SDN enabled flexible optical data center network with dynamic bandwidth allocation based on photonic integrated wavelength selective switch

Citation for published version (APA):

DOI:
10.1364/OE.388759

Document status and date:
Published: 16/03/2020

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain.
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.
Abstract: Optical switching techniques featuring the fast and large capacity have the potential to enable low latency and high throughput optical data center networks (DCN) to afford the rapid increasing of traffic-boosted applications. Flexibility of the DCN is of key importance to provide adaptive and dynamic bandwidth to handle the variable traffic patterns generated by the heterogeneous applications while optimizing the network resources. Aiming at providing the flexible bandwidth for optical DCNs, we propose and experimentally investigate a software-defined networking (SDN) enabled reconfigurable optical DCN architecture based on novel optical top of rack (OToR) switch exploiting photonic-integrated wavelength selective switch. Experimental results show that the optical bandwidth per link can be automatically reallocated under the management of the deployed SDN control plane according to the variable traffic patterns. With respect to the network with inflexible interconnections, the average packet loss of the reconfigurable DCN decreases 1 order of magnitude and the server-to-server latency performance improves of 42.2%. Scalability investigation illustrates limited (11.7%) performance degradation as the reconfigurable network scale from 2560 to 40960 servers. Both the numerical and experimental assessments validate the proposed DCN with reconfigurable bandwidth feature and lower latency variations with respect to the inflexible DCNs.

1. Introduction

As the hubs of the content-centric Internet, data centers (DCs) power the traffic-boosting applications, such as the cloud computing, Internet of Things (IoT) and big data, by hosting hundreds of thousands of servers [1,2]. The proliferation of these applications increases DC traffic on a steep growth reaching 25 percent annually [3,4]. However, the current electrical switches based intra-DC interconnections are facing technical challenges because the implementation of high-bandwidth electrical switches is limited by the ASIC I/O bandwidth as the result of the scaling issue of ball grid array (BGA) package [5,6]. Stacking the ASIC boards in a multi-tier structure could increase the switching bandwidth but at the expenses of extra latency and costly complex interconnections, leading to high cost and power consumption switching architectures. Therefore, to accommodate the tremendous increase in amounts of traffic, the next-generation DCNs are expected to evolve towards new switching technologies with high bandwidth and architectures to upturn the network performance. Benefitting from the data rate and format transparency, switching the data traffic in the optical domain with high bandwidth is gaining momentum as the potential solution for significantly scaling up DCNs [7,8,9]. The very large bandwidth offered by the optical switches also allow for flattening the network architecture then avoiding large latency caused by hierarchical electrical switching structures [10]. Moreover,
the large reduction of power-consuming O/E/O conversions at the optical switches significantly improves the energy and cost efficiency [11].

A multitude of optical DCNs scenarios have been proposed leveraging various optical switching technologies, such as HiFOST, OPSquare and WaveCube architectures based on semiconductor optical amplifiers (SOAs) [12,13] or micro-electro-mechanical systems (MEMS) [14], arrayed waveguide grating routers (AWGRs) built LIONS [15], or a combination thereof architecture present in [16] deploying wavelength-selective switches (WSSs). However, in all these above-mentioned optical DCNs, the optical bandwidth between top of racks (ToRs) is determined by the pre-provisioned transceivers (TRXs) at the ToRs, which cannot be reallocated on-demand to serve the dynamic DC traffic once the network is built. For most practical scenarios, the rigid bandwidth allocation is not optimal for supporting 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 [17]. Moreover, the bandwidth requirements for each ToR dynamically vary as the applications run. Therefore, even for optical DCNs with high capacity, the rigid bandwidth appears to be either overprovisioned or insufficient for the running applications.

One solution is to use intelligent workload-place algorithms to allocate network-bound application components to physical infrastructure with suitable bandwidth connectivity [18]. Nevertheless, the flexible deploying mechanism of workload placement on the whole-infrastructure dramatically expands the complexity of the network control and management in particular for large scale optical networks. On the other side, another approach is to flexibly reconfigure network interconnections providing dynamic network bandwidth to handle application components with variable traffic communications. If the network interconnection could “shape-shift” in such fashion, this could considerably simplify the complicated workload placement problem. Several reconfigurable optical DCNs architectures such as ProjecToR, FireFly and OSA supporting flexible bandwidth and capacity allocation have been proposed [19,20,21]. However, the ProjecToR and FireFly based on the wireless connections have to guarantee the line-of-sight between the TRX pairs. This limits the scalability of these networks to be deployed in multiple places. Additionally, it is hard to fast and accurately align the wireless lines with the target receivers, causing high packet loss. The scalability of wired OSA architecture is limited by the switch port radix, just supporting 2560 servers. POPI and POTORI DCN architectures can provide dynamic bandwidth to serve the various traffic [22,23]. However, the hundreds of microseconds server-to-server latency do not satisfy the latency-sensitive applications.

In this work, we propose and experimentally assess a software-defined network (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 configuring flexibility and excellent cost efficiency. Based on the collected network traffic statistics of data plane, the deployed TRXs per link and the PIC-WSS at OToR can be automatically and elastically configured by the SDN control plane to provide the dynamic optical bandwidth in real-time. This enables an on-demand optical bandwidth allocation mechanism between OToRs links to be adapted to the dynamic traffic matrix generated by the heterogeneous applications. Experimental results show that the PIC-WSS introduces less than 0.5 dB penalty at bit error rate (BER) of 1E-9 and the SOA based optical switch can fully compensate the WSS loss avoiding costly and power-consuming EDFAs. Two kinds of network traffic scenarios are utilized in the simulation model and experimental network to investigate network performance. The network performance in terms of packet loss and latency have been assessed for the reconfigurable network with respect to the inflexible interconnecting network. The latency CDF of the reconfigurable DCN has also been measured to validate the convergence of the latency performance. Moreover, based on
experimental parameters, the numerical investigation on the OMNeT++ simulation model is carried out to investigate the scalability of the reconfigurable DCN.

2. Reconfigurable optical DCN based on PIC-WSS

The proposed reconfigurable optical DCN architecture is shown in Fig. 1. The novel FPGA implemented OToR has been developed for this architecture with respect to our previous proposed OPSquare DCN [13,24,25]. 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). Single-hop link is interconnected by the IS for OToRs locating in the same cluster, and at most two-hop communication is sufficient to forward the traffic between OToRs residing in different clusters. Benefiting from this cluster-divided architecture, multi-path is supported for the connections of each pair OToRs, improving the network fault-tolerance. At each OToR, p and q transceivers with corresponding electrical buffers are deployed to connect the IS and ES for the intra-cluster and inter-cluster communications, respectively. The allocation of p and q can be dynamically assigned on-demand according to desired optical bandwidth and intra-cluster/inter-cluster traffic ratio.

![Fig. 1. Proposed reconfigurable optical DCN architecture. IS: intra-cluster optical switches; ES: inter-cluster optical switches.](image)

The traffic generated by the servers can be classified into three categories (intra-OToR, intra-cluster, inter-cluster) and will be processed by the Ethernet switch at each OToR as shown in Fig. 2(a). 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 buffer associated with the receiver is deployed inside the Ethernet switch. 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 variable applications are dynamically 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. To automatically configure the PIC-WSS and the allocations of p + q TRXs, the OpenDaylight (ODL) and OpenStack platforms are deployed as the SDN controller connecting the IS/ES controllers and OToRs by means of the extended Open Flow (OF) protocol and SDN-agents [26]. Cooperating with OF protocol, these SDN-agents enable the communications between the southbound interface (SBI) of ODL controller and the PCIe interfaces of FPGA-based OToR as shown in Fig. 2(a), whereby bridging monitoring/reporting and configuration mechanisms at
both sides. These agents translate the OF commands generated from the ODL controller into a set of FPGA implemented actions through the proprietary interfaces and vice versa. In particular, 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 number of different wavelengths (and thus the aggregated bandwidth per link) at the integrated WSS outputs will be reassigned accordingly to ES and to IS as well.

![Fig. 2.](image)

Fig. 2. (a) Schematic of the novel OToR. TX: Transmitter, RX: Receiver, PIC WSS: photonic integrated wavelength selective switch, MUX: multiplexer; (b) Structure of the PIC-WSS.

The configuration of the \( 1 \times 2 \) PIC-WSS with 100 GHz spaced 40 channels is shown in Fig. 2(b). The WSS consists of one arrayed waveguide grating (AWG), 40 single stage \( 1 \times 2 \) Mach-Zehnder Interferometer (MZI) switches, and 39 wavelength couplers on silica platform [27,28]. The AWG is utilized as wavelength demultiplexing (for the WDM signal at the input port) as well as multiplexing (for the signal looped back after the MZI switches and wavelength couplers). The WSS adopts loop-back configuration for preventing center wavelength mismatch and decreasing transmission loss and crosstalk caused by waveguide crossing. The input port WDM channels are demultiplexed by the AWG. Each separated channel is switched by the \( 1 \times 2 \) MZI switch using thermo-optics effect to one of the two output ports in order to forward the signal to output port A or B of the WSS. The adjacent wavelength signals are coupled by corresponding wavelength coupler and looped back to the AWG, which is now operated as multiplexing. Therefore, the looped back signals will be forwarded to the output port A or output port B based on the selection made by the \( 1 \times 2 \) MZI switch. The peak loss is less than 8.1 dB and the average crosstalk from the two output ports are \(-23.0\) dB and \(-40.7\) dB, respectively.

The optical packets will be delivered to the destination OToRs via the IE/ESs on the WDM optical links. Meanwhile, the copy of the optical packets are stored in the electrical buffer at each OToRs. An optical label indicating the destination information of the corresponding optical packet is attached to the packet header. The fast SOA-based optical switch (IS/ES) illustrated in Fig. 3 features a modular structure and the WDM traffic coming from the same OToR is processed at each module [29]. Once the optical packet arriving at the ES/IS, the label signals will be separated from the optical packets and be processed on-the-fly at the label extractor, while the data payloads are sent to the select and broadcast \( 1 \times N \) switch. With the extracted label bits, the switch controller resolves the packet contention and accordingly controls the SOA switch.
gates inside the 1×N switch for the packet payloads forwarding. To solve the packet contention caused packet loss, Optical Flow Control (OFC) signals (ACK/NACK) generated at the ES/IS nodes will be sent back to the corresponding OToRs. If the OToR receives a positive ACK (no contention and packet successful forwarded), the stored packet will be released from the buffer. In response to a negative acknowledgement (NACK) which means the packet has been dropped due to the contention, the stored packet will be retransmitted and go through the same procedures again [30,31]. The SOA with nanoseconds switching speed works as fast optical gate and also as amplifier to compensate for the broadcasting architecture caused splitting losses and the WSS loss. The combiners at the output of the optical switch aggregate the identical wavelength to the same destining OToR.

![Fig. 3. Schematic of the SOA based optical switch. COM: Combiner.](image)

3. Experimental demonstration and results

The experimental set-up to validate the reconfigurable optical DCN is shown in Fig. 4. 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 mean value of the injected current into the SOA gates is 120 mA to compensate the 8.1 dB losses of the PIC-WSS. The SOA based switch features nanoseconds reconfiguration time benefitting from the fast (3 nanoseconds) response time. The ODL and OpenStack based SDN control plane connects the switch controllers and the OToRs via the OF agents implementing the OF protocol. The SPIRENT Ethernet Testing Center emulating 24 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. The realistic DCN traffic volume (Traffic-A: 50% intra-OToR, 37.5% IC and 12.5% EC traffics; Traffic-B: 50% intra-OToR, 12.5% IC and 37.5% EC traffics) are employed in this assessment [32,33]. For the demonstration of the dynamic optical bandwidth allocation for the application with Traffic-A model, in the initial configurations of the TRXs and the PIC-WSS (Case-1), the TRXs with \(\lambda_{1, 2} (q = 2)\) and \(\lambda_{3, 4} (p = 2)\) are allocated by the SDN control plane to forward the traffic of EC and IC, respectively. The Ethernet switch inside the OToR monitors the traffic ratio (37.5% IC and 12.5% EC) and reports this information to the SDN control plane via the OF link. The monitored statistics is illustrated in Fig. 4. The SDN controller runs the Bandwidth Computing Engine using the monitored traffic volume and sends the OF commands within 125 milliseconds to OToR to provide more bandwidth for the IC communication (see Fig. 4). Therefore the OToR automatically reconfigures the PIC-WSS so that \(\lambda_2\) 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 $\lambda_1$ ($q' = 1$) connects with the OToR$_3$ for the EC communication. These optical bandwidth allocations are automatically operated under the management of the SDN control plane without manual operations.

First, BER measurements are performed to quantify possible signal degradation caused by the PIC-WSS. Figure 5 shows the BER curves and eye diagrams for the optical links between the OToR$_1$, OToR$_2$ and OToR$_3$ for these two kinds of configuring cases (Case-1 and Case-2). Error free operation has been obtained with less than 0.5 dB penalty at BER of 1E-9 before/after the optical link configurations. This confirms that the ASE noise of SOA gate and PIC-WSS cause a very limited deterioration. It is worth to notice that the inherent 8.1 dB losses of PIC-WSS can be compensated by the SOA switch gates and thereby no EDFA is required in this proposed reconfigurable DCN.

The network performance (packet loss and server-to-server latency) for the Traffic-A model before and after the wavelength reconfiguration to validate the improvements of the flexible optical bandwidth is shown in Figs. 6(a) and 6(b). For the inflexible IC link of Case-1, where the
optical bandwidth is not adaptable for the traffic volume, the packet loss increases dramatically after 0.4 load and 0.12 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 0.06 at the high load of 0.8 for both the IC and EC links. After reconfiguring $\lambda_2$ 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 after the wavelength reallocations for the deployed application traffic model A decreases the packets buffer queuing time at the OToR. This also explains the 68.94% improvements of latency performance for IC link of Case-2 (1.91 $\mu$s) with respect to Case-1 (6.15 $\mu$s) at the load of 0.6.

Fig. 6. Packet loss (a) and server-to-server latency (b) for Traffic-A before and after the bandwidth reallocations.

To emulate the real network operation environment where deploys various applications with traffic pattern switchover, the Traffic-B (50% intra-OToR, 12.5% IC and 37.5% EC traffics) is generated by the SPIRENT Ethernet Testing Center to emulate the new application deployed and previous application (generating Traffic-A) switchover. Following the same reconfiguration procedures from Case-1 to Case-2, the SDN control plane will monitor the change of the traffic ration and send OF commands to automatically reassign the optical bandwidth (from Case-2 to Case-3) to adapt the new traffic pattern B. For this new configuration (Case-3) of Traffic-B, the wavelength $\lambda_1$ ($p'' = 1$) is connected with OToR2 for the IC communication, while the $\lambda_2, 3, 4$ ($q'' = 3$) connects with the OToR3 providing more bandwidth to EC traffic.

Figures 7(a) and 7(b) show the packet loss and server-to-server latency for the Traffic-B model before (Case-2) and after (Case-3) the wavelength reconfiguration. With respect to inflexible Case-2 where the bandwidth is unmatched with the new traffic model B, the packet loss and latency performance of the reconfigured Case-3 with adaptable bandwidth provisioning improve one magnitude order and 80%, respectively, at the high load of 0.6. The network with interconnection of Case-3 achieves an average packet loss of 0.06 for the IC and EC links, and an average server-to-server latency of 5.0 $\mu$s. We also count the server-to-server latency for all the optical packets of Case-2 and Case-3 when the load of Traffic B is 0.6. The Cumulative Distribution Function (CDF) of server-to-server latency is shown in Fig. 8. The results indicate that the server-to-server latency of IC/EC links for the reconfigured Case-3 (with adaptable bandwidth allocation) have low variations with respect to the mean value of 1.85 $\mu$s and 2.43 $\mu$s, respectively. As a comparison, the latency distribution of IC link for the inflexible Case-2 with insufficient optical bandwidth provisioning is more dispersed, with latency distribution from 3.5 $\mu$s to 50 $\mu$s.
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 platform of the DCN is implemented with the experimentally measured parameters. The servers are programmed to generate Ethernet frames with the length varying from 64 bytes to 1522 Bytes at the load from 0 to 1. Frame arrival time model is built based on the ON/OFF periods length (with/without data packets forwarding). The data packets on each ON period are randomly destined to one of the possible servers under the dedicated traffic pattern. The preamble length of optical packets is set as 1E3 Bytes and the delay of the physical realization of label processing and switch controlling is port-count independent and has been measured as 40 ns in total. The average distance between the OToR and ES/IS are 50 meters, therefore, the round trip time (RTT) between the OToR and ES/IS is 560 ns, which includes transmission over the 2×50 m distance and 60 ns delay caused by the label processor as well as flow control operations. 20 \((p + q = 20)\) WDM transceivers are deployed at each OToR and each transceiver equipping with 50 KB buffer operates at 10 Gb/s. 40 servers are grouped in the same rack and the server operating rate is 10 Gb/s. The Traffic-A, Traffic-B and the adaptable TRX configuration Case-2 and Case-3 are used in the simulation model, respectively. The average packet loss ratio and server end-to-end latency as a function of number of servers are shown in Fig. 9(a). Firstly, the performance of the simulated network with the same server numbers (24) as
the experimental setup has been investigated to validate the OMNeT++ simulation model. The packet loss and latency performance illustrate that the 24 servers simulation (24-Sim) platform are matching with the 24 servers experimental (24-Exp) network. The numerical results validate only average 11.7% performance degradation as the reconfigurable network scale from 2560 to 40960 servers. The packet loss is less than 1E-6 and the end-to-end latency is below 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. The CDF of the network latency for 40960 servers and traffic pattern B has also been assessed for the inflexible (Case-2 and load 0.6) and reconfigured (Case-3 and load 0.6) optical interconnections. The results shown in Fig. 9(b) prove that 90% packets for the DCN with reconfigurable and adaptable optical bandwidth converges at the mean value. The latency of EC link with optical bandwidth provisioning of Case-2 distributes from 5 µs to 85 µs.

Fig. 9. (a) Packet loss and server-to-server latency for different network scale; (b) Cumulative Distribution Function (CDF) of latency of network deploying 40960 servers for two cases connections of Traffic B at the load of 0.6.

4. Conclusions

We propose and experimentally assess a reconfigurable optical DCN with adaptable optical bandwidth provisioning based on a novel OToR with PIC-WSS deployment. Experimental performance assessment of OToR to OToR link shows below 0.5 dB penalty at BER of 1E-9 and full compensation of the PIC-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. At the load of 0.6, experimental assessments confirm 0.015 packet loss and 1.80 µs end-to-end latency can be achieved for the adaptable bandwidth reallocation, which is one order of magnitude and 42.2% improvements, respectively, with respect to the network with inflexible interconnections. We built an OMNeT++ simulation model based on experimental parameters. Numerical results prove that the proposed optical reconfigurable DCN features only limited (11.7%) performance degradation as the network scale from 2560 to 40,960 servers, by automatically providing adaptable bandwidth to two kinds of network traffic pattern. Network performance for 40,960 servers achieves 1E-6 packet loss and less than 3 µs end-to-end latency at the load of 0.3. The numerical assessments of latency CDF for large scale (40,960 servers) network also validate the much lower latency variations (90% packets converged) for bandwidth adaptable network with respect to the inflexible interconnections.

Funding

Horizon 2020 Framework Programme (PASSION. 780326).
Disclosures
The authors declare no conflicts of interest

References