Experimental Demonstration of Optical Polling Flow Control for Contention Resolution in Optical Data Center Networks

Xuwei Xue, Fu Wang, Bitao Pan, Fulong Yan, Kristif Prifti, Xiaotao Guo, Rafael Kraemer, Shaojuan Zhang, Nicola Calabretta

IPI-ECO Research Institute, Eindhoven University of Technology, the Netherlands, x.xue.1@tue.nl

Abstract A novel Optical Polling Flow Control technique for packet-contention resolution in optical DCNs is experimentally assessed. By flexibly reducing the packet-retransmission and HOL blocking to decrease the latency and packet-loss, the proposed technique achieves 6.5E-3 packet-loss and 7.7 μs deterministic latency at 0.6 traffic load.

Introduction

Recently, optically switched data center networks (DCNs) with ultra-high bandwidth have been extensively investigated to cope with the continuous traffic growth in data centers (DCs) [1, 2]. However, in optical DCNs, packet contention occurs at the switch fabric whenever two or more packets at same time slot have the same destination. In electrical switches, the availability of electrical random-access memory (RAM) is used to solve the contention. However, there is no equivalent optical RAM for optical switches and, thus, high packet loss is one of the main issues for optically switched DCNs [3].

Considering the short distances between the top of rack (ToR) in DCs, the most feasible technique for contention resolution is based on flow control (FC) with packet retransmission mechanism, exploiting electrical RAMs at the ToRs [4]. AO-NACK, MPNACK and Optical Flow Control (OFC) [5-7] are typical FC technique, where the conflicted packets are detected and negative acknowledgement (NACK) signals are sent back to the source racks to trigger the retransmission of the conflicted packets stored in the electrical RAMs. However, the retransmission of conflicted packets in those FC techniques introduces extra (nondeterministic) end-to-end latency. Besides the retransmission issue, the head-of-line (HOL) blocking phenomenon, which is the head packet in the buffer line blocks the packets stored behind of it in the same buffer queue, could accelerate the buffer overflow and thus, cause a large number of packet loss [8]. Thus, the packets stored in the buffer, especially for the most-occupied buffer, need to be released in time to maintain a relatively low occupation ratio and thus, prevents the packet loss.

In this work, we propose and experimentally demonstrate software-defined networking (SDN) enabled Optical Polling Flow Control (OPFC) technique for packet contention resolution, decreasing packet retransmission and HOL blocking. In OPFC, a buffer schedule sequence is calculated at the SDN controller, providing a polling order to select traffics in buffer block to be sent out. Based on the polling order, the selected packets to be sent out at each ToR have different destination. If the occupation ratio of the selected buffer block is above a pre-defined threshold, the packets stored in the selected block will be transmitted, preventing packet contention (retransmission). Otherwise, to overcome the HOL caused packet loss, the packets stored in the most-occupied buffer block will be sent out.

Principle of Optical Polling Flow Control

The proposed OPFC contention resolution technique is developed at cluster scale, and thus, it can be applied to various optical switching DCNs architectures organized in cluster units. Without loss of generality, the operation of the proposed OPFC technique is demonstrated on the OPSquare topology [9], as illustrated in Fig. 1(a). Therefore, a cluster of the OPSquare architecture as shown in Fig. 1(b) has been used to show the principle of OPFC.

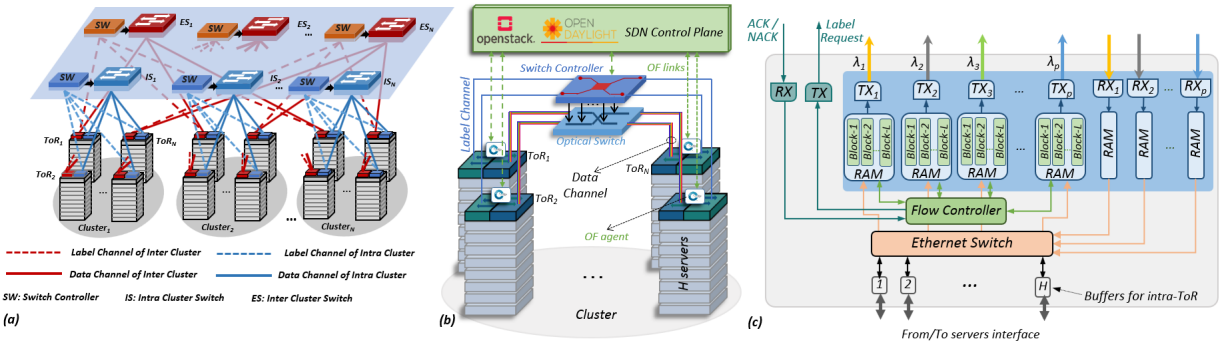

Fig. 1: (a) OPSquare DCN architecture; (b) cluster unit deploying OPFC technique, OF: Open Flow; (c) schematic of the ToR.

Ethernet frames destining to servers in different racks (inter-rack traffics) are stored in the electrical RAM at ToR (Fig. 1(c)). The frames with the same destination are stored in the same buffer block. At each time slot, only one buffer block can be selected and the copy of its stored frames will be grouped to generate an optical data packet. As shown in Fig.1 (b), the aggregated data packets are sent to the optical switch via the optical channels. Meanwhile, the label signals indicating the destination of the corresponding data packets are delivered to the switch controller via the label channels. Based on the received label signals, the switch controller checks the packet contention and accordingly configures the optical switch to forward the data packets. If the packet is successfully forwarded, an ACK signal is sent to the corresponding ToR from the controller to release the stored frames in electrical RAM. Otherwise, a NACK signal is sent back to retransmit of the conflicted packets.

The FPGA-based ToRs and switch controller can monitor in real-time the count of buffer blocks and the number of ToRs as well as traffic load, and in turn report them to the SDN control plane via the OF agents [10]. Based on this monitored information, application engines developed at the SDN controller schedule the buffer selection order for all the ToRs. The calculated buffer selection order is a polling sequence, which means the optical packets selected to be sent out at each rack are intrinsically destined to different ToRs at each time slot, thereby, preventing the packet contention. For the *i*-th ToR at time slot *t*, the calculated polling order $P'(i,t)$ is:

$$P'(i,t) = (i+t)\ mod\ (L+1) + 1 \qquad (1)$$
$$i \in [1,N], t \in [1,L]$$

L is the number of buffer blocks for each transmitter, *N* is the number of ToRs at each cluster unit. Once the ToR receives the calculated polling order, it will first check the occupation ratio of the selected buffer block. If the occupation ratio $B(P'(i,t))$ is equal to or above the predefined threshold Bs ($Bs \le B(P'(i,t))$), the final block selected order $S(i,t)$ for the *i*-th ToR at time slot *t* will be the same with the polling order $P'(i,t)$, that is

$$S(i,t) = P'(i,t) \qquad (2)$$

This is the polling order case. If all the buffer selection at each ToR follow this polling order case, there will be no contended packets at the switch (no packets need to be retransmitted) because all packets are destined to different racks. If the occupation ratio $B\big(P'(i,t)\big)$ of the buffer block selected by the polling order is less than the threshold Bs ($B\big(P'(i,t)\big) < Bs$), the final selected block $S(i,t)$ will be the j_0-th block, which has the highest buffer occupation ratio ($B(j_0,t) = Max(B(j,t))$), that is

$$S(i,t) = j_0,\quad B(j_0,t) = Max\big(B(j,t)\big) \qquad (3)$$
$$i \in [1,N], t \in [1,L], j \in [1,L]$$

The working mode from Eq. (3) is the HOL order case. If the ToR follows this HOL order case, the packets stored at the most-occupied buffer block will be sent out to alleviate the HOL blocking and then to reduce the buffer overflow caused packet loss. The threshold value Bs determines the case selection (polling order case or HOL order case). For the proposed OPFC technique, the Bs value can be dynamically reconfigured at the FPGA-based ToR, based on the monitored traffic load.

Experimental Assessments

The set-up illustrated in Fig. 2(a) is built to experimentally assess the OPFC technique. As benchmark, the network performance based on the OFC [7] and HPACR [11] techniques are also experimentally investigated. The setup consists of 4 FPGA-based ToRs and each one equips with a 10 Gb/s optical channel to deliver the optical data packets with a length of 2600 bytes. Three buffer blocks of 4096 bytes each are deployed inside the ToR. The 4×4 SOA based optical switch is deployed to interconnect these 4 ToRs. The SPIRENT Ethernet Testing Center emulating 4 servers at 10 Gb/s generates Ethernet frames

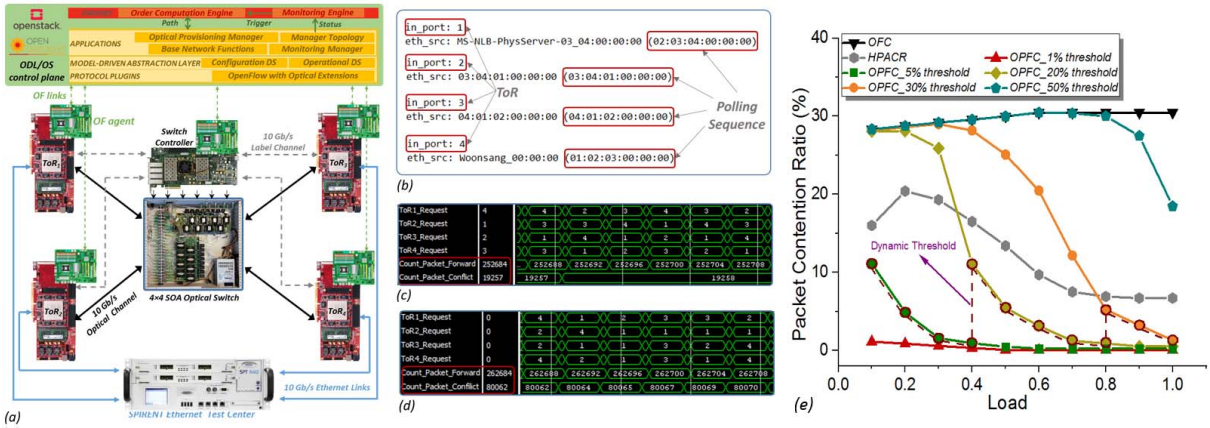

Fig. 2: (a) Experimental set-up; Monitored (b) Open Flow signals at SDN controller, label signals at switch controller of (c) OPFC scheme and (d) OFC scheme; (e) packet contention ratio for OFC, HPACR and OPFC schemes.

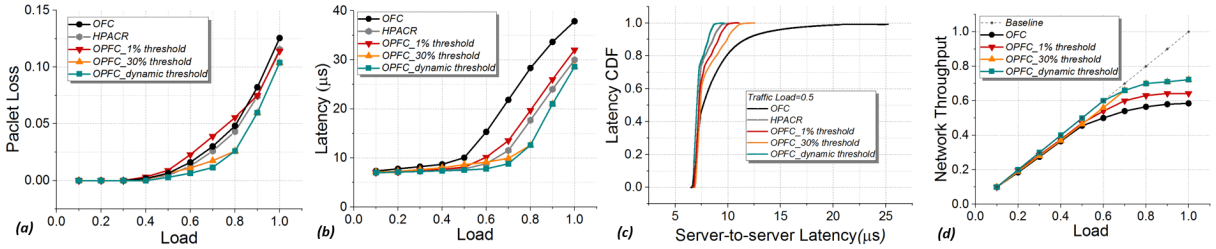

Fig. 3: Network performance in terms of (a) packet loss; (b) server-to-server latency; (c) latency CDF; (d) network throughput.

with variable load. The number of forwarded/conflicted packets is recorded at the switch controller and reported to the SDN plane.

In OPFC technique, the SDN control plane generates the polling sequences for these 4 ToRs to schedule the buffer block selection (Fig. 2(b)). As can be seen from these sequences (highlighted by the red box), there is no packet contentions at each time slot. Based on the number of forwarded packets (Fp) and conflicted packets (Cp) monitored at the switch controller, the packet contention ratio (PCR) is defined as Cp/Fp. The measured PCR of the OPFC technique is 7.62% (19257/252684) (Fig. 2(c)). For OFC scheme, the PCR is 30.48% (80062/262684) (Fig. 2(d)), which is much higher than that obtained with the OPFC technique. The PCR measured as function of the traffic load with the OFC, HPACR and OPFC techniques is shown in Fig. 2(e). The buffer threshold value determines the case selection (polling order case or HOL order case). Changing the threshold from 5% to 50% (5%, 20%, 30%, 50%, respectively), the average PCR of OPFC scheme increases because the ToRs gradually step into the HOL order case. To alleviate the HOL blocking and to decrease the PCR at the same time, the OPFC scheme dynamically reconfigure the threshold value based on the monitored traffic load. Various threshold values have been experimentally investigated. Based on the network performance, the optimal sets of the threshold with the change of traffic load is settled as follows: For the load lower than 0.4, the threshold is set as 5% and the threshold is 20% for the load between 0.4 and 0.8; when the load is higher than 0.8, the threshold value is reconfigured as 30%. The average PCR of OPFC scheme with dynamic threshold configuration is 4.7%.

The packet loss performance is then investigated as shown in Fig. 3(a). The frequent packet retransmissions in the OFC scheme block the forwarding of the packets stored behind the retransmitted packet. A large number of packets are lost for this scenario, especially for the high traffic load. The HPACR technique provides lower packet contention ratio compared with the OFC, but the lack of an effective mechanism to alleviate the HOL blocking still deteriorates the packet loss performance. With the SDN updated polling buffer selection order to reduce the packet retransmission and with the capability to alleviate the HOL blocking following the variation of traffic load, the OPFC technique with dynamic threshold configuration has the best performance in terms of packet loss. The packet loss of 6.5E-3 is achieved for dynamic threshold case at the 0.6 traffic load. Benefitting from the fewer packet contentions and the rapidly buffer releasing. The OPFC technique achieves 7.7 μs server-to-server latency at the load of 0.6 as illustrated in Fig. 3(b), improving 49.7% and 12.5%, respectively, compared with the OFC protocol (15.3 μs) and HPACR scheme (8.8 μs).

The latency Cumulative Distribution Function (CDF) is shown in Fig. 3(c) for a traffic load of 0.5. Due to the fewer packet retransmission (less unpredict latency), the server-to-server latency of OPFC based networks with dynamic threshold configuration has low variations (ranging from 6.8 μs to 9.2 μs) with respect to the mean value (7.4 μs). The network throughput performance is also investigated and shown in Fig. 3(d). With respect to the OFC and HPACR, the throughput of OPFC with dynamic buffer threshold at the load of 0.8 improves of 23.9% and 6%, respectively.

Conclusions

We have experimentally assessed a novel SDN-enabled packet contention resolution for optical DCNs. This scheme proactively prevents the contention caused time-consuming packet retransmission, and flexibly alleviates the HOL blocking caused packet loss. Based on the monitored traffic load, the buffer threshold can be dynamically reconfigured to balance the network performance. Experimental results prove the packet contention ratio of OPFC (5.7%) scheme decreases 24.42% and 6.9% with respect to the schemes of OFC (30.12%) and HPACR (12.6%), respectively. A packet loss of 6.5E-3 and the deterministic server-to-server latency of 7.7 μs is achieved for the OPFC technique at the load of 0.6. At the same load, the network throughput increases from 0.55 (HPACR) to 0.59 (OPFC).

Acknowledgements

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

References

[1] Cisco, "Forecast and Methodology 2016-2021. White Paper. Cisco Systems," ed: Inc, 2018.

[2] S. Huang, B. Guo, and Y. Liu, "5G-Oriented Optical Underlay Network Slicing Technology and Challenges," *IEEE Communications Magazine*, vol. 58, no. 2, pp. 13-19, 2020.

[3] 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," i*n 2018 Optical Fiber Communications Conference and Exposition (OFC)*, pp. 1-3: 2018.

[4] P. Andreades, "Control Plane Hardware Design for Optical Packet Switched Data Centre Networks," UCL (University College London), 2020.

[5] R. Proietti, Y. Yin, R. Yu, X. Ye, C. Nitta, V. Akella, and S. B. Yoo, "All-optical physical layer NACK in AWGR-based optical interconnects," *IEEE Photonics Technology Letters*, vol. 24, no. 5, pp. 410-412, 2011.

[6] X. Yu, H. Gu, K. Wang, M. Xu, and Y. Guo, "MPNACK: an optical switching scheme enabling the buffer-less reliable transmission," *in International Conference on Optoelectronics and Microelectronics Technology and Application*, vol. 10244, p. 102440A, 2017.

[7] W. Miao, S. Di Lucente, J. Luo, H. Dorren, and N. Calabretta, "Low latency and efficient optical flow control for intra data center networks," *Optics express*, vol. 22, no. 1, pp. 427-434, 2014.

[8] K. Keykhosravi, H. Rastegarfar, and E. Agrell, "Multicast scheduling of wavelength-tunable, multiqueue optical data center switches," *Journal of Optical Communications Networking*, vol. 10, no. 4, pp. 353-364, 2018.

[9] X. Xue, F. Yan, B. Pan, and N. Calabretta, "Flexibility assessment of the reconfigurable OPSquare for virtualized data center networks under realistic traffics," *in 2018 European Conference on Optical Communication (ECOC)*, pp. 1-3, 2018.

[10] X. Xue, 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.

[11] F. Wang, B. Liu, X. Xue, L. Zhang, F. Yan, E. Magalhães, Q. Zhang, X. Xin, and N. Calabretta, "Demonstration of SDN-Enabled Hybrid Polling Algorithm for Packet Contention Resolution in Optical Data Center Network," *Journal of Lightwave Technology*, vol. 38, no. 12, pp. 3296-3304, 2020.