Experimental Demonstration of an AWGR-based Nanoseconds Optical Switching DCN

Yuanzhi Guo, Xuwei Xue, Bingli Guo, Daohang Dang, Yisong Zhao, Rui Ding, Jiapeng Zhao, Changsheng Yang and Shanguo Huang

State Key Laboratory of Information Photonics and Optical Communications, Beijing University of Posts and Telecommunications Beijing 100876, China. x.xue@bupt.edu.cn

Abstract: An arrayed waveguide grating router based nanoseconds optical switching data center network is experimentally demonstrated and investigated. Experimental assessments validate the system achieves error-free communication with 465 ns server-to-server latency even at load of 0.9.

1. Introduction

The rapid growth of traffic volume brought by cloud computing and 5G related services has imposed a great pressure on data center networks (DCNs) [1]. To meet the pressing requirements on the DCN in terms of the low latency, high bandwidth and high power-efficiency, DCN must be scalable to support both high data rate and on-demand connected endpoints [2]. Current DCNs employ multi-tier high radix electrical switches to interconnect top-of-rack (ToR) switches and deploy multi-rooted topologies providing equal bandwidth to every endpoint. Electrical switches based DCNs require high radix of multi-stage optical to electrical to optical (O/E/O) conversions to avoid hierarchical multi-layer architectures resulting high server-to-server latency and high power consuming. Additionally, due to the limit of Ball Grid Array package technique, the problem of I/O bandwidth improvement on the electrical switching prevent the implementation of high radix electrical switch at high data rates.

Benefiting from the high bandwidth, less O/E/O conversion operations and transparent format transmission, optical switches are exploited to overcome the limitations, such as the inadequacy of the bandwidth and tremendous power consumption, of electrical networks [3]. Several optical switching technologies have been proposed to validate the feasibility in DCN, such as OPSquare [4] and FOScube [5]. OPSquare is a flat DCN architecture based on the parallel intra/inter-cluster SOA-based optical switching networks. FOScube is then extended to the three parallel interconnect, intra/inter clusters and between super-clusters. However, these schemes are microsecond-level server-to-server latency simulation verification. The experimental demonstration and assessments of nanoseconds-level server-to-server all-optical switching DCN with practical data center traffic has not carried out yet, due to the lack of customized hardware to support the fast optical switching. In this paper, a nanoseconds all-optical switching based on a novel 25Gbps FPGA-implemented top of racks (ToRs) with fixed wavelength of WDM transceivers to interconnect AWGR is proposed and experimentally assessed under variable traffic volume. Experimental results validate that switching 25Gbps data packets with error-free nanoseconds-level communication is 465 ns server-to-server latency at the load of 0.9.

2. AWGR based nanoseconds optical switching DCN

AWGR is a passive optical component with the characteristics of the cyclic wavelength routing and can realize fully connected contention-free optical network. Benefiting from the high capacity, low processing delay and insertion loss, AWGR combining with the WDM transmitters is a potential solution as the optical switches in DCNs

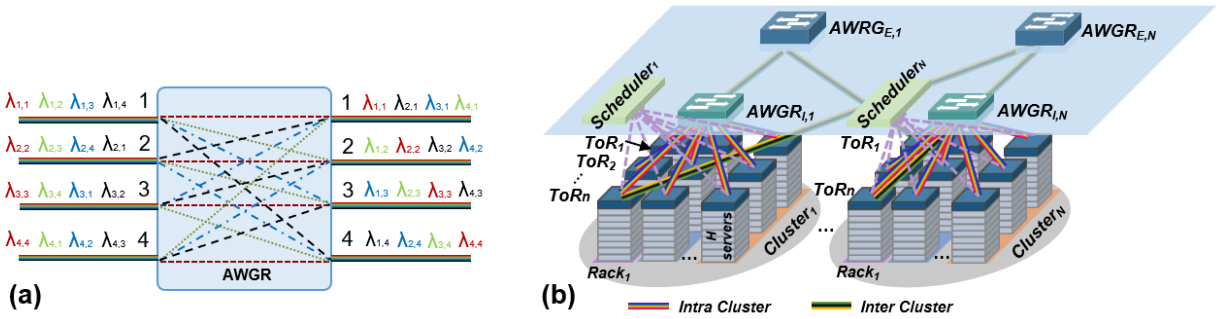

Fig. 1. (a) Switching principle of a 4x4 AWGR. (b) Structure of the AWGR-based all optical switching DCN.

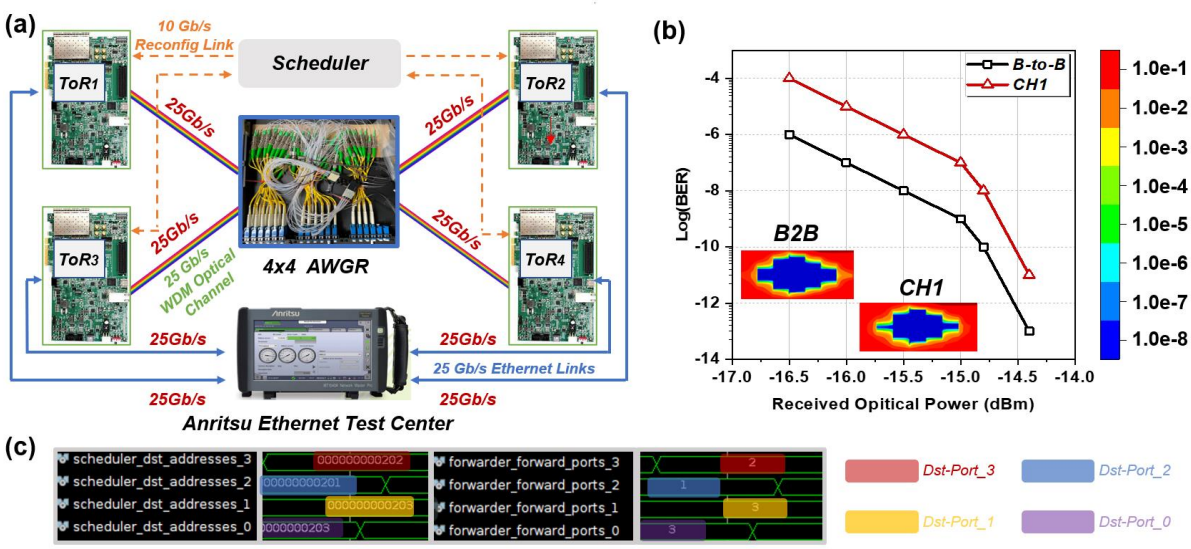

Fig. 2. (a) Experimental set-up of the AWGR based all optical intra-cluster DCN. (b) Network BER performance. (c) Signals monitored at the FPGA-implemented ToR.

compared to traditional MEMS optical switches. Fig. 1(a) shows the structure of a 4x4 AWGR. Different wavelength signals in the 4 input ports are routed to different output ports according to the cyclic wavelength routing. Combining with WDM technology, a non-blocking 16-channel communication link is realized with only 4 wavelengths.

The proposed nanoseconds all optical switching network system is demonstrated in Fig. 1(b). AWGR-based fast passive optical switches including the FPGA-implemented ToRs as well as fixed wavelength transceivers (TRXs) have been deployed to fully support the functionality of nanoseconds all optical switching. N racks coupled with H servers are aggregated into one cluster and there are N clusters in the proposed DCN system. The inter-cluster AWGR ($AWGR_E$) and intra-cluster AWGR ($AWGR_I$) are exploited to forward the inter-cluster and intra-cluster traffic, respectively. The i-th ToR in each cluster is interconnected by the i-th $AWGR_{E,i}$ ($1 \le i \le N$). Each $AWGR_I$ in cluster interconnects all ToRs via the WDM optical channels. Every cluster has a global FPGA-based scheduler ($Scheduler_N$) to reconfigure the start of the time slot and wavelength allocations, based on the collected buffer ratio of ToR and the topology information. The scheduler accordingly provides the adaptable optical bandwidth to AWGR connectivities and achieves contention-free data transmission in the optical domain. $AWGR_E$ and $AWGR_I$ connected sub-network are the same independent network in the proposed optical DCN. Therefore, in order to simplify the elaboration, the following set-up is only executed to fully elaborate the feasibility of the system within the intra-cluster when traffic is scheduled.

3. Experimental assessments and discussions

As shown in Fig. 2(a), the experimental set-up used to evaluate the proposed system consists of 4 FPGA-implemented ToRs equipped with 25 Gbps fixed wavelength of WDM TRXs. Fixed wavelength TRXs modulate data to the wavelength required for that packet to switch from the ingress to egress of AWGR in terms of the cyclic wavelength routing. Additionally, Fixed wavelength of WDM TRXs is more suitable for small-scale experimental assessment because it can save the reconfiguration delay caused by TWCs. Anritsu Ethernet Test Center is exploited to generating Ethernet frames with controllable and variable traffic load, emulating 4 servers with 25 Gbps Ethernet links. The ToR is implemented by a FPGA with the Kintex UltraScale+ KCU116 board and support up to 4 25 Gbps channels. The functionality of scheduler is implemented by the experiment scenarios due to the Anritsu Ethernet Test Center to simply the feasibility of the experiments.

First, the bit error rate (BER) performance is investigated to quantify the potential optical signal degradation during transmission in the AWGR switching fabric. As illustrated in Fig. 2(b), back-to-back (B-to-B) as the reference line is the scenarios that 2 direct-connected FPGA through 25Gbps TRX and optical fiber. When AWGR is connected between the two FPGA board, the BER result (CH1) indicates slight performance degradation compared to the result of B-to-B and confirms error-free transmission with power penalty less than 0.5dB at BER of 1E-9.

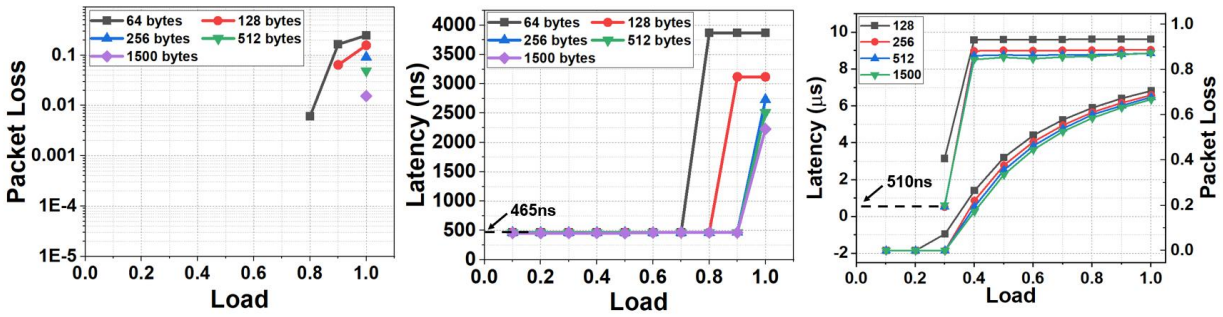

Fig. 3. Network performance of (a) packet loss (b) server-to-server latency.
(c) The scenario of 3 Txs sending traffic to 1 Rx.

Traffic generated from Anritsu Ethernet Test Center flows to ToRs and AWRG. Fig. 2(c) shows the 4-port ToR schedule and forward the incoming data traffic to the expected output. When port 3 Tx receive packet, the internal scheduler extracts the destination MAC address from the Ethernet frame, for example, Dst-Port_3 (means the destination MAC address extracted from port 3 and the result of the corresponding switched port will appear at the forwarder_forward_ports_3) highlights two aspects of information while the former is the destination address and the latter is the port to egress. The forwarders then obtain the destination port by checking the look up table.

Two experimental scenarios are carried out to assess the proposed network. First, the intra-cluster sub-network is demonstrated to validate the network performance in the practical scenario. we assume ToR1 transmits random destination packets to ToR2, ToR3 or ToR4 at variable traffic load. Fig. 3(a) and (b) shows the network performance comparison between 5 different sizes of packet. In real DCN, the size of the packet is mainly concentrated in two places of 200 and 1500 bytes [6]. Therefore, the results of 200- and 1500-bytes packets, error-free with 465ns server-to-server latency except at high traffic load of 1.0, show the system is feasible in DCN. 64-bytes packet is error-free with 465ns under the moderate traffic load (<0.7). Consequently, the results of 64-bytes packets can meet the infrequent chain-building communication such as the instant message service. The buffer read pointer rate is slower than the logic processing rate, as a result, faster process on the small packet (<200 bytes) leads to high packet loss and the large packet needs more cycle to process and can easily match the read rate.

Finally, an extreme case is exploited to test the upper limit of the proposed system. ToR2, ToR3 and ToR4 simultaneously send packets with identical traffic load to ToR1 to emulate the extreme scenario of multi-Txs to one Rx in practical network. The network performance of the extreme scenario is shown in Fig. 3(c). Under low traffic load (<0.4), all tested packets except for 128-bytes packets achieve error-free with 510ns transmission latency. As the traffic load increases (>0.4), for ToR1, its practical load of received traffic has exceeded 1.0, so more than 40% packets dropped. Integrated Logic Analyzer (ILA), a hardware debug tool used to monitor the FPGA-based ToR1, indicates that the buffer is fully occupied. This introduces the competition resolution problem, thus resulting in the high packet loss.

4. Conclusions

We propose and experimentally assess a AWGR based nanoseconds optical switching DCN, exploiting 25G FPGA-implemented ToRs and passive AWGR switches. Experimental assessments indicate that the proposed system can achieve error-free with 465ns latency at the load of 0.9 for various packet size. For the extreme case of multi-Txs sending traffic to one Rx scenario, experimental results show the system can solve contention with error-free traffic forwarding and 510ns latency at the load of 0.3.

References

[1] W. Xia, P. Zhao, Y. Wen, and H. Xie, "A Survey on Data Center Networking (DCN): Infrastructure and Operations," IEEE Communications Surveys & Tutorials, vol. 19, no. 1, pp. 640–656, 1st Quart., 2017.
[2] K.-I. Sato, "Realization and application of large-scale fast optical circuit switch for data center networking" J. Lightw. Technol. vol. 36 no. 7 pp. 1411-1419 Apr. 2018.
[3] Xue, X., Calabretta, N. "Nanosecond optical switching and control system for data center networks" Nature Communications, vol. 13, pp. 2257, 2022.
[4] F. Yan, W. Miao, O. Raz, and N. Calabretta, "Opsquare: A flat DCN architecture based on flow-controlled optical packet switches," Journal of Optical Communications and Networking, vol.9, no. 4, pp. 291-303, Apr. 2017.
[5] F. Yan X. Xue and N. Calabretta, "FOScube: a Scalable Data Center Network Architecture Based on Multiple Parallel Networks and Fast Optical Switches" 2018 European Conference on Optical Communication (ECOC) pp. 1-3.
[6] Benson T, Akella A, Maltz DA, "Network traffic characteristics of data centers in the wild." in Proceedings of the 2010 ACM Special Interest Group on Data Communication (SIGCOMM) Conference, ACM, pp. 267-280, 2010.