Experimental Assessments of a Flexible Optical Data Center Network Based on Integrated Wavelength Selective Switch

X. Xue(1), F. Nakamura(2), K. Prifti(1), B. Pan(1), F. Yan(1), F. Wang(1), X. Guo(1), H. Tsuda(2), N. Calabretta(1)
(1) IPI-ECO Research Institute, Eindhoven University of Technology, Eindhoven, the Netherlands
(2) Graduate School of Science and Technology, Keio University, Yokohama, Japan.
x.xue.1@tue.nl

Abstract: A novel bandwidth-reconfigurable optical DCN exploiting photonic-integrated WSS is experimentally assessed. Results show that optical bandwidth can be automatically reallocated according to the traffic patterns with 1.75μs end-to-end latency and 0.015 packet-loss at 0.6 load.

1. Introduction

The next-generation data centers (DCs) are expected to evolve towards new switching technologies and architectures to upturn the performance in terms of throughput, latency and scalability. Several DC networks (DCNs) switching the traffic in the optical domain have been proposed providing high capacity and low latency interconnections benefitting from the data rate and format transparency [1, 2]. However, the optical bandwidth of these optical DCNs between top of racks (ToRs) is determined by the deployed transceivers at the ToRs and cannot be reallocated on-demand to serve the dynamic DC traffic once the network is deployed. For most practical scenarios, the static bandwidth allocation is not optimal for handling the dynamic traffic patterns generated by the heterogeneous applications. Only a few ToRs are operated at high capacity at a certain time in the practical DCNs while bandwidth and capacity of other ToRs are underutilized. Moreover, the bandwidth requirements for each ToR dynamically varies as the applications run. Therefore, even for optical DCNs with high capacity, the rigid bandwidth appears to be either overkill or insufficient for the running applications. In this work, we propose and experimentally assess an SDN controlled flexible optical DCN architecture based on SOA-based optical switches and a novel reconfigurable optical ToR (OToR) employing a 40-λ 1×2 photonic integrated wavelength selective switch (PIC-WSS). The deployed PIC-WSS provides high flexibility and excellent cost efficiency. By automatically and elastically controlling the deployed TRXs and the PIC-WSS, the total capacity of the fixed TRXs at each OToR can be dynamically reallocated. This enables a flexible optical bandwidth allocation between OToRs links to be adapted to the dynamic traffic matrix generated by the heterogeneous applications. Experimental results show that the PIC-WSS introduces <0.5 dB penalty at bit error rate (BER) of 1E-9 and the SOA based optical switch fully compensates the WSS loss avoiding costly and power consuming EDFAs. The network performance assessments confirm that 1.75 μs end-to-end latency and 0.015 packet loss at load of 0.6. Numerical investigation of the DCN scalability based on experimental parameters indicates the proposed optical DCN can be scaled out to 40960 servers with 0.043 packet loss and 6.4 μs latency at the load of 0.5.

2. Reconfigurable Optical DCN Based on PIC-WSS

The proposed reconfigurable optical DCN architecture is shown in Fig. 1(a). The novel FPGA implemented OToR has been developed for this architecture with respect to our previous proposed OPSquare DCN [3]. Each OToR interconnects H-server in every rack, and N racks are grouped into one cluster. The N N×N intra-cluster optical switches (IS) and N N×N inter-cluster optical switches (ES) are dedicated for intra-cluster and inter-cluster communication, respectively. The *i*-th OToR in each cluster is interconnected by the *i*-th ES (1≤*i*≤N). The traffic generated by the servers can be classified into three categories (intra-OToR, intra-cluster, inter-cluster) and will be

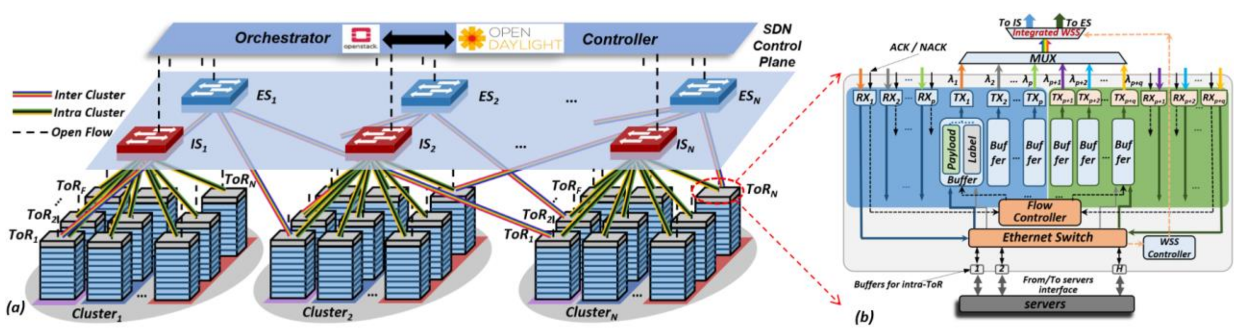

Fig. 1: (a) Reconfigurable optical DCN architecture. IS: intra-cluster optical switches; ES: inter-cluster optical switches. (b) Schematic of the novel OToR exploiting PIC-WSS. TX: Transmitter; RX: Receiver; MUX: multiplexer.

processed by the Ethernet switch at each OToR as shown in Fig. 1(b). For the intra-OToR traffic, the frames will be directly forwarded to the destination servers in the same rack. While for the intra-cluster (IC) and inter-cluster (EC) communication, the frames will be forwarded to the electrical buffer associated with the *p* IC transmitter (TXs) or *q* EC TXs with different wavelength to be aggregated into the optical data packets. The *p+q* TXs define the total output bandwidth of the OToR to serve the IC and EC traffic. It is worth to notice that the traffic pattern and volume between the IC and EC links is variable as the dynamic applications are deployed. By elastically assigning the ratio of the total TXs associated with the IC and EC connections, the optical bandwidth of the IC and EC communications can be therefore adapted to the variable traffic on the IC and EC links. This is implemented by controlling the PIC-WSS to select which of the *p+q* TRXs will be switched at the two outputs of the WSS towards the ES and IS. The OpenDaylight (ODL) and OpenStack platforms are deployed as the SDN controller connecting the IS/ESes and OToRs by means of the extended Open Flow (OF) protocol [4]. The FPGAs-based OToRs can collect data traffic statistics (the traffic volume ratio of IC and EC link), and send such information through OF links to the SDN controller. On the other side, the SDN controller can update the set of actions of OToRs in real-time to automatically reallocate the amounts of *p* and *q*. E.g., if more IC traffic is generated by the newly deployed application, some of the *q* EC TRXs are reassigned by the flexible OToR to serve the IC traffic. The output wavelength (and thus the bandwidth per link) of the integrated WSS will be reassigned accordingly to ES and to IS as well. The PIC-WSS shown in Fig. 2 employs loopback configuration with only one arrayed waveguide grating (AWG) for avoiding waveguide crossings and center wavelength mismatch, decreasing transmission loss and cross talk [5]. The number of wavelength channels is 40 with 100 GHz channel spacing and less than 8.1 dB transmission loss.

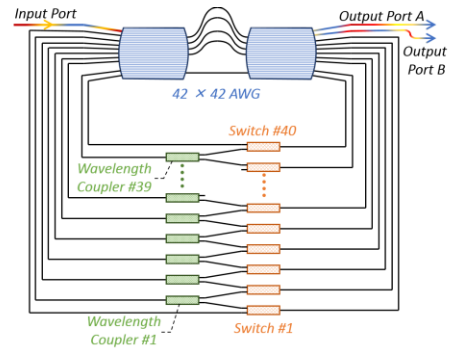

Fig. 2. Structure of the PIC-WSS.

3. Experimental Demonstration and Results

The experimental set-up to validate the reconfigurable optical DCN is shown in Fig. 3. It consists of 3 FPGA-based OToRs and each one is equipped with 4 ($p+q$ = 4) 10 Gb/s WDM TRXs (1541.30 nm, 1542.10 nm, 1542.90 nm, 1543.70 nm). Two 4×4 SOA based optical switches are utilized to forward the IC and EC traffic, respectively. The ODL and OpenStack based SDN control plane connects the switch controllers and OToRs via the OF agents implementing the OF protocol. The SPIRENT Ethernet Testing Center emulating 8 servers at 10 Gb/s generates Ethernet frames with variable and controllable load. Ethernet frames are generated between 64 and 1518 bytes with an average size of 792 bytes. A realistic DCN traffic volume with 50% intra-OToR, 37.5% IC and 12.5% EC traffics are employed in this assessment [6]. First, BER measurements are performed to quantify possible signal degradation caused by the PIC-WSS. Fig. 4(a) shows the BER curves and eye diagrams for the 4 WDM TXs connecting the $OToR_1$ and $OToR_3$ through the optical switch. Error free operation has been obtained with <0.5 dB penalty at BER

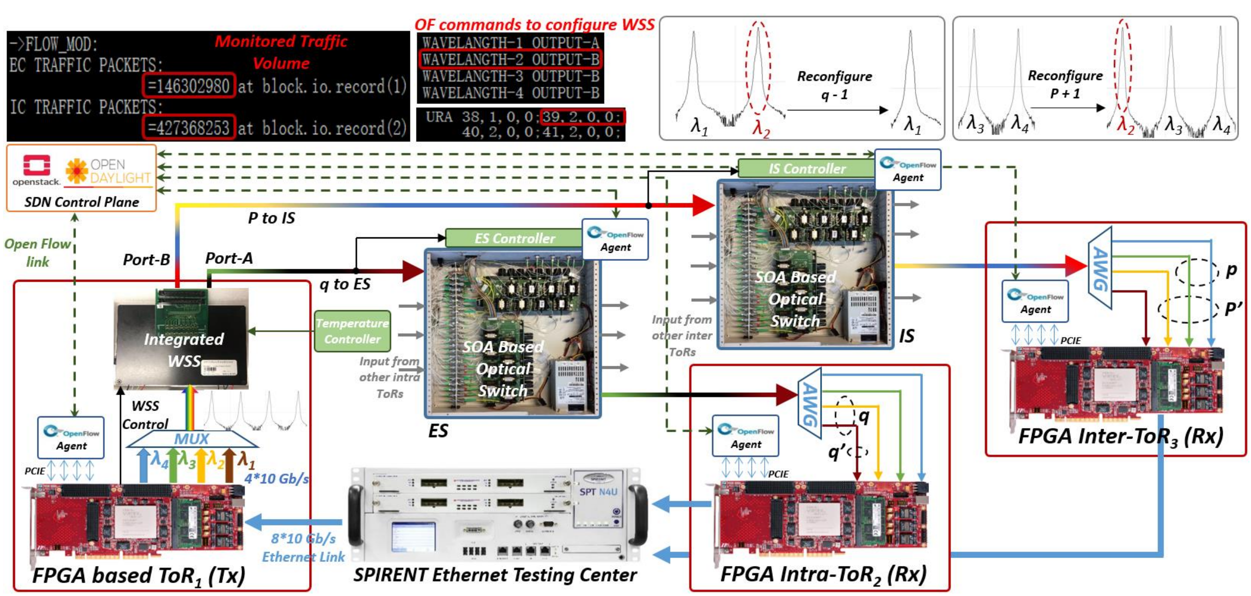

Fig. 3. Experimental set-up of the reconfigurable optical DCN based on the PIC-WSS.

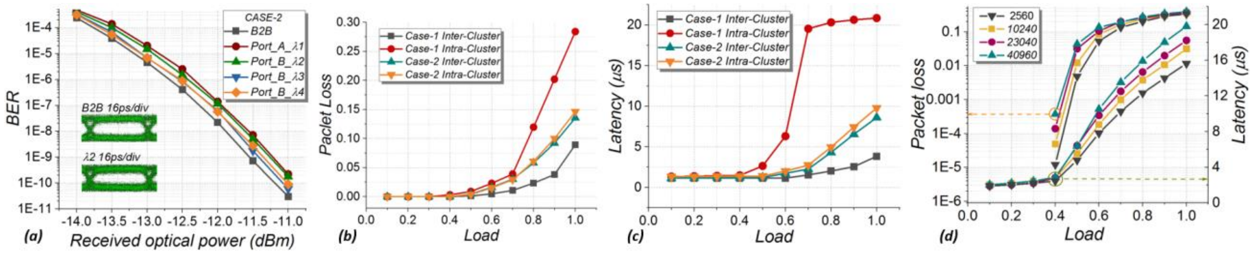

Fig. 4. Experimental results of the reconfigurable optical DCN based on the OToR with PIC WSS.

of 1E-9 after the optical links confirming that the PIC-WSS does not cause any deterioration. The 8.1dB WSS losses are compensated by the SOA switch gates and no EDFA are required in the link. For the demonstration of the dynamic optical bandwidth allocation, in the initial configuration (Case-1), the TRXs with $\lambda_{1,2}$ (q = 2) and $\lambda_{3,4}$ (p = 2) are allocated to forward the traffic of EC and IC, respectively. The Ethernet switch inside the $OToR_1$ monitors the traffic ratio (37.5% IC and 12.5% EC) and reports this information to the SDN controller via the OF link. The SDN controller runs the Bandwidth Computing Engine using the monitored traffic volume and sends the OF commands to $OToR_1$ to provide more bandwidth for the IC communication (see Fig. 3). Therefore the $OToR_1$ reconfigures the WSS so that λ_2 is now used to increase the IC bandwidth. For this new configuration (Case-2), the wavelength $\lambda_{2,3,4}$ (p' = 3) are connected with $OToR_2$ providing more bandwidth to IC traffic, while the λ_1 (q' = 1) connects with the $OToR_3$ for the EC communication.

The network performance (packet loss and latency) before and after the wavelength reconfiguration to validate the improvements of the flexible optical bandwidth is shown in Fig. 4(b) and (c). For the IC link of Case-1, the packet loss increases dramatically after 0.4 load and 0.14 packet loss is measured at load of 0.8. Comparatively, the packet loss of the Case-2 after the automatic optical bandwidth reallocation to serve the traffic volume is 0.05 at the high load of 0.8 for both the IC and EC links. After reconfiguring λ_2 to the IC link, the average network performance is significantly improved with respect to the Case-1. This is because the initial IC bandwidth of Case-1 (p = 2) is insufficient to support the high (37.5%) IC traffic volume, while the EC bandwidth (q = 2) is in excess for the 12.5% EC traffic. The adaptable bandwidth for the deployed application traffic decreases the packets buffer queuing time. This explains the improved latency for IC link of Case-2 (1.23 µs) with respect to Case-1 (19.55 µs) at load of 0.5. Finally, the network performance as a function of the DCN scalability with the novel OToR to implement the flexible optical bandwidth allocation has been numerically investigated. An OMNeT++ simulation model of the DCN is implemented with the experimentally measured parameters like Ethernet switch processing time and buffer queuing time. The typical same traffic pattern (50% intra-OToR, 37.5% IC and 12.5% EC traffic) and the adaptable TRX configuration Case-2 (p = 3 and q = 1) are used in the simulation. The packet loss ratio and server end-to-end latency as a function of number of servers are shown in Fig. 4(d). The packet loss is less than 1E-6 and the end-to-end latency is < 3 µs at load of 0.3 for the large scale (40960 servers) network, which indicates the good scalability of the proposed reconfigurable optical DCN.

4. Conclusions

We propose and experimentally assess a flexible optical DCN based on a novel optical ToR with PIC-WSS. Experimental optical link performance shows <0.5dB penalty at BER of 1E-9 and full compensation of the WSS insertion loss by the SOA gates. Enabled by the ODL and OpenStack based SDN control plane, the automatic optical bandwidth reallocation has been demonstrated. Experimental assessments confirm 0.015 packet loss and 1.75 µs end-to-end latency can be achieved at the load of 0.6 for the adaptable bandwidth reallocation. OMNeT++ simulation based on experimental parameters shows that the proposed DCN can scale to 40960 servers with 1E-6 packet loss and <3 µs end-to-end latency at the load of 0.3.

5. Acknowledgements

The authors want to thank the EU H2020 Passion (G.A. 780326) project for partially supporting this work.

6. References

[1] K. Chen, "WaveCube: A scalable, fault-tolerant, high-performance optical data center architecture", INFOCOM, 2015.
[2] P. Bakopoulos, "NEPHELE: An end-to-end scalable and dynamically reconfigurable optical architecture for application-aware SDN cloud data centers",IEEE Communications Magazine, 2018.
[3] X. Xue, "Flexibility assessment of the reconfigurable OPSquare for virtualized data center networks under realistic traffics", ECOC, 2018.
[4] X. Xue, "Experimental Assessment of SDN-enabled Reconfigurable OPSquare Data Center Networks with QoS Guarantees", OFC, 2019.
[5] Y. Ikuma, "Integrated 40 λ 1×2 Wavelength Selective Switch Without Waveguide Crossings", IEEE PTL, 2013.
[6] S. Kandula, "The nature of data center traffic: measurements & analysis", ACM SIGCOMM conference on Internet measurement, 2009.