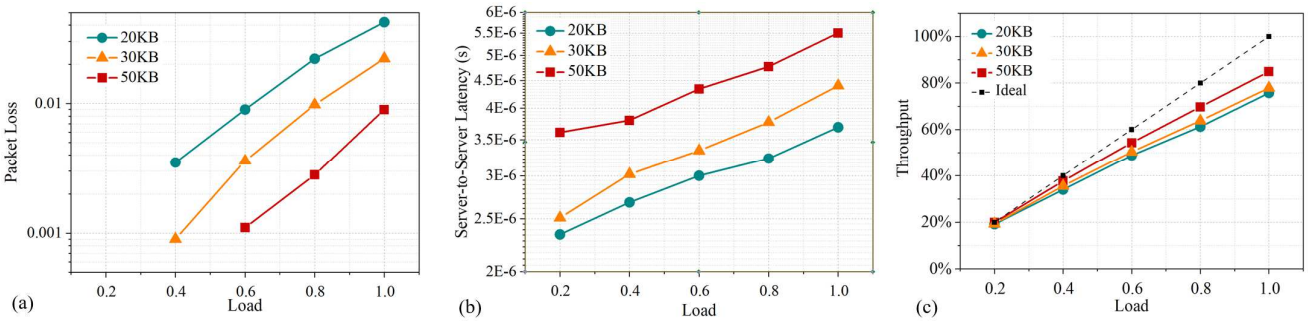

Fig. 3: Network performance of (a) Packet loss (b) Server-to-Server latency (c) Throughput.

sends the traffic statistics of each buffer block at the end of the N^{th} timeslot to the Scheduler to request the corresponding wavelength for the $N+1^{th}$ timeslot. Seconds, the Scheduler implements the aforementioned algorithm based on scheduling requests to arrange adaptable wavelength for traffic in each most-occupied buffer block of the $N+1^{th}$ timeslot under the premise of no contention.

III. Performance Investigations and Discussions

An OMNeT++ simulation model is programmed to numerically investigate the network performance of the proposed HoeSDCN structure. The traffic ratio of intra-rack, intra-cluster and inter-cluster in this numerical model are set to 50%, 37.5% and 12.5%, respectively. Each rack groups 40 servers and the link rate of servers is 20 Gb/s. To maintain the same oversubscription (1:1) with the experimental setup, 8 WDM tunable transceivers operating at 100 Gb/s are deployed at each ToR. The size of buffer block in each ToR can be flexibly allocated based on the network traffic load. The servers are programmed to generate Ethernet frames with the length varying from 64 bytes to 1518 bytes, under a carriable traffic load from 0 to 1. The Ethernet frames are randomly destined to one of the possible servers at each timeslot. The buffer size of this model is set as 20KB, 30KB, 50KB, respectively.

Fig.3 shows the numerical performance of the HoeSDCN structure with the present DisP strategy. For the higher traffic load, the buffer blocks at the ToR will be easier to fill out. Moreover, the heavy traffic load will enhance the traffic burst feature, which indicates the traffic sent to different destinations is from various sources most time. For the network deploying smaller buffer, such as 20KB, the buffer blocks that has more traffic will lose packets. As a comparison, the larger buffer with enhanced DisP strategy ensures that traffic in the most-occupied buffer blocks can be scheduled in time under the premise of no contention, so that the packet loss is much lower. Fig.3 (a) illustrates that the packet loss of HoeSDCN can be reduced significantly for buffer size of 50KB under the DisP scheme. The packet loss of network with 50KB buffer block decreases about 50% than the network with 20KB buffer because the smaller buffer is hard to store the burst traffic with random destinations. The most-occupied buffer for the case of 50KB can be released in time so that the traffic queuing time is lower. However, due to the data storing delay, the HoeSDCN deploying 50KB buffer, compared with the network without 30KB and 20KB buffer as shown in Fig. 3(b), the server-to-server latency increases by 23.0% and 30.9% at the load of 0.6, respectively. Benefitting from the lower packet loss, the throughput of HoeSDCN network with 50KB buffer improves by 12.26 % at traffic load of 0.8, as shown in Fig. 3(c), compared with the network deploying 20KB buffer.

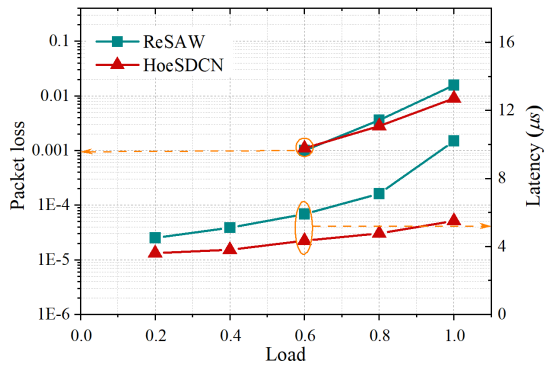

Fig. 4: Comparison with ReSAW structure in terms of packet loss and Server-to-Server latency.

The HoeSDCN structure deploying DisP scheme is also compared with ReSAW, which is a hybrid optical/electrical switching DCN with rigid optical bandwidth allocation [6], to investigate the network performance. The comparison in terms of Server-to-Server latency and packet loss is illustrated in Fig. 4. The latency and packet loss of ReSAW structure is higher due to the static network scheduling caused bandwidth over-provision or competition, while the presented HoeSDCN strategy featuring flexible bandwidth allocation reduces 32.8% latency and 22.23% packet loss when the load is 0.8.

IV. Conclusions

To fully utilize the high bandwidth of optical switches and the flexibility of the electrical switches, a hybrid optical/electrical switching DCN named HoeSDCN is proposed in which the designed DisP scheduling scheme can dynamically allocate the optical bandwidth to adapt the burst network traffic. Numerical investigations validate that the HoeSDCN structure decreases 22.23% packet loss and 32.8% latency, compared with the typical ReSAW network deploying static bandwidth allocation scheme.

Acknowledgment

This work was partially supported by the National Natural Science Foundation of China (62101065, 62220106002, 62125103, 62171059) and National Key Research and Development Program of China (2018YFB1801702).

REFERENCES

[1] H. Ballani, P. Costa, I. Haller, K. Jozwik, K. Shi, B. Thomsen, and H. Williams, "Bridging the last mile for optical switching in data centers," in *2018 Optical Fiber Communications Conference and Exposition (OFC)*, 2018, pp. 1-3: IEEE.

[2] H. Ballani, P. Costa, R. Behrendt, D. Cletheroe, I. Haller, K. Jozwik, F. Karinou, S. Lange, K. Shi, and B. Thomsen, "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*, 2020, pp. 782-797.

[3] Z. Zhao, X. Xue, B. Guo, Y. Zhao, X. Zhang, Y. Guo, W. Ji, R. Yin, B. Chen, S. Huang, and Networking, "ReSAW: a reconfigurable and picosecond-synchronized optical data center network based on an AWGR and the WR protocol," *Journal of Optical Communications Networking,* vol. 14, no. 9, pp. 702-712, 2022.

[4] Y. Yin, R. Proietti, X. Ye, C. J. Nitta, V. Akella, and S. J. I. J. o. S. T. i. Q. E. Yoo, "LIONS: An AWGR-based low-latency optical switch for high-performance computing and data centers," vol. 19, no. 2, pp. 3600409-3600409, 2012.

[5] X. Xue and N. Calabretta, "Nanosecond optical switching and control system for data center networks," *Nature communications,* vol. 13, no. 1, p. 2257, 2022.

[6] Z. Zhao, X. Xue, B. Guo, Y. Zhao, X. Zhang, Y. Guo, W. Ji, R. Yin, B. Chen, S. J. J. o. O. C. Huang, and Networking, "ReSAW: a reconfigurable and picosecond-synchronized optical data center network based on an AWGR and the WR protocol," vol. 14, no. 9, pp. 702-712, 2022.