Automatically reconfigurable optical data center network with dynamic bandwidth allocation

Journal of Optics

PAPER • OPEN ACCESS

Automatically reconfigurable optical data center network with dynamic bandwidth allocation

OPEN ACCESS
IOP Publishing

Journal of Optics

J. Opt. **23** (2021) 114003 (10pp)

https://doi.org/10.1088/2040-8986/ac29cb

Automatically reconfigurable optical data center network 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], Yu Wang[1], Chongjin Xie[3] and Nicola Calabretta[1]

[1] IPI-ECO Research Institute, Eindhoven University of Technology, Eindhoven, The Netherlands
[2] Alibaba Cloud, Alibaba Group, Hangzhou, People's Republic of China
[3] Alibaba Cloud, Alibaba Group, Sunnyvale, CA, United States of America

E-mail: x.xue.1@tue.nl

Received 23 June 2021, revised 16 August 2021
Accepted for publication 24 September 2021
Published 22 October 2021

Abstract

The rapid increasing traffic in data centers (DCs) puts tremendous pressure to the present multi-tier network architectures and electrical switching techniques. Switching traffic in the optical domain featuring ultra-high bandwidth, therefore, has been intensively investigated to build the high capacity data center networks (DCNs). To handle the variable traffic pattern in DCs, the network reconfigurability with adaptable optical bandwidth allocation is of key importance to flexibly assign the optical bandwidth. To this end, we propose and experimentally evaluate a software-defined networking enabled reconfigurable optical DCN with dynamic bandwidth allocation in this work, based on novel optical top of racks exploiting a wavelength selective switch. Experimental assessments show that the proposed solution can automatically reallocate the optical bandwidth in real-time to adapt the dynamic traffic pattern. Compared with the conventional optical DCN with static bandwidth provision, the end-to-end latency performance of the reconfigurable scheme with adaptable bandwidth allocation improves of 58.3% and the average packet loss decreases one order of magnitude. Moreover, the reconfigurable optical DCN features deterministic latency performance, with much lower time variations of packets delivery completion. Based on the experimental parameters, the simulation platform is also built to validate the good scalability of the proposed reconfigurable DCN. Numerical results illustrate the negligible performance degradation (11%) as the network scales from 2560 to 40 960 servers.

Keywords: optical interconnects, data center network, optical switches, reconfigurable architectures, flow control, software defined networking

(Some figures may appear in color only in the online journal)

[*] Author to whom any correspondence should be addressed.

1. Introduction

Recently, with the escalation of traffic-boosting services and applications, such as Internet of Things, cloud computing and high definition streaming, traffic bandwidth growth in data centers (DCs) exceeds that of wide-area telecom networks and even outpaces the bandwidth growth rate of electrical switch application-specific integrated circuits [1]. Current electrical switches are expected to hit the bandwidth bottleneck due to the inability to increase the pin-density on the ball grid array packaging technique [2]. To overcome this bandwidth bottleneck issue of electrical switches, optically switching the traffic has been considerably investigated as a future-proof solution supplying ultra-high bandwidth [3]. Benefiting from the optical transparency, the optical switch with high bandwidth is independent of the bit-rate and data-format of the traffic [4]. Moreover, migrating of the switching functionality from electrical to optical domain removes the power- and time-consuming optical–electrical–optical (O–E–O) conversions and eliminates the dedicated electronics circuits for various-format modulation, hence, significantly decreasing cost and processing delay [5]. Additionally, the unlimited bandwidth offered by the optical switching techniques also seamlessly supports the employment of wavelength-division multiplexing (WDM) technology, enabling the flexible and high-efficiency bandwidth utilization [6].

Leveraging various optical switch techniques, a multitude scenarios of optical data center networks (DCNs) have been proposed and numerically investigated, such as micro-electro-mechanical systems (MEMSs) based WaveCube [7], semiconductor optical amplifiers (SOAs) built HiFOST and OPSquare [8, 9] and LIONS based on arrayed waveguide grating (AWG) routers [10], or a combination thereof architecture deploys wavelength-selective switches (WSSs), presented in [11]. In all these aforementioned optically switched schemes, once the network is built, 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 the optical bandwidth cannot be reallocated on-demand to adapt the variable DC traffic volume. Nevertheless, the fixed bandwidth provision is not suitable for most DCN scenarios, not only due to the low-efficiency network resource utilization but also because the unadaptable bandwidth cannot guarantee network performance. Only a few links between the ToRs, as reported in [12], require high bandwidth in a certain time, while most bandwidth of other links between ToRs is underutilized in DCNs. Additionally, the bandwidth requirements between ToRs are also dynamically varying as the hosted applications and services switchover. Thus, the static bandwidth allocation appears to be ether 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/service components to network infrastructure with adaptable bandwidth interconnections [13, 14]. However, the implementation of these mechanisms needs to consider the overall network infrastructure to flexibly allocate workload placement, significantly increasing the complexity of network management and control, particularly for large-scale optical DCNs. By dynamically reconfiguring the network interconnections, another potential solution is to provide the elastic bandwidth to serve the application/services components generating variable traffic volumes. 'Reconfiguring' the network interconnections in this approach could considerably simplify the complexity with respect to the complicated workload-placing issue. Several flexible DCNs providing dynamic bandwidth allocation, such as FireFly, ProjecToR and OSA, have been proposed and investigated [15–17]. Nevertheless, the FireFly and ProjecToR architectures based on wireless interconnections requires the paired TRXs to maintain the line-of-sight. This requires stable placing environment, limiting the large-scale deployment of such networks in multiple places. Moreover, the wireless lines are hard to be accurately and fast aligned between the paired TRXs, which, therefore, can cause a large number of packet loss. The OSA network, with limited 2560 servers connected, is not suitable for the building of central DCNs, where the network needs to be scaled out to 10 000 servers.

A flexible optical DCN architecture based on fast optical switches has been proposed and numerically investigated in [18]. In this work, a software-defined network (SDN) enabled reconfigurable DCN with dynamic optical bandwidth allocation has been proposed and experimentally assessed, based on photonic integrated SOA-based optical switches and a novel reconfigurable ToR. The SDN control plane can monitor the network traffic of the connected data plane in real time. Based on the collected statistics, the deployed TRXs per optical link and the associated wavelength selective switch (WSS) at each ToR as well as the photonics integrated optical switch can be dynamically configured to provide the adaptable optical bandwidth. With this mechanism, the optical bandwidth between any ToRs can be accordingly reconfigured to adapt and serve the variable traffic matrix. Experimental and numerical assessments have been used to evaluate the reconfigurability, network performance and network scalability. Two sorts of traffic volumes are generated in experimental setup and simulation model to investigate the network performance. Experimental results prove that at the bit error rate (BER) of 1×10^{-9}, the photonics integrated switch chip introduces less than 1.0 dB penalty for all the channels. At 0.6 traffic load, the network performance of 0.013 packet loss and 3.25 μs server-to-server latency are achieved for the network with adaptable bandwidth reallocation. The simulation model built based on the experimental parameter also proves the fine scalability of the reconfigurable DCN.

2. SDN-enabled reconfigurable optical DCN

The novel reconfigurable optical DCN is demonstrated in figure 1. The SOA-based fast optical switches including the field-programmable gate array (FPGA)-implemented switch controller as well as novel FPGA-implemented ToRs with dedicated optical interfaces have been developed to fully

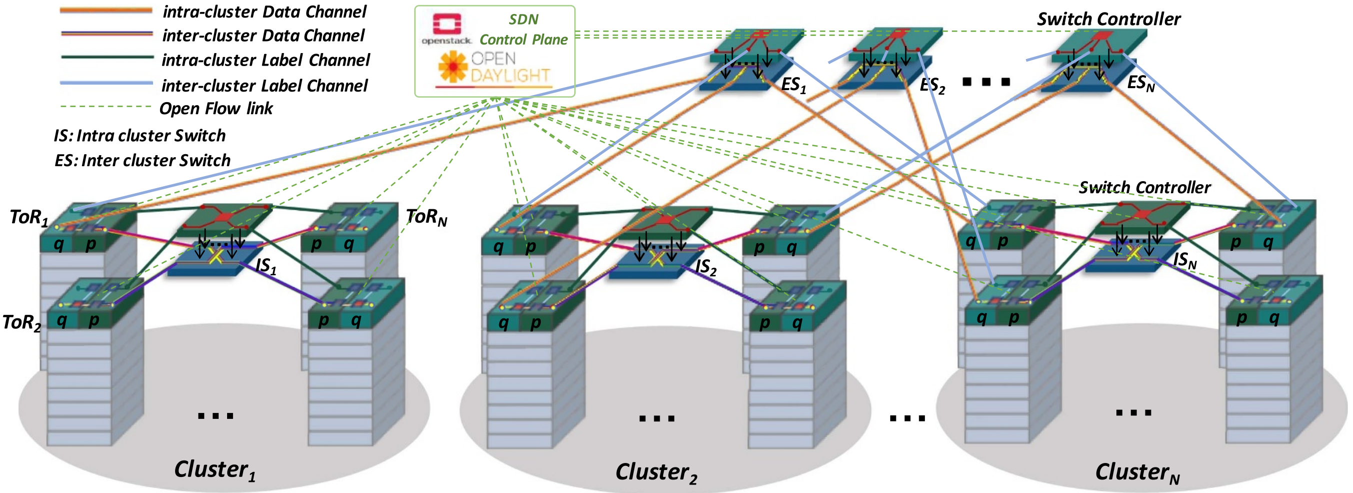


Figure 1. Reconfigurable optical DCN architecture. IS: intra-cluster optical switches; ES: inter-cluster optical switches.

support the reconfigurable operations, different from our previous works [9, 19, 20]. In each rack, the ToR interconnects H-server, and N racks form one cluster. To fulfill the inter-cluster and intra-cluster traffic, the N inter-cluster optical switches (ES) and N intra-cluster optical switches (IS) both with a radix of N are employed. More specifically, the ith ES is responsible for the communications among the ith ToR in each cluster ($1 \leqslant i \leqslant N$). The data packets are carried on the data channels to implement the intra-cluster and inter-cluster communication. Meanwhile, the destination of the data packets is delivered on the label channels to the switch controllers to accordingly control the forward of the corresponding data packets. For intra-cluster communications where ToRs are located in same cluster, single-hop passing IS is sufficient. At most two-hops of optical switches are needed for inter-cluster communications with ToRs residing in different clusters. Moreover, benefiting from multi-path of the inter-cluster communications, the architecture supports load balancing, improving the network fault-tolerance.

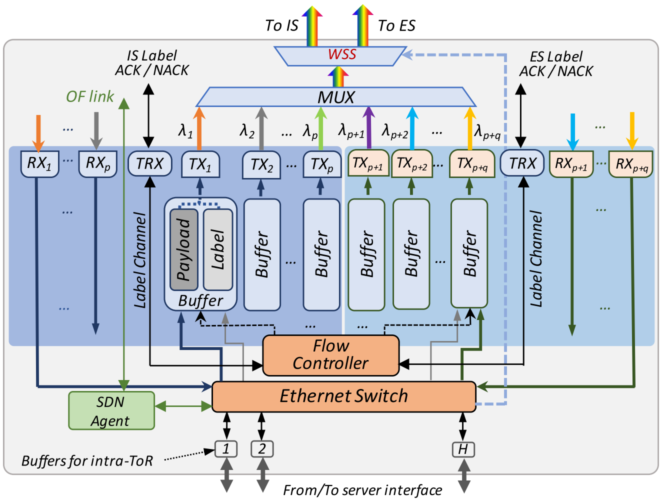


Figure 2. Schematic of the novel optical ToR exploiting WSS. TX: transmitter; RX: receiver; MUX: multiplexer; WSS: wavelength selective switch.

2.1. Novel optical ToR with dedicated optical interfaces

The FPGA implemented ToR with dedicated optical interfaces is shown in figure 2, The ToR is equipped with $p + q$ TRXs, each with private electrical buffers. We use p TRXs to connect the ToR with the IS for intra-cluster communication, while the other q TRXs are utilized for inter-cluster communications by interconnecting the ToR to the ES. The number of p and q can be flexibility adjusted based on the desired bandwidth requirement of intra-cluster and inter-cluster traffic amounts. We classify the traffic from servers into three categories (intra-ToR, intra-cluster and inter-cluster). The three kinds of traffic are firstly processed by the Ethernet switch once Ethernet frames generated at servers arrive at ToR. The frames of the intra-ToR traffic are directly forwarded to the servers locating in the same rack. As to the inter-cluster (EC) and intra-cluster (IC) traffic, the ToR forwards the frames to the electrical buffer of corresponding transmitter (TX) based on the destinations. The Ethernet frames with the same destination are aggregated as the optical data packet. The total uplink bandwidth of the ToR is determined by the $p + q$ TXs with different wavelength. Note that the traffic ratio of the IC and EC changes dynamically as the running of heterogeneous applications. By elastically allocating the number of TXs on the IC and EC connections at each ToR, the IC and EC bandwidth can be therefore adjusted to adapt the variable traffic volumes on the IC and EC links. Meanwhile, by controlling the WSS, certain amounts of TXs are dynamically switched to one of the two outputs of the WSS, connecting with the IS and ES, respectively.

2.2. SDN control plane

An OpenStack based network service orchestrator is deployed as the virtual infrastructure manager and the workload composer. Moreover, to automatically allocate the $p + q$ TRXs and to configure the optical switches, the OpenDaylight (ODL) are deployed as the SDN controller connecting ToRs and optical

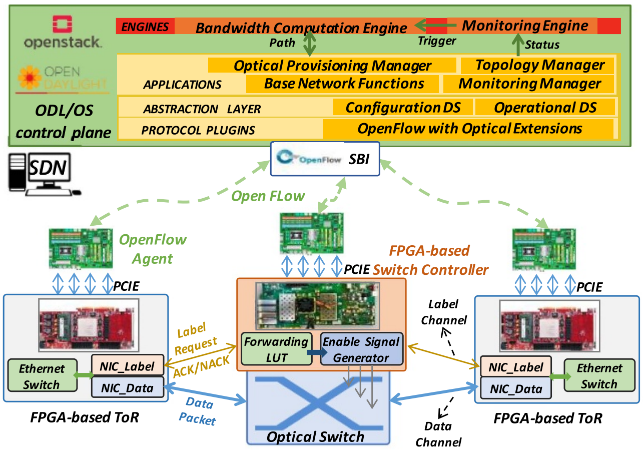


Figure 3. Detailed interconnections of the reconfigurable optical DCN. NIC: network interface controller.

switch (ES and IS) controllers by means of SDN-agents and the extended OpenFlow (OF) protocol [21]. Cooperating with OF protocol, these agents enable the connections between the ODL southbound interface and the peripheral component interconnect express (PCIE) interfaces of FPGA-based ToR and switch controller as shown in figure 3, bridging monitoring/reporting and configuring functionalities at both sides. The SDN-agents gather the monitored information from the FPGA-based ToRs and report it to the SDN controller for further processing. In details, the monitored data is first sent through the PCIE interface of the FPGA to the SDN-agent driven by the direct memory access technique. Then the dataflow from PCIE is translated to the OF packets and further forwarded to the SDN controller. The SDN controller collects and processes all the monitored data and accordingly reconfigure the network. As shown in the SDN control plane of figure 3, the physical distribution information and data plane layout are stored in the topology manager (TM). The bandwidth computation engine (BCE) of the OpenStack-based orchestrator provides the ODL controller with a ToR-to-ToR capacity computation service for bandwidth allocation, based on network topological information from the TM. The optical provisioning manager module sends the flow configuration messages (FlowMod) to the OF Agents according to the calculated ToR-to-ToR bandwidth assignment. The FlowMod messages then configure the underlying devices (i.e. ES, IS, ToR and WSS) to set the specific WSS and TXs configurations. Furthermore, the statistics of optical data plane (the traffic volume of EC and IC links) are collected by the monitoring manager and aggregated into ToR-to-ToR level metrics in real time. Such aggregated information is collected by the monitoring engine at OpenStack to trigger the bandwidth reallocation to adapt the traffic volume and therefore, guaranteeing the expected low packet loss and latency. For instance, if a higher amount of EC traffic is generated as the running of the DC applications, part of the p IC TRXs will be automatically reallocated at the novel optical ToR to serve the newly added EC traffic. The amounts of wavelengths (and thus the aggregated optical bandwidth per link) at the WSS outputs, connecting the IS and ES, will be reassigned as well to provide the dynamic optical bandwidth to serve the variable EC/IC traffic.

2.3. Optical flow control (OFC) technique

The key difference of optical switches with respect to electrical switches is the lack of buffers. The electrical switches employ random access memories (RAMs) to buffer the conflicted data packets. Once packet contention occurs at electrical switch node, the packets lost the contention in current time slot would still get chance (retransmit) in the following time slots. Since no practical optical buffers exist, the conflicted optical packets at the optical switch would be dropped, thereby resulting in high packet loss [22]. Thus, to prevent the packet contention caused packet loss, an OFC technique is developed in the proposed reconfigurable network. In the operation of OFC, the Ethernet frames generated by the servers are first processed at the Ethernet switch at ToR. The frames of IC and EC traffic with the same destination are aggregated as the optical data packet. The original of optical data packets are stored in the associated electrical buffer of $p + q$ TXs, as illustrated in figure 2. By the p/q WDM optical data channels, the copy of the original packets is delivered to the destination via the IE/ESs. As shown in figure 3, a label request signal indicating the destination of the associated optical data packet is generated at the ToR. The label signals are delivered to the switch controller through the label channels to manage the forwarding of the corresponding data packets. After processing the label requests, the switch controller sends an acknowledgement (ACK) signal (indicating packet is successfully forwarded by the optical switch) or unacknowledgement (NACK) signal (indicating packet is forwarded to the un-destined ToRs) back to corresponding ToRs. Based on the received ACK or NACK signal, the ToR releases the stored original packet from the electrical RAM (ACK signal) or trigger the data packet processor to retransmit the optical copy of original packet (NACK signal).

2.4. SOA-based optical switch

Due to the short time constant, the SOA has less ideal transmission properties in terms of lower saturation output power, noise figure and pattern dependence compared with the fiber amplifiers [23]. The SOA, however, has advantages of small size, low cost and high potential for photonic integration. Moreover, the coverage of low-loss bands allows for the transmission at 1300 nm, which is dispersion-free in standard single-mode fiber. Additionally, the SOA gates in the optical switch can compensate for the broadcasting caused splitting losses, by amplifying the optical power. Thus, in this reconfigurable optical DCN, the SOA has been used as the selecting gate element in the $N \times N$ optical switches, to provide the fast (<1 ns) switching on/off operation and compensation to the broadcasting loss [24]. The schematic of the SOA-based $N \times N$ switch node used to interconnect the ToRs is shown in figure 4 with a modular structure. Each switch connects N ToRs, and the N switch modules handle the traffic from each ToR independently. $p + q$ WDM channels, with wavelength

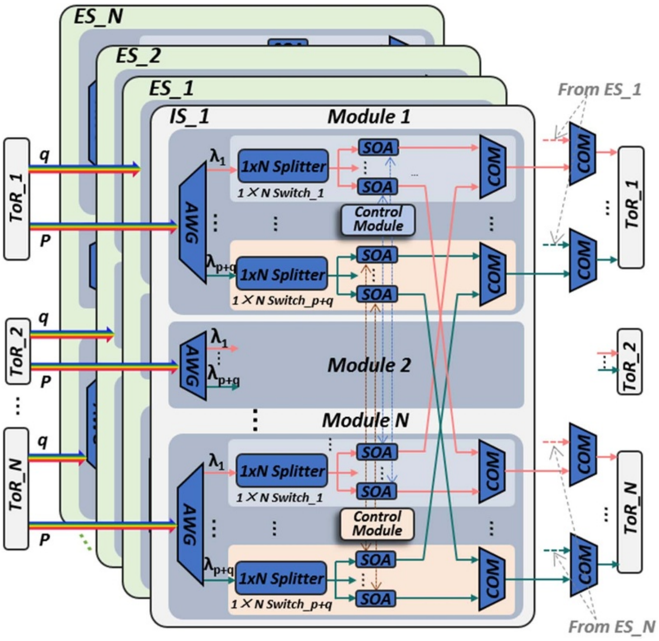


Figure 4. Schematic of the SOA based optical switch. COM: combiner.

centered at $\lambda1$, $\lambda2$, ..., $\lambda p + q$, carry the data packets aggregated at each ToR to different destinations crossing the SOA switch. At the switch module, AWG is first placed at the module input to disaggregate the $p + q$ WDM traffic. Afterwards, the packets are fed into the $1 \times N$ optical switch node. As the implementation of OFC technique, the label request signals indicating the destination information of the associated data packets are delivered to the switch controller on the label channel. Extracting the destination information and the stored look-up table, the switch controller then controls the $1 \times N$ optical switch to forward the optical packets to the right output port. Finally, the combiners at the output ports aggregate the identical wavelength from different sources to the same destining ToR.

3. Experimental evaluation

Based on the broadcast-and-select structure of SOA-based fast optical switch as shown in figure 4, a 6×4 mm^2 fabricated photonic switch chip [25] integrating four optical modules, schematically illustrated in figure 5, is utilized in this experimental setup. The photonic switch chip with four independent modules can be used to implement the 4 IS or 4 ES for 4 ToRs inter-cluster or intra-cluster interconnections. An 800 μm booster SOA at the input stage of each module is placed to compensate the 1:4 splitter caused 6 dB losses and to compensate the AWGs caused partially losses. 1×2 multimode interferometers are cascaded to realize the passive 1:4 splitter. The WDM inputs can be processed and forwarded to the target output by the mentioned optical switching modules. In addition, the switching module supports multicast and broadcast due to its broadcast-and-select structure.

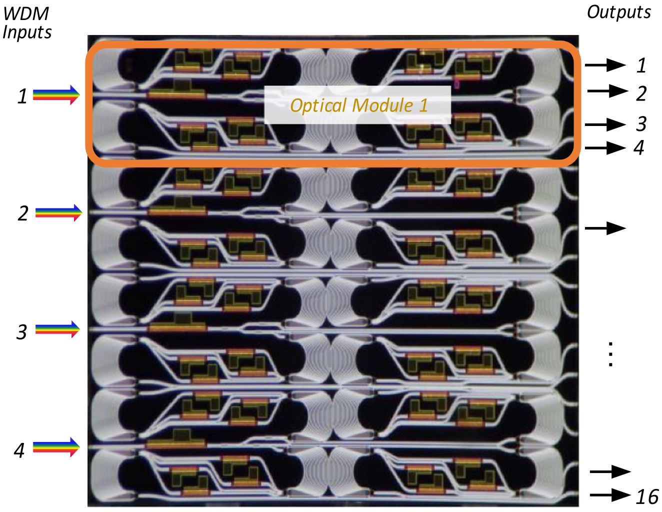


Figure 5. Photonic integrated WDM switch chip.

The switch works in a parallel way and each module works independently, making the switching time and complexity of entire switch equal to the single module and independent of the port-count.

An experimental setup for validating the proposed reconfigurable optical network is shown in figure 6. The setup consists of 6 FPGA-based ToRs, in which 3 ToRs (ToR1, ToR2 and ToR3) are equipped with 4 ($p + q = 4$) 10 Gb s^{-1} WDM TRXs (1543.70, 1542.90, 1542.10, 1541.30 nm). 4 ToRs (ToR1, ToR4, ToR5 and ToR6) connecting to the same optical switch (EC) are utilized to generate the packet contention and then to validate the designed OFC technique. 2 4×4 SOA based photonic switch chips are used to work as the EC and IC switches. To compensate for the splitter losses, the mean value of 120 mA current is injected into the SOA gates. The chip is working in 25 centigrade temperature that is stabled by a watercooler. The ODL based SDN control and OpenStack based service orchestration plane 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^{-1}. Ethernet frames with an average size of 792 bytes are generated randomly between 64 and 1518 bytes.

3.1. BER performance

First, the BER performance of each WDM channel at 10 Gb s^{-1} are measured to quantify the possible signal degradation caused by the photonic integrated switch chip. The fourth module (ES in experimental setup), as one of the four identical modules, has been selected for the BER performance assessment. Figure 7 shows the BER curves versus the received power. The back-to-back curve is also included for reference. Error-free operations with less than 0.5 dB penalty have been measured at BER of 1×10^{-9} for Channel 1 (CH1) and Channel 3 (CH3) at 10 Gb s^{-1} data rates for the case of

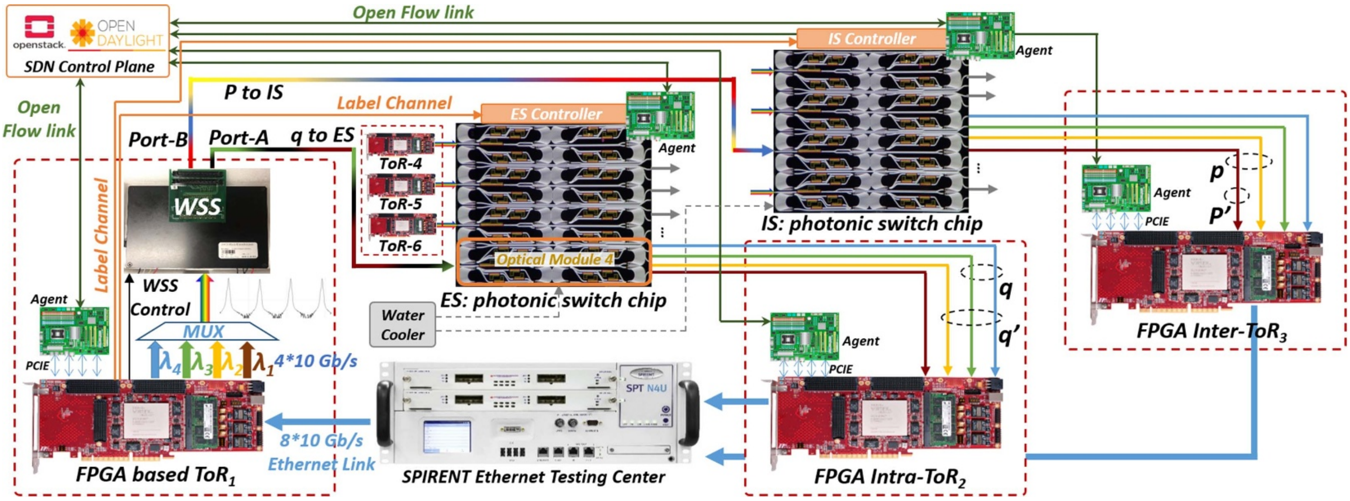


Figure 6. Experimental set-up of the SDN enabled reconfigurable optical DCN.

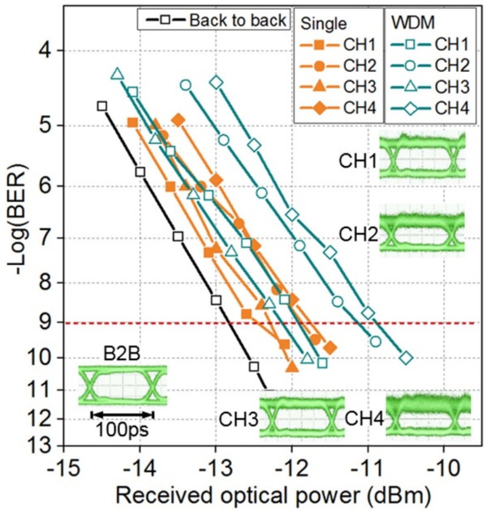


Figure 7. BER curves for single channel and WDM channel input.

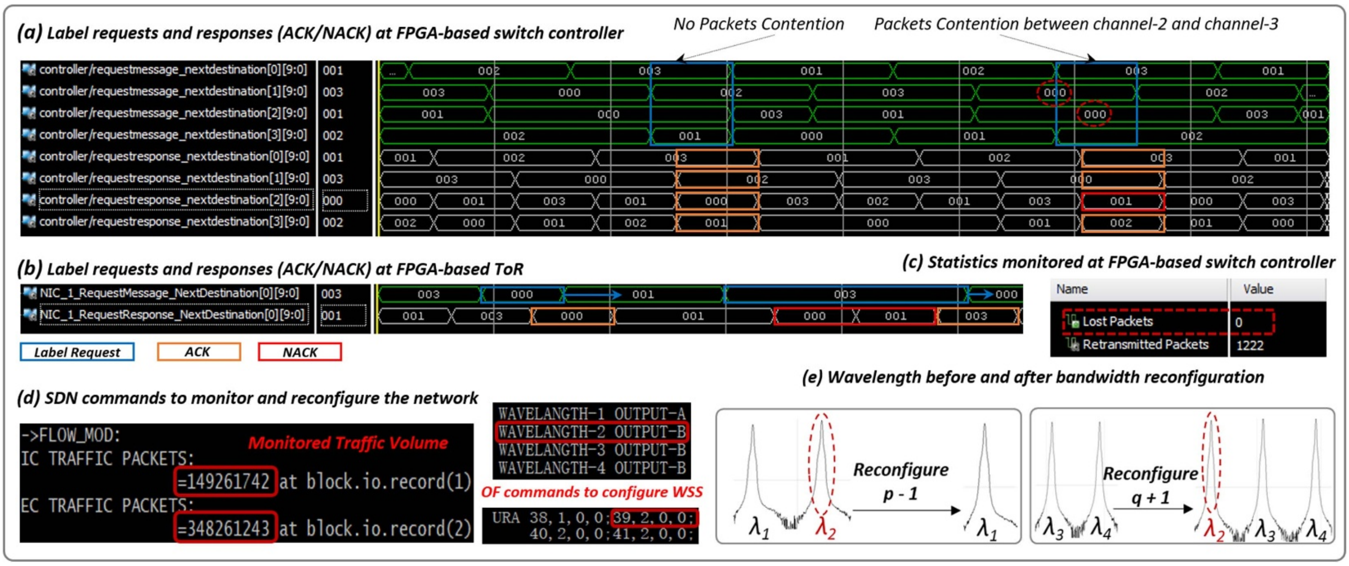
single-channel input. While for Channel 2 (CH2) and Channel 4 (CH4), the penalty is around 1 dB. The signal degradation mainly due to accumulated noise for CH2 and CH4, which is shown and confirmed by the captured eye diagrams. When all the four WDM input channels are fed into the switch, the BER results indicate slight performance degradation with an extra penalty of around 0.5 dB for CH1 and CH3 and 1 dB for CH2 and CH4 compared with single wavelength operations.

3.2. Evaluation of OFC technique

The OFC technique developed to prevent the packet contention caused packets loss at the switch node has also been demonstrated and evaluated. Heavy traffic load is generated at the SPIRENT Ethernet Testing Center to introduce more packets contention. ToR1, ToR4, ToR5 and ToR6 are connected to the switch of ES. Optical label requests (RequestMessage_NextDestination signals (RM)) of ToR1, ToR4, ToR5 and ToR6 as shown in figure 8(a), carrying the destination (destined ports) of the associated data packets are sent to the switch controller at every time slot. Based on the received label request signals, the switch controller executes the contention resolution and generate the switch enable signals accordingly. Afterwards, the label responses (RequestResponse_NextDestination (RR)) are sent to the corresponding ToRs from the switch controller, as a result of the contention resolution. For instance, as illustrated in figure 8(a), when there is no contention at one time slot (e.g. RM: 003, 002, 000, 001), the RR (003, 002, 000, 001) signals, same with the RM signals, will be generated and sent back to the corresponding ToRs. While in case of contention at one time slot (e.g. RM: 003, 000, 000, 002), where the traffic form two ToRs has the same destination (RM: 000, 000), the RR (003, 000, 001, 002) signals, different with the RM signals, are sent back to the corresponding ToRs. If the ToR receives the ACK signal (RR = RM) meaning the packet forwarded successfully, as shown in figure 8(b), the traffic stored in the electrical buffer will be released and a new RM signal will be sent out in the next time slot. Otherwise, if the ToR receives the NACK signal (RR /= RM) meaning the packet dropped at the switch, the stored packet and associated RM signal will be retransmitted in the next time slot. The monitored statistics (counts of retransmitted and lost packets) at the FPGA-based switch controller are shown in figure 8(c), which confirms that the OFC technique implements zero packets loss on the optical links.

3.3. Dynamic bandwidth allocation

According to the traffic distribution reported in [26, 27], two kinds of DC traffic are generated in this experiment by adjusting the MAC address of Ethernet frames in SPIRENT. Traffic-A consists of average 50% intra-ToR, 15% IC and 35% EC traffic, while Traffic-B has average 50% intra-ToR, 35% IC and 15% EC traffic. At start, the model of Traffic-A is


(a) Label requests and responses (ACK/NACK) at FPGA-based switch controller
No Packets Contention
Packets Contention between channel-2 and channel-3
(b) Label requests and responses (ACK/NACK) at FPGA-based ToR
Label Request
ACK
NACK
(c) Statistics monitored at FPGA-based switch controller
(d) SDN commands to monitor and reconfigure the network
Monitored Traffic Volume
OF commands to configure WSS
(e) Wavelength before and after bandwidth reconfiguration
Reconfigure p - 1
Reconfigure q + 1
λ1
λ2
λ3
λ4

```
controller/requestmessage_nextdestination[0][9:0]    001
controller/requestmessage_nextdestination[1][9:0]    003
controller/requestmessage_nextdestination[2][9:0]    001
controller/requestmessage_nextdestination[3][9:0]    002
controller/requestresponse_nextdestination[0][9:0]   001
controller/requestresponse_nextdestination[1][9:0]   003
controller/requestresponse_nextdestination[2][9:0]   000
controller/requestresponse_nextdestination[3][9:0]   002
```

```
NIC_1_RequestMessage_NextDestination[0][9:0]    003
NIC_1_RequestResponse_NextDestination[0][9:0]   001
```

Name	Value
Lost Packets	0
Retransmitted Packets	1222

```
->FLOW_MOD:
IC TRAFFIC PACKETS:
                   =149261742 at block.io.record(1)
EC TRAFFIC PACKETS:
                   =348261243 at block.io.record(2)
```

```
WAVELANGTH-1 OUTPUT-A
WAVELANGTH-2 OUTPUT-B
WAVELANGTH-3 OUTPUT-B
WAVELANGTH-4 OUTPUT-B
```

```
URA 38, 1, 0, 0;39, 2, 0, 0;
    40, 2, 0, 0;41, 2, 0, 0;
```

Figure 8. Label request and response signals at (a) switch controller and (b) ToR; (c) statistics monitored at switch controller; (d) SDN commands to monitor and reconfigure the network; (e) wavelength before and after bandwidth reconfiguration.

generated and deployed in this setup. In the dynamic bandwidth allocation to traffic-A case, the initial configurations (Case-1) allocates the $\lambda1$, 2 ($p = 2$) for the IC traffic, and the $\lambda3$, 4 ($q = 2$) for the EC traffic, respectively. The ToRs monitor the traffic distribution and report this statistic to the SDN control plane in real time on the OF links. The traffic statistic (counts of EC and IC packets) received by the SDN controller is illustrated in figure 8(d). After processing the statistic in the BCE, the SDN controller sends out the OF commands to ToR1 within 125 ms to allocate more bandwidth resource (by reassigning the TRXs and reconfiguring the WSS, see figure 8(e)) for the EC traffic. Benefiting from the automatic reallocation mechanism, the underutilized wavelength $\lambda2$ is now applied to serve the EC traffic. In the new bandwidth configuration (Case-2) after reallocation, the wavelength $\lambda2$, 3, 4 ($q' = 3$) are connected with ToR2 providing more bandwidth to EC traffic, while the $\lambda1$ ($p' = 1$) is connected with the ToR3 for the IC traffic. Without the manual intervention, the SDN controller enables automatic reallocation of optical bandwidth based on the traffic statistic from the data plane.

Figures 9(a) and (b) depict the network performance comparison between the fixed case (Case-1) and reconfigurable case (Case-2) based on the Traffic-A model. The packet loss and server-to-server latency are set as network performance criteria to validate the dynamic optical bandwidth allocation. 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 with the Traffic-A. Utilizing the automatic 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, the network performance of the Case-2 outperforms the Case-1 after reallocating the wavelength $\lambda2$ to the EC link. The reason is that the initial EC bandwidth ($q = 2$) is overloaded under the Traffic-A model (35% EC traffic), whereas the IC bandwidth ($p = 2$) is underutilized with respect to the 15% EC traffic. Therefore, the wavelength reallocation relieves the EC link load and guarantees the network performance of the IC link.

The packets completion time of the Case-2 decreases as well due to the bandwidth reallocation, and the latency of Case-2 on the EC link (6.98 μs) achieves 67.28% improvements compared with the Case-1 (21.33 μs) at the load of 0.8. To analyze the latency distribution, we count the server-to-server latency of 40 000 optical packets under a load of 0.6 for Case-1 and Case-2, respectively. The cumulative distribution function (CDF) of latency is shown in figure 9(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 and 4.42 μs, respectively, while the latency distribution of EC link in the Case-1 varies from 3.5 to 50 μs.

By customizing the SPIRENT to adjust the MAC address, new Traffic-B (50% intra-ToR, 35% IC and 15% EC traffics) is generated and deployed to emulate the real DC operation environment, where the traffic distribution is varied (e.g. from Traffic-A to Traffic-B) as the hosted applications switchover. Similar with the reallocation procedure from Case-1 to Case-2, the global SDN controller receives the new network statistic and sends OF commands to reallocate the optical bandwidth (from Case-2 to Case-3). The bandwidth configuration of Case-2 suitable for the Traffic-A maintains the original connections, while the reconfigured new Case-3 adjusts the bandwidth according to the new traffic pattern B. To adapt the traffic distribution of the Traffic-B model, the wavelengths $\lambda1$, 2, 3 ($p'' = 3$) are accordingly allocated to IC links, and the wavelength $\lambda4$ ($q'' = 1$) is assigned to connect with the ToR2 for the EC traffic. The network performance comparison is illustrated in figures 9(d) and (e) for the Case-2 and Case-3. The network of Case-3 achieves an average end-to-end latency of 4.5 μs and packet loss of 0.01 for the EC and IC links at a load of 0.6. Compared with the Case-2 configuration that is

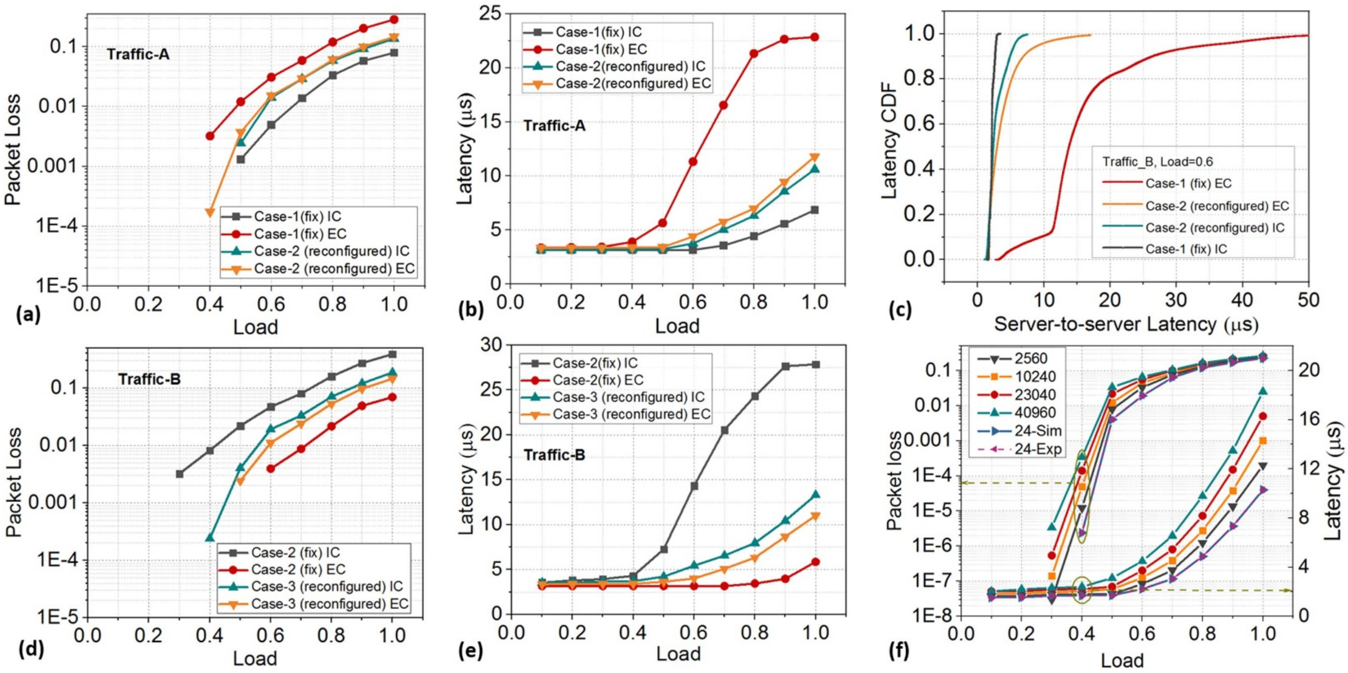


Figure 9. Network performance (packet loss and server-to-server latency) for various traffic pattern and bandwidth configurations.

mismatched for the new traffic model B, the Case-3 improves 65% latency performance, and a packet loss reduction of over one magnitude order.

3.4. Scalability investigation

Finally, we numerically investigate the network performance of the proposed DCN as the network scale-out, equipping the novel flexible ToR to implement the dynamic optical bandwidth allocation. A simulation platform for the proposed DCN based on OMNeT++ is built, adopting the parameters measured in the experiments. The Ethernet frames generated at the servers are programmed with the length varying from 64 to 1518 bytes with controllable and variable traffic load from 0.1 to 1. The model of frame arrival time is designed based on the ON/OFF periods length (with or without data packets forwarding). The Ethernet frames randomly destine to any possible servers in each ON period, constrained by the dedicated traffic pattern. The length of the aggregated optical packets is 2080 ns and the interval packets time is 200 ns. The preamble of the optical data packets is set as 1×10^3 bytes to recover the clock and data. The delay (43.4 ns) of the label processing is port-count independent and is set through experimental measurement. The total TRX latency and Ethernet medium access control/physical layer (MAC/PHY) processing time at the switch controller (Xilinx Virtex-7 VC709) is 354.3 ns. Thus, the total processing delay at the switch controller is 397.7 ns. The fiber length between the ToR and switch nodes are 50 m. The total TRX latency and Ethernet MAC/PHY processing time at the FPGA-based ToR (Xilinx Ultrascale XCVU095) is 167 ns. Thus, the round-trip time on the label channel is 1064.7 ns (354.3 ns + 43.4 ns + 167 ns + 250 ns + 250 ns). 4 ($p + q = 4$) 50 Gb s^{-1} TRXs are deployed at each ToR and each TRX is equipped with 50 KB electrical buffer. Each rack groups 40 servers with 10 Gb s^{-1} link. The Traffic-A and its adaptable TRX configuration Case-2, Traffic-B and TRX configuration Case-3 as discussed in the experimental section are simulated in the OMNeT++ model, respectively.

The average packet loss ratio and server-to-server latency as a function of network size are shown in figure 9(f). Firstly, the performance of the simulated network with the same network scale (24 server) as the experimental network has been investigated to validate the built simulation model. The network performance in terms of both packet loss and latency illustrates that the simulation results (24-Sim) matches with the experimental results (24-Exp). As the reconfigurable network scales from 2560 to 40 960 servers, the numerical results validate only average 11% latency performance degradation. At a load of 0.3, the packet loss is less than 1×10^{-5} and the server-to-server latency is below 3.5 μs for the large scale (40 960 servers) network, which indicates the good scalability of the proposed reconfigurable optical DCN.

4. Conclusion

We propose and experimentally evaluated an SDN enabled optical DCN with reconfigurable optical bandwidth provisioning, based on photonic integrated fast SOA-based switches and novel optical ToRs deploying WSS. The switch controller and ToRs are implemented by the FPGA and the 4 × 4 photonic switch chips are utilized to interconnect the ToRs. Experimental assessments of ToR-to-ToR link shows below 2 dB penalty at BER of 1×10^{-9} for all the channels of the photonic switch chip. The developed OFC technique solves the packet contention and implements zero packet loss at the switch node. Enabled by the SDN control plane, the dynamic optical bandwidth assignment has been implemented to adapt the variable traffic volume. At the 0.6 traffic load, the reconfigurable network with adaptable bandwidth reallocation achieves the 3.25 μs server-to-server latency and 0.013 packet loss, which is 58.3% and one order of magnitude improvements, respectively, compared with the network with fix interconnections. The latency CDF validates the reconfigurable network featuring deterministic latency performance, with much lower latency variations (90% packets converged). OMNeT++ simulation model is also built based on the experimental parameters to investigate the scalability of the reconfigurable DCN. Network size of 40 960 servers achieves less than 3.5 μs server-to-server latency and 1×10^{-5} packet loss at the load of 0.3. The numerical results prove that the network performance only deteriorates 11% as the network scale from 2560 to 40 960 servers.

Data availability statement

The data generated and/or analyzed during the current study are not publicly available for legal/ethical reasons but are available from the corresponding author on reasonable request.

Funding

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

ORCID iDs

Xuwei Xue https://orcid.org/0000-0003-3059-1764
Kristif Prifti https://orcid.org/0000-0002-8481-5391
Chongjin Xie https://orcid.org/0000-0003-4732-3037

References

[1] Ghiasi A 2015 Large data centers interconnect bottlenecks *Opt. Express* **23** 2085–90

[2] Dorren H, Wittebol E H, de Kluijver R, de Villota G G, Duan P and Raz O 2015 Challenges for optically enabled high-radix switches for data center networks *J. Lightwave Technol.* **33** 1117–25

[3] Testa F and Pavesi L 2017 *Optical Switching in Next Generation Data Centers* (Berlin: Springer)

[4] Calabretta N, Wang W, Ditewig T, Raz O, Agis F G, Zhang S, de Waardt H and Dorren H 2010 Scalable optical packet switches for multiple data formats and data rates packets *IEEE Photonics Technol. Lett.* **22** 483–5

[5] Fiorani M, Aleksic S, Casoni M, Wosinska L and Chen J 2014 Energy-efficient elastic optical interconnect architecture for data centers *IEEE Commun. Lett.* **18** 1531–4

[6] Liu Y, Yuan H, Peters A and Zervas G 2016 Comparison of SDM and WDM on direct and indirect optical data center networks *ECOC 2016; 42nd European Conf. on Optical Communication* (VDE) pp 1–3

[7] Chen K, Wen X, Ma X, Chen Y, Xia Y, Hu C, Dong Q and Liu Y 2017 Toward a scalable, fault-tolerant, high-performance optical data center architecture *IEEE/ACM Trans. Netw.* **25** 2281–94

[8] Yan F, Xue X and Calabretta N 2018 HiFOST: a scalable and low-latency hybrid data center network architecture based on flow-controlled fast optical switches *IEEE/OSA J. Opt. Commun. Netw.* **10** 1–14

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

[10] Yin Y, Proietti R, Ye X, Nitta C J, Akella V and Yoo S 2012 LIONS: an AWGR-based low-latency optical switch for high-performance computing and data centers *IEEE J. Sel. Topics Quantum Electron.* **19** 3600409

[11] Zhu Z, Zhong S, Chen L and Chen K 2015 Fully programmable and scalable optical switching fabric for petabyte data center *Opt. Express* **23** 3563–80

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

[13] Hamilton J 2009 Data center networks are in my way *Stanford Clean Slate CTO Summit*

[14] Zhang J, Huang H and Wang X 2016 Resource provision algorithms in cloud computing: a survey *J. Netw. Comput. Appl.* **64** 23–42

[15] Ghobadi M, Mahajan R, Phanishayee A, Devanur N, Kulkarni J, Ranade G, Blanche P-A, Rastegarfar H, Glick M and Kilper D 2016 Projector: agile reconfigurable data center interconnect *Proc. 2016 ACM SIGCOMM Conf.* pp 216–29

[16] Hamedazimi N, Qazi Z, Gupta H, Sekar V, Das S R, Longtin J P, Shah H and Tanwer A 2014 Firefly: a reconfigurable wireless data center fabric using free-space optics *Proc. 2014 ACM Conf. on SIGCOMM* pp 319–30

[17] Chen K, Singla A, Singh A, Ramachandran K, Xu L, Zhang Y, Wen X and Chen Y 2013 OSA: an optical switching architecture for data center networks with unprecedented flexibility *IEEE/ACM Trans. Netw.* **22** 498–511

[18] Nakamura F, Prifti K, Pan B, Yan F, Wang F, Guo X, Tsuda H and Calabretta N 2020 Experimental assessments of a flexible optical data center network based on integrated wavelength selective switch *Optical Fiber Communication Conf.* (Optical Society of America) p W1F. 5

[19] Wang F, Agraz F, Pages A, Pan B, Yan F, Spadaro S and Calabretta N 2019 Experimental assessment of SDN-enabled reconfigurable OPSquare data center networks with QoS guarantees *2019 Optical Fiber Communications Conf. and Exhibition (OFC)* (IEEE) pp 1–3

[20] Nakamura F, Prifti K, Pan B, Yan F, Wang F, Guo X, Tsuda H and Calabretta N 2020 SDN enabled flexible optical data center network with dynamic bandwidth allocation based on photonic integrated wavelength selective switch *Opt. Express* **28** 8949–58

[21] Wang F, Agraz F, Pagès A, Pan B, Yan F, Guo X, Spadaro S and Calabretta N 2020 SDN-controlled and orchestrated OPSquare DCN enabling automatic network slicing with differentiated QoS provisioning *J. Lightwave Technol.* **38** 1103–12

[22] Yao S, Yoo S B and Dixit S 2003 A unified study of contention-resolution schemes in optical packet-switched networks *J. Lightwave Technol.* **21** 672

[23] Turkiewicz J P 2006 Applications of O-band semiconductor optical amplifiers in fibre-optic telecommunication networks, Eindhoven University of Technology

[24] Stabile R 2017 Towards large-scale fast reprogrammable SOA-based photonic integrated switch circuits *Appl. Sci.* **7** 920

[25] Calabretta N, Miao W, Prifti K and Williams K 2016 System performance assessment of a monolithically integrated WDM cross-connect switch for optical data centre networks *ECOC 2016; 42nd European Conf. on Optical Communication* (VDE) pp 1–3

[26] Benson T, Akella A and Maltz D A 2010 Network traffic characteristics of data centers in the wild *Proc. 10th ACM SIGCOMM Conf. on Internet Measurement* pp 267–80

[27] Velan P, Medková J, Jirsík T and Čeleda P 2016 Network traffic characterisation using flow-based statistics *NOMS 2016–2016 IEEE/IFIP Network Operations and Management Symp.* (IEEE) pp 907–12