Experimental Assessments of a Flexible Optical Switch and Control System with Dynamic Bandwidth Allocation

Xuwei Xue[1], Kristif Prifti[1], Bitao Pan[1], Sai Chen[2], Xiaotao Guo[1], Fulong Yan[2], Shaojuan Zhang[1], Chongjin Xie[3] and Nicola Calabretta[1]

[1] Electro-Optical Communications Group, Eindhoven University of Technology, Eindhoven 5612 AZ, Netherlands, x.xue.1@tue.nl
[2] Alibaba Cloud, Alibaba Group, Hangzhou, China
[3] Alibaba Cloud, Alibaba Group, Sunnyvale, CA, USA.

Abstract *A flexible optical switch and control system with dynamic bandwidth allocation is experimentally assessed to enable reconfigurable DCNs. Experiments validate the flexible system achieves 3.15 µs end-to-end latency (improving 48.61%) and decreases one order magnitude of packet loss, compared with the scheme with fixed connections.*

Introduction

Recently, optically switching of datacenter traffic has been considerably investigated as a future-proof solution supplying ultra-high bandwidth to overcome the bandwidth bottleneck issue of electrical switches [1]. Leveraging various optical switching techniques, several scenarios of optical data center networks (DCNs) have been proposed, such as OPSquare, LIONS and OSA [2]. In all these optically switched architectures, the optical bandwidth between any top of racks (ToRs) is fixed because the bandwidth is determined by the amounts of the pre-deployed transceivers (TRXs). This means that the optical bandwidth between links cannot be reallocated on-demand to adapt the variable data center (DC) traffic volume. Only a few links between ToRs, as reported in [3], require high bandwidth in a certain time, while most bandwidth of other links is underutilized. Additionally, the bandwidth requirements between ToRs are also dynamically varying as the hosted applications and services switchover. Thus, the fixed bandwidth allocation appears to be insufficient or overprovisioned for the DC applications, even for the optically switched network with high capacity.

To overcome this issue, intelligent workload-placing mechanisms have been investigated to assign network-bound application components to network infrastructure with adaptable bandwidth interconnections [4]. However, these schemes need to consider the overall network infrastructure to flexibly allocate workload placement, significantly increasing the complexity of network management and control, particularly for large-scale DCNs.

In this work, a flexible optical switch and control system based on a novel optical ToR exploiting a wavelength selective switch (WSS) to dynamically allocate the optical bandwidth is proposed and experimentally assessed to adapt the variable DC traffic volume. Based on the monitored traffic statistics of the network, the software-defined networking (SDN) control plane can dynamically configure each ToR in order to deploy a variable optical bandwidth per optical link. Experimental results confirm a low packet loss and 3.15 µs server-to-server latency at 0.5 traffic load.

Flexible optical switch and control system

The proposed switch and control system is demonstrated in Fig. 1(a). The SOA-based fast optical switches including the FPGA-implemented switch controller as well as flexible FPGA-implemented optical ToRs with dedicated optical interfaces have been developed to fully support the reconfigurable operations. The label signals indicating the destination information of the optical data packets are sent to the switch

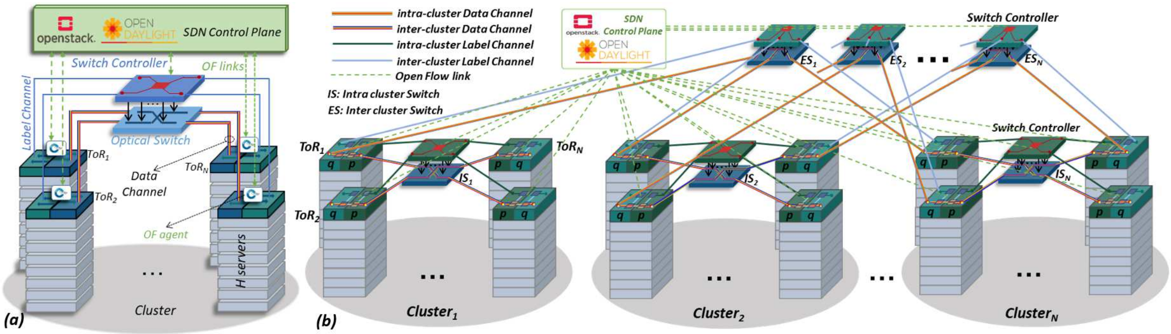

Fig. 1: (a) Flexible optical switch and control system. (b) Proposed system based optical data center network.

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

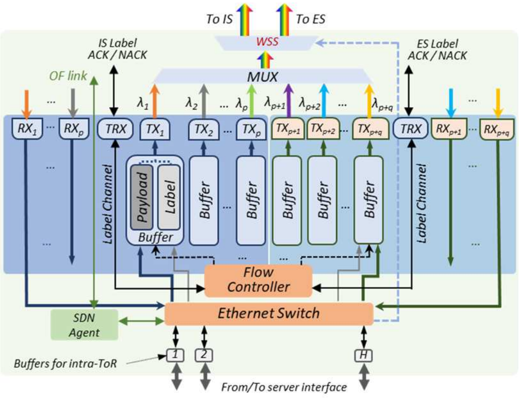

Fig. 2: Schematic of reconfigurable optical ToR.

controller which thereby control the forwarding of the corresponding data packets at the SOA switch. To automatically allocate the multi TRXs at ToRs and to configure the optical switches, the OpenDaylight (ODL) are deployed as the SDN controller connecting ToRs and switch controllers by means of SDN-agents and the extended OpenFlow (OF) links [5]. The SDN-agents gather the monitored traffic statistics from the FPGA-based ToRs and report them to the SDN controller for further processing.

Large-scale optical DCNs can be built based on this proposed switch and control system as structurally shown in Fig. 1(b), where inter-cluster optical switch (ES) is employed to connect ToRs locating in different clusters, while intra-cluster optical switch (IS) is dedicated for connecting ToRs inside the same cluster. To support this large-scale network, *p*+*q* transceivers (TRXs) are deployed at each ToR, where each TRX equips dedicated electrical buffer. As shown in Fig. 2, *p* TRXs are used to connect the ToR with the IS for intra-cluster communication, while other *q* TRXs are utilized for inter-cluster communications by connecting ToRs to the ES. The traffic generated by servers is classified into three categories (intra-ToR, intra-cluster and inter-cluster). The three kinds of traffic are firstly processed by the Ethernet switch at each ToR. Based on the destination MAC address, the frames of the intra-ToR traffic are directly forwarded to the servers locating in the same rack. As to the intra-cluster (IC) and inter-cluster (EC) traffic, the Ethernet switch forwards the frames to the electrical buffer of corresponding *p* and *q* transmitter (TXs). The ODL based SDN controller can monitor the traffic volume ratio of the IC and EC traffic. The number of *p* and *q* are then flexibly adjusted under the control of SDN controller, based on the desired bandwidth requirement of IC and EC traffic. In details, the SDN send OpenFlow commands to update the MAC address look-up table (LUT) of the Ethernet switch and thereby altering the forwarding ports of the Ethernet frames to different buffers. Meanwhile, the FPGA-based optical ToR configures the wavelength selective switch (WSS) to accordingly assign these *p* and *q* wavelength to corresponding outputs of WSS, connecting with the IS and ES, respectively.

Experimental demonstration and results

As illustrated in Fig. 3(a). The setup used to evaluate the proposed system consists of 6 FPGA-based ToRs, in which 3 ToRs (ToR$_1$, ToR$_2$ and ToR$_3$) are equipped with 4 (*p*+*q* = 4) 10 Gb/s WDM TRXs. 4 ToRs (ToR$_1$, ToR$_4$, ToR$_5$ and ToR$_6$) connecting to the EC are utilized to generate the packet contention and then to evaluate the practical network scenario. 2 6x4 mm^2 fabricated 4×4 photonic switch chip [6] integrating 4 optical modules is set in this setup as the IS and ES switch for inter-cluster and intra-cluster connections. The ODL based SDN controller connects the FPGA implemented ToRs and switch controller via the OF links and OF agents. The SPIRENT Ethernet Test Center generates Ethernet frames with controllable and variable traffic load, emulating 24 servers at 10 Gb/s.

First, the bit error rate (BER) performance is investigated to quantify the possible signal degradation caused by the switch chip. As shown in Fig. 3(b), error-free operations with less than 0.5 dB penalty have been measured at BER of 1E-9 for Channel 1 (CH 1) and Channel 3 (CH 3) at 10 Gb/s data rates for the case of single-channel input. The BER results indicate slight performance degradation for WDM channels with an extra penalty of around 0.5 dB for CH1 and CH 3 and 1 dB for CH 2 and CH 4 compared with

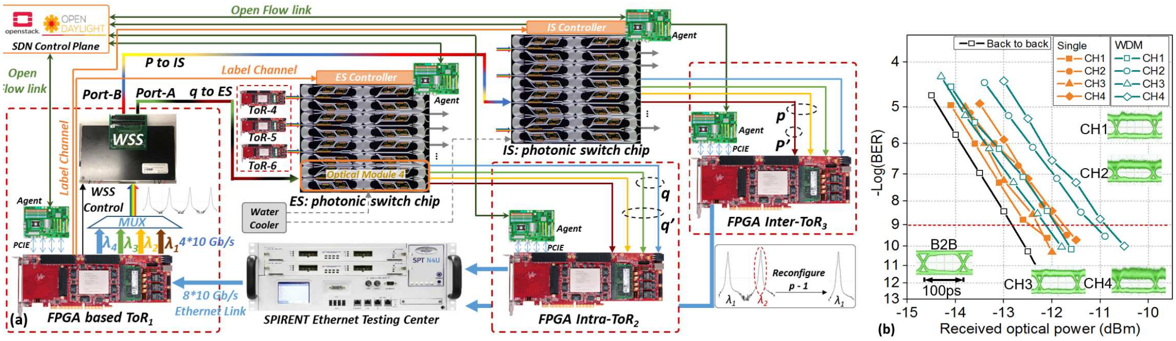

Fig. 3: (a) Experimental set-up of the optical switch and control system. (b) BER curves of the photonic switch chip.

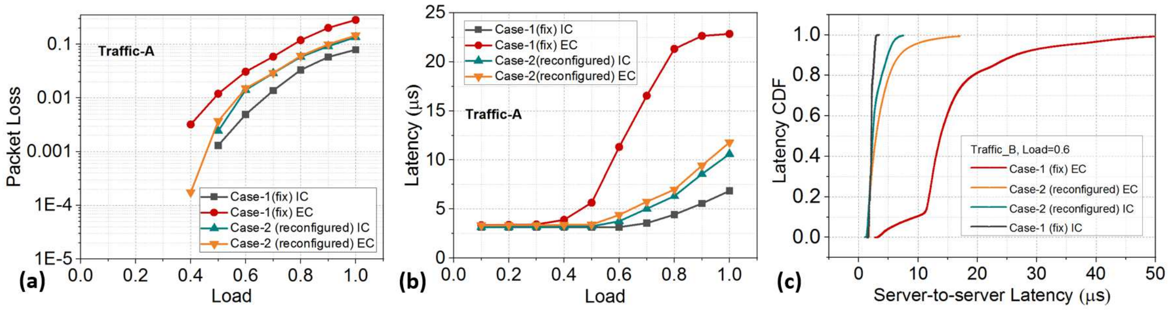

Fig. 4: Network performance (a) packet loss (b) latency (c) latency CDF for various bandwidth configurations.

single wavelength operations.

Traffic-A consisting of average 50% intra-ToR, 15% IC and 35% EC traffic is generated in this experiment by adjusting the MAC address of Ethernet frames in SPIRENT. At start, the initial configurations (Case-1) allocates the $\lambda_{1,\,2}$ (p = 2) (overprovisioned) for the IC traffic, and the $\lambda_{3,\,4}$ (q = 2) (unsufficient) for the EC traffic, respectively. The SDN controller monitors the traffic volume and sends OF commands to ToR$_1$ (by updating LUT) to allocate more bandwidth for EC traffic. Benefiting from this fdynamic reallocation mechanism, the underutilized wavelength λ_2 is now applied to serve the EC traffic. In the new bandwidth configuration (Case-2) after reallocation, the wavelength $\lambda_{2,\,3,\,4}$ (q' = 3) are connected with ToR$_2$ providing more bandwidth to EC traffic, while the λ_1 (p' = 1) is connected with ToR$_3$ for the IC traffic.

Fig. 4(a) and (b) depict the network performance comparison between the fixed case (Case-1) and reconfigured case (Case-2). Due to the underutilized IC and overloaded EC bandwidth in the Case-1, the packet loss of EC links increases dramatically after a load of 0.4. The EC link of the Case-1 performs a packet loss of 0.012 at load of 0.5. Utilizing the bandwidth reallocation procedure, the Case-2 achieves a packet loss of 0.002 at a load of 0.5 for both the EC and IC links. Obviously, after reallocating the wavelength λ_2 to the EC link, the network of the Case-2 with adaptable bandwidth provisioning outperforms the Case-1. Meanwhile, the latency performance of Case-2 on the EC link (3.15 µs) achieves 48.61% improvements compared with the Case-1 (6.13 µs) at the load of 0.5. To analyze the latency distribution, the server-to-server latency of 40000 optical packets is counted under a load of 0.6 for Case-1 and Case-2, respectively. The Cumulative Distribution Function (CDF) of latency is shown in Fig. 4(c). It is validated that, compared with the fixed Case-1, the IC/EC link server-to-server latency of Case-2 performs low variations. The mean latency distribution of IC and EC links in the Case-2 is 3.65 µs and 4.42 µs, respectively, while the latency distribution of EC link in the Case-1 varies from 3.5 µs to 50 µs.

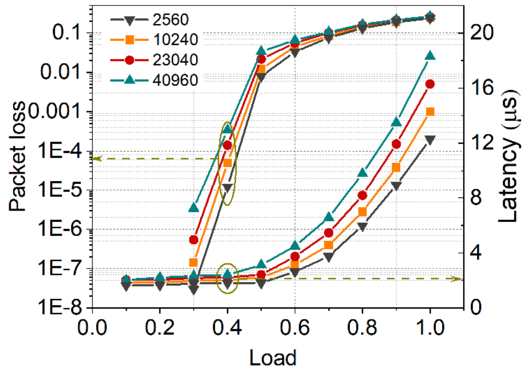

Fig. 5: Network performance of large-scale network.

Finally, we numerically investigate the network performance of the proposed switch and control system based DCNs as the network scale-out. A simulation platform based on OMNeT is built, adopting the parameters measured in experiments. As shown in Fig. 5, as the network scales from 2560 to 40960 servers, the numerical results validate only average 11% latency performance degradation. At a load of 0.3, the packet loss is less than 1E-5 and the server-to-server latency is below 3.5 µs for the large scale (40960 servers) network, which indicates the good scalability of the proposed flexible optical switch and control system.

Conclusions

We propose and experimentally evaluated a flexible SDN enabled optical switch and control system, based on novel optical ToRs deploying WSS. Enabled by the SDN controller, the dynamic optical bandwidth assignment has been implemented to adapt the variable traffic volume. At the 0.5 traffic load, the flexible system with adaptable bandwidth achieves the 3.15 µs server-to-server latency and 0.002 packet loss, which is 48.61% and one order of magnitude improvements, respectively, compared with the network with fixed interconnections. The latency CDF validates this flexible sytem featuring deterministic latency performance, with much lower latency variations (90% packets converged). The numerical investigation valditates the proposed system can support the large-scale reconfigurable network with only 11% performance deterioration as the network scale from 2560 to 40960 servers.

Acknowledgements

This work was supported by Alibaba Group through Alibaba Innovative Research Program.

References

[1] F. Testa and L. Pavesi, Optical switching in next generation data centers. Springer, 2017.

[2] X. Xue, F. Yan, K. Prifti, F. Wang, B. Pan, X. Guo, S. Zhang, and N. Calabretta, "ROTOS: A Reconfigurable and Cost-Effective Architecture for High-Performance Optical Data Center Networks," *Journal of Lightwave Technology*, vol. 38, no. 13, pp. 3485-3494, 2020.

[3] S. Kandula, J. Padhye, and V. Bahl, "Flyways to de-congest data center networks," 2009.

[4] J. Zhang, H. Huang, X. J. J. o. N. Wang, and C. Applications, "Resource provision algorithms in cloud computing: A survey," *Journal of Network Computer Applications*, vol. 64, pp. 23-42, 2016.

[5] F. Wang, F. Agraz, A. Pagès, B. Pan, F. Yan, X. Guo, S. Spadaro, and N. Calabretta, "SDN-controlled and orchestrated OPSquare DCN enabling automatic network slicing with differentiated QoS provisioning," *Journal of Lightwave Technology*, vol. 38, no. 6, pp. 1103-1112, 2020.

[6] K. Prifti, X. Xue, N. Tessema, R. Stabile, and N. Calabretta, "Lossless photonic integrated add-drop switch node for metro-access networks," *IEEE Photonics Technology Letters*, vol. 32, no. 7, pp. 387-390, 2020.