Flow-controlled and Clock-distributed Optical Switch and Control System

Flow-Controlled and Clock-Distributed Optical Switch and Control System

Xuwei Xue, Bitao Pan, Xiaotao Guo, and Nicola Calabretta

***Abstract*—Switching the traffic in the optical domain has been considerably investigated as a future-proof solution to overcome the intrinsic bandwidth bottleneck of electrical switches in data center networks (DCNs). However, due to the lack of fast and scalable optical switch control mechanism, the lack of optical buffers for contention resolution, and the complicated implementation of fast clock and data recovery (CDR), the practical deployment of fast optical switches in data centers (DCs) remains a big challenge. In this work, we develop and experimentally demonstrate for the first time a flow-controlled and clock-distributed optical switch and control system, implementing 43.4 ns optical switch configuration time, less than 3.0E-10 packet loss rate resulting from the packet contention, and 3.1 ns fast CDR time. Experimental results confirm that zero buffer overflow caused packet loss and lower than 3 μs server-to-server latency are achieved for network deploying a smaller electrical buffer of 8192 bytes at a traffic load of 0.5. Real servers running the Transmission Control Protocol (TCP) traffic generating and monitoring tools are exploited in this switch and control system as well, validating its capability of running practical DCNs services and applications with full TCP bandwidth.**

***Index Terms*—Switch control, optical data center network, contention resolution, fast clock and data recovery (CDR).**

I. Introduction

WITH the large-scale deployment of high-traffic applications, such as high-definition streaming, cloud computing and 5G services, traffic growth in data centers (DCs) outpaces the bandwidth growth rate of application-specific integrated circuits (ASICs) electrical switch [1]. Because the Ball Grid Array (BGA) packaging technique is difficult to increase the pin-density, current ASICs electrical switches are expected to hit the bandwidth bottleneck in two generations from now [2]. As a future-proof solution supplying unlimited bandwidth, optical switching techniques have been considerably investigated to overcome this bandwidth bottleneck of electrical switches [3]. Being independent of the data-format and bit-rate, the optical switch can provide theoretical unlimited bandwidth benefiting from the optical transparency [4]. Moreover, switching the traffic in the optical domain removes the power-consuming and time-cost optical/electrical/optical (O/E/O) conversions at the switch nodes. Optical switching could also eliminate the dedicated circuits and devices for various format modulation, hence, significantly decreasing the cost expenses as well as processing delay [5].

Due to the inability of information processing at the optical switches, a suitable and fast control mechanism is required for the switch control and fast-forwarding of the data traffic to effectively harness the promises held by the high-bandwidth optical switches [6]. Typically, an optical label or header, carrying the destination information of the data packet, is associated with the data packet to be processed at the switch controller [7]. Based on the received destination, the controller will accordingly configure the optical switch and then to forward the data packets. Short switch configuration time compromising both hardware switching time and controlling overhead is essential as it determines the packet completion time and network throughput. Thus, to not inflate the completion time of short data packets and to fully utilize the nanoseconds-level hardware switching time, the controller should handle the label signals and configure the switch in nanoseconds-magnitude. Besides this, the controlling overhead should be independent of the network scale. Given the scale of present DC networks (DCNs), which typically comprises a few thousand racks and hundreds of thousands servers, the switch control needs to be performed in parallel and independently for every optical switch, not for network-wide scale schedule. In addition, for the label control mechanism, the edge nodes connected with the optical switch must be precisely time-synchronized (ideally few nanoseconds) to align the label signals and corresponding data packets [8]. Even an accordingly sized interpacket gap can compensate for the inaccuracy of time synchronization, however, this would reduce the overall throughput. Therefore, the implementation of fast and scalable switch control is a big roadblock for the deployment of optical switches in DCNs, requiring a nontrivial amount of ingenuity and custom hardware support.

Packet contentions occur at the switch node whenever two or more packets try to leave the switching fabric at the same time on the same output port. Random access memories (RAM) is deployed for electrical switches to buffer the data packets that lost the contention and then to prevent the packet loss. The lack of optical buffer is one of the main architectural differences between optical switches and electrical switches. Due to the nonexistence of optical RAM, the conflicted packets that lost the contention at optical switch nodes would be dropped and thereby resulting in high packet loss [9]. Thus, to build the optically switched DCNs, packet

Manuscript received June 22, 2021; revised December 21, 2021; accepted February 24, 2022. Date of publication March 4, 2022; date of current version May 18, 2022. This work was supported by the EU-H2020 project PASSION project (G.A.780326). The associate editor coordinating the review of this article and approving it for publication was A. Fumagalli. *(Corresponding author: Bitao Pan.)*

The authors are with the Electro-Optical Communications Group, IPI Research Institute, Eindhoven University of Technology, 5612 AZ Eindhoven, The Netherlands (e-mail: b.pan@tue.nl).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCOMM.2022.3156613.

Digital Object Identifier 10.1109/TCOMM.2022.3156613

contention resolution is another unsolved critical challenge that needs to be addressed. Despite several approaches have been proposed to overcome such issue, based either on optical fiber delay lines (FDLs) [10], wavelength conversion [11] or deflection routing [12], none of them is practical for large-scale DCNs, due to the fixed buffering time (FDLs), extra hardware deployment (wavelength conversion) and management complexity (deflection routing).

Unlike the point-to-point synchronized connections between any paired electrical switches, optical switches create only momentary physical links between sources and destination nodes [13]. Therefore, in a packet-based optically switched network, where the clock frequency and phase of the data signal vary packet by packet, new physical connections (optical links) are created every time the switch configuration changes. This implies that the receiver has to continuously adjust the local clock (frequency and phase) to properly sample the incoming packets and then to recover the data. As no valid data can be received before the clock and data recovery (CDR) has completed, the longer this process takes (hundreds of nanoseconds for off-the-shelf transceivers), the lower the network throughput will be, particularly for the intra data center scenarios where many applications produce short traffic packets [14]. Burst-mode CDR receivers with nanoseconds data recovery time based on gated oscillators or over-sampling have been extensively investigated in passive optical network. These techniques, however, increase the cost and complexity of the transceiver design, and need to be redeveloped and re-evaluated for higher link data rates, not suitable for large-scale DCNs [8].

The aforementioned challenges in terms of fast switch control, packet contention resolution and fast CDR locking have been the roadblocking to the deployment of fast optical switches in DCNs. To overcome these issues, the optics and networking communities have been investigating various independent techniques, but most of them focus on one specific issue. Moreover, each community address the challenges and problems from their own perspective. Some techniques are the promising solutions for one certain challenge but they could also introduce other issues. For instance, the technique of clock phase caching is used to implement the fast CDR in [14], however, to correct the clock phase jitter along the Tx-Rx path, the extra time overhead of phase-shifting at each transmitter has offset the advantages resulted from fast CDR. Thus, the solutions need to be synergistically planed and have less resource occupation and implementation complexity. In this work, we propose and experimentally demonstrate a flow-controlled and clock-distributed optical switch system to comprehensively solve all these issues. The nanoseconds optical switch and control system is based on a combination of a new label control and synchronization mechanism to achieve a fast system switch control of 43.4 ns, a novel Optical Flow Control technique to prevent contention and the need of an optical buffer, achieving less than 3E-10 packet loss, and a clock distribution mechanism implementing 3.1 ns CDR time, preventing deployment of complex burst mode CDR receivers. All these developed techniques are synergistically implemented on the optical label channels. This practical and low complexity scheme could significantly decrease network resource occupation. The proposed switch system can be used to build the large-scale optical DCNs. Network performance, such as the packet loss and latency, under the Ethernet frame size, buffer dimensioning and traffic distribution has been experimentally investigated. Experimental assessments validate that the switch system implements lower than 3 μs server end-to-end latency and zero buffer overflow caused packet loss at the traffic load of 0.5. Real servers deployed in this system, running the TCP traffic generating and monitoring tools, validates the capability of running practical services and applications with full TCP throughput.

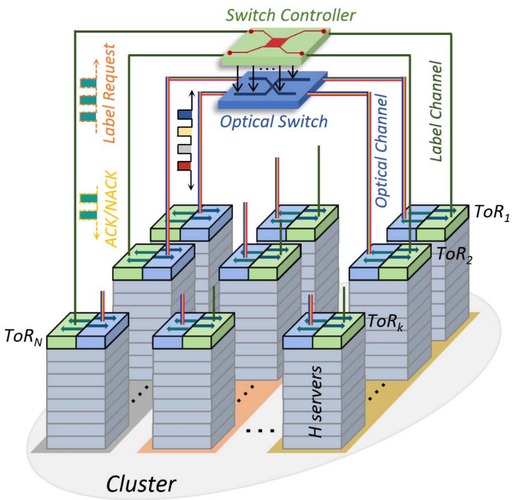

Fig. 1. Flow-controlled and clock-distributed optical switch system deployed at the cluster unit.

II. Principle

The proposed flow-controlled and clock-distributed optical switch and control system is organized in a cluster scale, not developed for the overall network. This enables the distributed system fully scalable, being independent of the network scale, even for DCNs comprising a very large number of racks and servers. Thus, the optical switch system can be applied to various optical DCN architectures organized in parallel and independent clusters granularity, such as LIONS, HiFOST and OPSquare topologies based on fast (nanoseconds) optical switches [15]–[18], and EOFS, OPMDC and OSA paradigms based on wavelength selective switches [19]–[21]. Without loss of generality and to simplify the explanation, a cluster unit deploying the flow-controlled and clock-distributed optical switch system as schematically illustrated in Fig. 1 has been used to show its principles. Each rack consists of H servers and each cluster groups N racks interconnected via the top of rack (ToR). An optical switch and the corresponding controller are used to interconnect all the racks via the data channels and label channels, respectively.

A. Label Control Mechanism

Field-programmable gate array (FPGA), benefitting from its flexible programmability, is utilized to implement the ToRs and optical switch controllers in the proposed system. Fig. 2 (a) schematically illustrates the functional blocks of

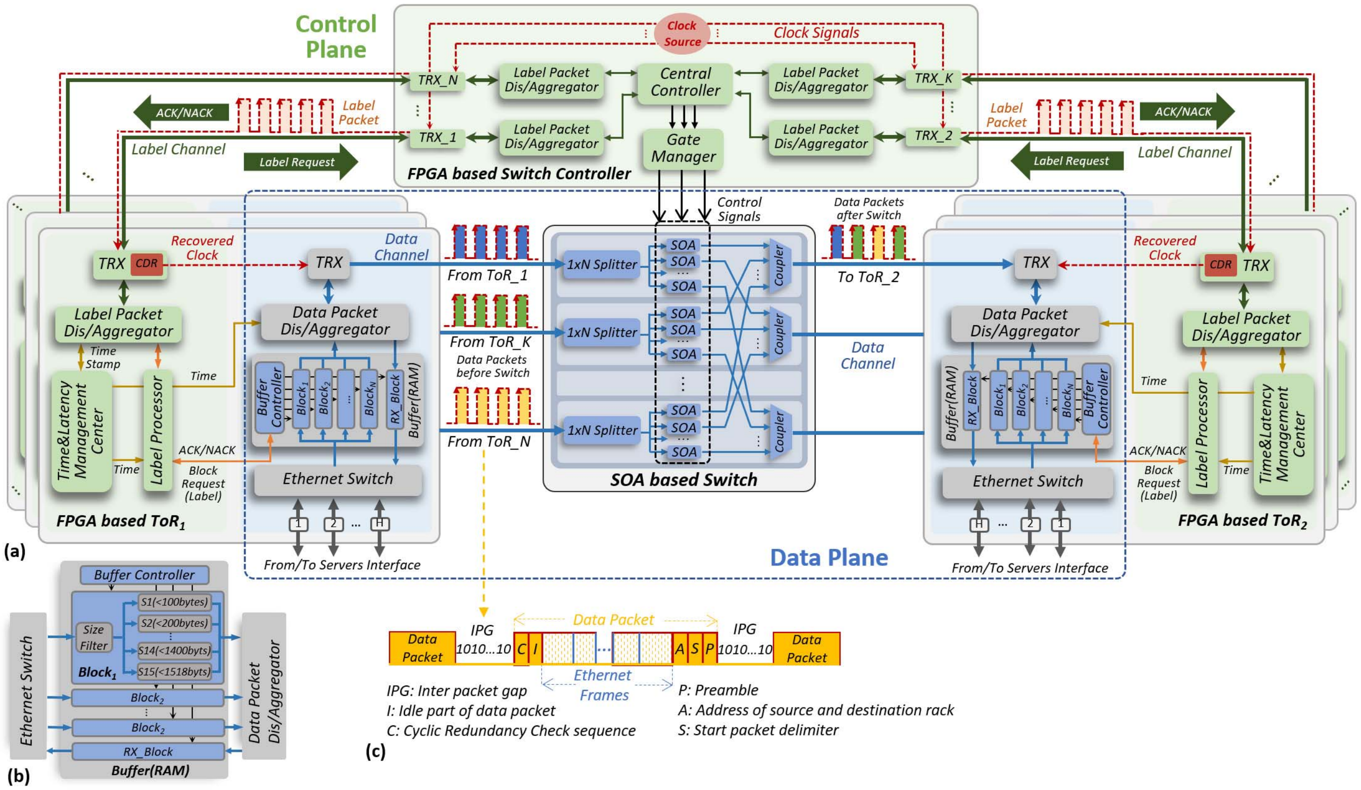

Fig. 2. (a) Functional blocks of the FPGA-based ToRs and switch controller in flow-controlled and clock-distributed optical switch and control system. (b) Details of the buffer block. (c) Components of the optical data packets.

the FPGA-based ToRs and switch controller. The Ethernet frames generated by servers in the same rack are first sent to and processed at the Ethernet switch in the ToR. Based on the destination MAC addresses, frames destining to servers in the identical rack (intra-rack traffic) are directly flowed to the intra-rack servers. While for frames destining to servers in other racks (inter-rack traffic), they are stored at the electrical buffer (RAM), where the buffer block stores the inter-rack traffic with the same rack destination, respectively. At every time slot, the copy of the Ethernet frames in the most-occupied buffer block will be selected and grouped to generate an optical data packet. The aggregated optical data packet at each ToR, as shown in Fig. 2(a), will be delivered to the optical switch via the data channel. Meanwhile, the serial number of the selected buffer block is processed as the label request packet, indicating the destination of the aggregated data packet. The forwarding priority of the data packet is also packaged in the label request packet for contention resolution at the switch. Via the label channel, the label request packet associated with the data packet is sent to the switch controller. Based on the received label requests from all the ToRs, the switch controller judges the packet contention and accordingly computes the optical-switch configuration in order to forward the optical data packets to the correct destinations. Note that the label request matrix can be processed within few nanoseconds to implement the fast switch control, benefiting from the parallel processing capability of FPGA-based switch controller.

The label control mechanism requires two independent channels (label and data channel) to implement the traffic control and forwarding, respectively. There would be a very small skew between the label control plane and data transmission plane. The different fiber length on these channels could further deteriorate this skew. To align the data packets and the corresponding label requests, the time of the edge nodes (ToRs) in the proposed switch and control system is accurately synchronized from the switch controller. At the initialization state of this system, the Time and Latency Management Center (Fig. 2 (a)) in each ToR sends the timestamp carrying the initial ToR time (T_{tx}) to the switch controller via label channels. Processed at the controller, the timestamps are sent back to the source ToRs, where the time (T_{rx}) of receiving the timestamp is recorded. Based on the known signal processing delay at the FPGA-based ToR and controller, and the time offset (T_{rx}-T_{tx}), the transmission delay on the label channels thus can be automatically calculated, even the length of the label channel fiber is various. Afterward, the central switch controller distributes its time to all the connected ToRs via the label channels. Compensated with the measured fiber delay and the FPGA processing delay, the time of all the ToRs is synchronized and coherent from the controller. The synchronized time is then used to align the transmission of the data packets and label request packets at each time slot to decrease the time skew.

Ethernet frames generated by servers with the same destination are stored in the identical buffer block at ToR. Each buffer block consists of several smaller buffers to store the frames with the similar length. As illustrated in Fig. 2(b), the Ethernet frames after the Ethernet switch will be first check by the Size

Filter to measure the frame length and accordingly sent to the corresponding smaller buffer. E.g, the frames of 148 bytes will be sent to the S2 buffer, which stores the frames with length from 100 bytes to 200 bytes. Thus, the Data Packet Aggregator can select the frames with suitable length to aggregate the optical data packet. Note that the length of the optical data packet in the proposed system can be flexible customized according to the network configurations. The components of the optical data packet are shown in Fig. 2(c). It consists of a 3 bytes preamble, 1 byte start packet delimiter, 4 bytes rack source/destination address, 4 bytes Cyclic Redundancy Check (CRC) sequence and the rest parts are the payload data (aggregated Ethernet frames). At the receiver side, the received optical data packet will be first processed at the Data Packet Dis-aggregator to extract the aggregated Ethernet frames which will be sent to the Ethernet Switch for packet forwarding in the destination rack.

B. Optical Flow Control (OFC) Technique

Packet contention occurs at the switch fabric whenever two or more packets from different source ToRs have the same output port at the same time slot. Due to the lack of optical buffer to store the packet that lost the contention, an Optical Flow Control (OFC) technique is developed and implemented between the ToR nodes and switch nodes to prevent the packet contention caused packet loss at the switch node. The Ethernet frames are stored in the electrical buffer at each ToR node and the copy of Ethernet frames in the most-occupied buffer block are selected to deliver at every time slot. According to the label control mechanism, the switch controller judges the packet connection and controls the packet forwarding based on the received label request matrix. Data packets with higher priority will be forwarded to the destination racks while the conflicted packets with lower priority will be forwarded to the racks with no destination requesting. This packet forwarding mechanism guarantees the receivers of data channels receiving the nonstop traffic flow at every time slot, being continuously active. After label request processing, the switch controller generates label response (Flow Control ACK/NACK) signals and sends back them to the corresponding ToRs on the bidirectional label channels, as shown in Fig. 2 (a). For the ToR receiving the ACK signal, which indicates the successful optical packet forwarding, the corresponding frames will be released from the electrical buffer block. Instead, if a ToR receives the NACK signal, which indicates the optical packet lost the contention and forwarded to the un-destined ToRs, the stored frames at the electrical buffer block will be retransmitted in the following time slot until the ToR receiving the ACK feedback. Thus, benefitting from the implementation of OFC technique, the conflicted packets that dropped at the undestined racks will be retransmitted in the following time slot to prevent the packet loss.

C. Clock Distribution Technique

Most of the locking time required by the clock and data recovery (CDR) is to adjust the variation in clock-frequency to sample the incoming data. Thus, even if the optical switch system takes nanoseconds reconfiguration time (including both hardware switching time and control overhead), the network throughput will still be very low due to slow CDR locking. To shorten the CDR time, the bidirectional label channels in our new approach are continuous links that are not only used to send the ACK/NACK signals from the switch controller to the ToRs, but they are also used to distribute the clock from the optical switch controller to ToRs to synchronize the system clock frequency. As shown in the FPGA-based switch controller of Fig. 2 (a), an onboard Clock Source is employed as a master clock to be distributed to the connected ToRs by the ACK/NACK signals. At each ToRs, the clock is recovered from the continuous ACK/NACK streaming by a conventional CDR receiver. The recovered clock (Rx-CLK) is then employed to drive the transceivers (TRXs) of the data channels. In this way, the clock with the same frequency is distributed and used in all the ToRs, not only to transmit the data packets and the label signals, but also to implement the nanoseconds recovery of the data packets. Indeed, once all the network ToRs have the clock with the same frequency, the receiver only needs to align the clock phase of the incoming data, which can be achieved within few tens of bits (few nanoseconds), preventing the need of a time-consuming clock frequency recovery.

Note that the CDR circuits at the conventional receivers need to receive continuous data traffic to maintain the recovered clock with good quality. To guarantee this, the optical switch controller, which has the full vision of the traffic from the ToRs, exploits the multicast capability of the optical switch to forward packets that lost contention to the un-destined ToRs to fill the empty slots. Additionally, the inter-packet gap (IPG) and idle parts of the packets, as well as the empty packets (a ToR has nothing at all to send) due to the lower traffic load, due to the lower traffic load are inserted with pulse transitions (“1010…1010”), as shown in the data packet pattern of Fig. 2 (c), to maintain the continuous stream of data, similar as in the Ethernet protocol. The data packet consists of several parts. First, the preamble consists of a sequence pattern of alternating 1 and 0 bits, allowing receivers on the data channel to easily synchronize their receiver’s clock phase, providing bit-level synchronization. The start packet delimiter is the eight-bit value that marks the end of the preamble, which is the first field of a data packet, and indicates the beginning of the packet. The address of source/destination rack is embedded for packet identification. The Cyclic Redundancy Check (CRC) is a 32-bit checksum calculated to provide error detection in the case of packet transmission collisions or link errors which could corrupt the data packet.

III. Experimental Results and Analysis

A. Experimental Setup

The set-up illustrated in Fig. 3 is built to experimentally demonstrate and assess the flow-controlled and clock-distributed optical switch and control system. The setup consists of 4 FPGA-based (Xilinx Virtex UltraScale+ VU9P) ToRs and each one equips with a 10.3125 Gb/s data channel to deliver the data packets. One SOA-based 4×4 optical

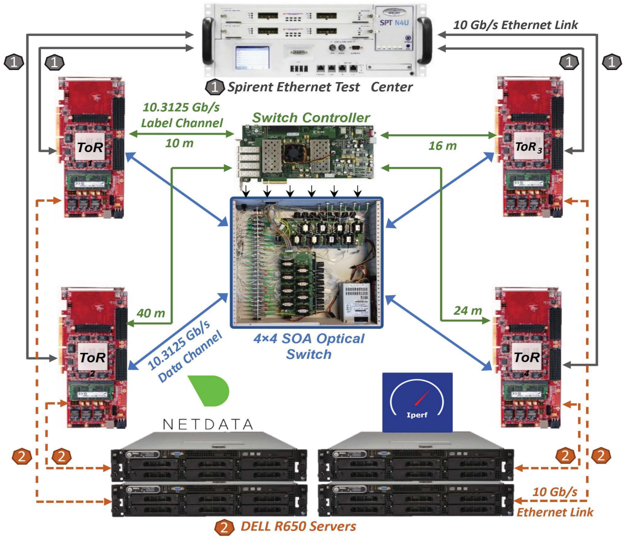

Fig. 3. Experimental set-up of flow-controlled and clock-distributed optical switch and control system deployed at the cluster unit with 4 ToRs.

switch [6] with corresponding FPGA-based (Xilinx VC709) controller are utilized to interconnect all the ToRs. Three buffer blocks, where the buffer size is controllable and variable due to the reprogramming capability of FPGA, are deployed inside each ToR. The FPGA-based ToR can monitor the occupation ratio (occupied bytes/overall buffer size) of each buffer block in real-time. At each time slot, the traffic stored in the most-occupied block will be selected to send out and the corresponding label request packet is also delivered to the switch controller via the label channel. The 10.3125 Gb/s label channels efficiently implement the label controlling mechanism, OFC technique and clock frequency distribution. The label signals and flow control signals (ACK/NACK) are transmitted between the FPGA-based ToRs and the FPGA-based switch controller. The fiber length between ToRs and switch nodes are different to validate the synchronized slot mechanism and the robustness of the switching system.

The experiment consists of two parts. First, to evaluate the fast label mechanism, OFC protocol and clock distribution as well as the system performance like the packet loss and latency, the Spirent Ethernet Testing Center emulating 8 servers (2 servers per rack) is connected in this setup, generating Ethernet frames at 10 Gb/s with controllable and variable load. The load here is defined as the ratio of average occupied bandwidth in a certain time on Ethernet link divided by the link capacity. Ethernet frames are generated between 64 and 1518 bytes with a controllable length (random or constant length). In DCN, 35% of the frames are control frames with size shorter than 200 bytes, and more than 45% frames are utilized to carry the application information with the size longer than 1400 bytes [22]–[24]. To match the practical network scenario, the length of the optical data packet in the experimental setup is set as 2600 bytes, in which it can aggregate one 1400 bytes frame, one 200 bytes frame and two frames with average length of 500 bytes. Note that the Spirent Ethernet Testing Center can also flexibly configure the traffic ratio of intra-rack and inter-rack by customizing the destination MAC address of generated frames. Second, to investigate the capability of running practical DCN applications, 4 servers (DELL R650) installing the *Netdata* and *iPerf* are connected in this setup to assess the real-time Transmission Control Protocol (TCP) traffic throughput performance [25, 26]. *iPerf* is a tool for creating TCP streams of specified bandwidth to various destinations in this setup. *Netdata* is distributed, real-time, TCP performance monitoring tool for systems and applications. *Netdata* has a highly efficient database storing long-term traffic-metrics for days, weeks, or months, all at 1-second granularity. Moreover, *Netdata* works in a master-slave mode, which helps the telemetry function implementation with distributed multiple servers. In this setup, *Netdata* is installed on each server to monitor the real-time TCP bandwidth at both source servers and destination servers.

B. Experimental Results and Analysis

1) Fast Switch Control and Packet Contention Resolution: The operation of the label controlling mechanism and the OFC technique have been first experimentally investigated to validate its effectiveness of nanoseconds switch control and packet contention resolution, respectively. Based on the label control mechanism, each ToR will deliver the label request signals, consisting of destination and forwarding priority of the associated data packets, to the switch controller at every time slot. As shown in Fig. 4, the FPGA-based switch controller judges the contention based on the received label signals and then accordingly sends the ACK (LabelResponse = LabelRequest) and NACK (LabelResponse $\neq$ LabelRequest) signal to the corresponding ToRs, respectively. Note that the priority order of data packets in this experimental case is set as '*1>2>3>4*' where packets with order '*1*' have the highest priority. In time slot N, the label requests of ToR$_1$ and ToR$_2$ is *3* that indicates the data packets from ToR$_1$ and ToR$_2$ are destined to same ToR$_3$, packet contention so that occurred at the optical switch. Given the higher priority, ToR$_1$ packet is forwarded to the correct destination ToR$_3$, while ToR$_2$ packet with lower priority is sent to ToR$_2$ to maintain the continuous traffic stream at receiver (the packet will be dropped at ToR$_2$, once verified that its correct destination is ToR$_3$ by checking its MAC address). According to the OFC technique, ToR$_1$ receives an ACK signal (see label response in Fig. 4) to release the stored packet in the buffer block, while ToR$_2$ receives the NACK signal to trigger the packet retransmission to prevent the packet loss. In next time slot *N+1*, ToR$_1$ sends out a new label request *2* while ToR$_2$sends again the label request *3* until receiving the ACK signal.

The monitored traffic traces in Fig. 4 show that the label request/response signals and the optical data packets are time-aligned at the switch controller and optical switch, respectively. This validates the accurate implementation of the synchronous slotted mechanism. The time of the ToRs is accurately synchronized from the switch controller in the proposed switch and control system to align the data packets and the corresponding label requests. Benefitting from this, the maximum skew between the switch control signal and data packet is 6.2 ns in the experimental setup with different fiber length

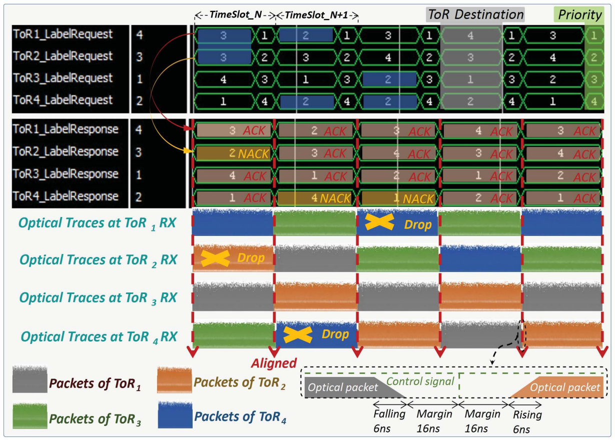

Fig. 4. Monitored label signals at the FPGA-based switch controller and received optical traces at ToRs. RX: receiver.

on the label channels. Thus, a short margin time between the switch control signal and optical packet is sufficient to absorb the skew and then to avoid the mismatching caused packet loss. As illustrated in Fig. 4, the overall switch and control time is 43.4 ns, which consists of 12.4 ns label processing time, 3 ns switch driver delay, 6 ns switch rising time, 6 ns switch falling time, and 16 ns margin time of the switch control signal with respect to the optical packet. Compared with the recent Micro-electromechanical systems (MEMS) switch based optical circuit switching network [27], [28], which has the microseconds magnitude of switch configuration time, the 43.4 ns switch and control time of the proposed system improves 2 orders of magnitude. Moreover, the fast switch and control mechanism allows a short (43.4 ns) inter-packet gap (IPG), further guaranteeing a high throughput.

2) Fast Clock and Data Recovery: To further improve the throughput, the clock and data recovery (CDR) needs to be completed as fast as possible. With this aim, the proposed switch and control system distributes the clock frequency from the optical switch controller to all the connected ToRs [28] to avoid the time-consuming frequency recovery. The empty slot is inserted with pulse transitions to maintain the receivers being active. Thus, the data can be correctly recovered in 3.1 ns, as it can be observed from the receiver data of RX_Data in Fig. 5. The receiver extracts the preamble (3 bytes) and the start packet delimiter (1 byte) within 1 clock cycle (3.1 ns). The 3.1 ns recovery time exploiting the frequency distribution is almost 30 times faster compared with the 92 ns recovery time achieved by the cost-high burst mode receiver [29]. Moreover, without the requirements of additional hardware, the clock distribution mechanism realizes fast CDR function based on the common commercial receivers. For this fast CDR scenario (point a in Fig. 5), the data channel throughput is 97.7%, where the data packet length is 2017 ns (2600 bytes), and the CDR time and the IPG is 3.1 ns and 43.4 ns, respectively. The CDR time of the system with no clock distribution is also measured as the reference. In contrast, around 906 ns are required to recover the correct data, mostly due to the time-consuming clock frequency adaption. Even with the short (43.4 ns) IPG, the network throughput is 51.7% (point c in Fig. 5), wasting half of the available bandwidth.

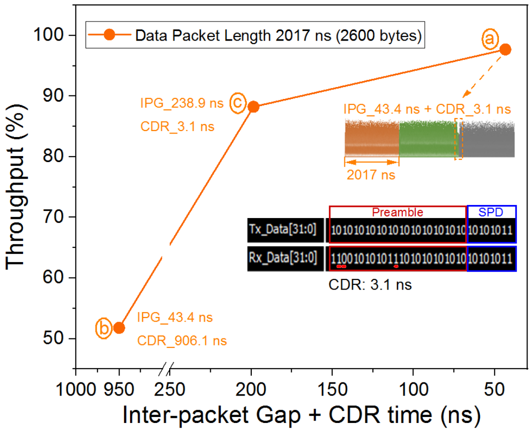

Fig. 5. Throughput against inter-packet gap and CDR locking time. SPD: start packet delimiter.

3) Packet Loss and Latency Performance: With the implementation of label control mechanism, optical flow control protocol and clock distribution, the system performance in terms of packet loss and end-to-end latency are also experimentally assessed under the various Ethernet frame size. The length (constant or random) of Ethernet frames can be customized at the Spirent Ethernet Testing Center. The average frame size is 792 bytes, randomly generated between 64 bytes and 1518 bytes. The size of each buffer block is 8192 bytes and size of aggregated data packet is 2600 bytes. The traffic ratio of intra-rack and inter-rack is set as 50%:50% to have an oversubscription of 1:1 at each ToR, where the Spirent emulates 2 servers with 20 Gb/s down (Ethernet) link and the 10 Gb/s up (optical) link. The loss of packets has two reasons: buffer overflow at ToR and packet contention at the optical switch node. Fig. 6(a) illustrates the packet loss performance as a function of the traffic load. The traffic load here is defined as the ratio of average occupied bandwidth on Ethernet links divided by the link capacity. The number ($\mathrm{C_{TX}}$) of sent out data packets to specific racks is recorded at each ToR and the received packets at the specific ToRs are counted ($\mathrm{C_{RX}}$) as well. Thus, the packet loss rate on the optical link can be calculated as ($\mathrm{C_{TX}}$ - $\mathrm{C_{RX}}$)/ $\mathrm{C_{TX}}$. Fig. 6(a) shows that the packet loss on the optical links is less than 3.0E-10 at the high load. When the load is less than 0.6, there is no packet loss on the optical links. The negligible packet loss validates the accurate operations of the OFC protocol which prevents the contention caused packet loss. The overall packet loss due to both the buffer overflow and the packet contention caused loss, are measured by the Spirent as well. As we can see from Fig. 6(a), the frame of 64 bytes has the highest overall packet loss. This is due to relatively more frames of small size are stuck by the head of line frame, subsequently being dropped once the buffer is overflowed. The system with high traffic load is more susceptible to lose the packet because traffics could be generated in bursts with a higher probability and be sent to a buffer block in a certain time at the high traffic load, which will cause the buffer overflowed.

The average server-to-server latency across the switch system has been also measured. First it is important to understand the fixed processing delay across the Spirent and FPGA-based

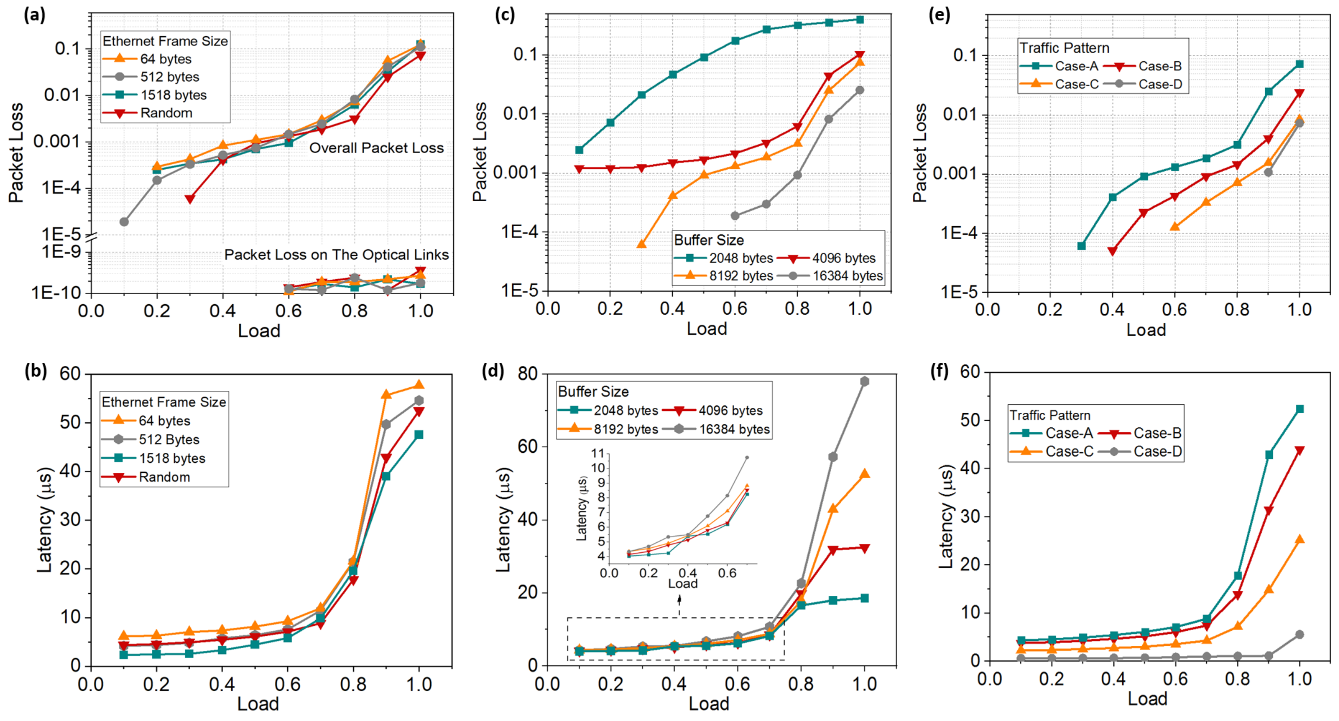

Fig. 6. Packet loss ratio and server end-to-end latency for different Ethernet frame size, buffer size and intra/inter-cluster traffic ratios.

TABLE I

DATA PROCESSING DELAY AT SPIRENT AND FPGA-BASED TOR

Processing Blocks	Delay (nanoseconds)
Spirent (TX+RX)	120
10G PCS/PMA (IP from Xilinx Ultrascale XCVU095)	
TX path	79
RX path	88
10G MAC (TX + RX)	228
Ethernet Switch (TX to RX)	36
Frame Buffering and Aggregating (Min/Max)	116/2348
Data Packet Generating	152
10G GTH (IP from Xilinx Ultrascale XCVU095)	
TX path	79.2
RX path	87.8
Data Packet Receiving	56
Frame Buffering and Disaggregating (Min/Max)	56/2288
Ethernet Switch (RX to TX)	36
Fiber Transmission Delay (64 meters)	320
Total (Min/Max)	1454/5310

Optical packet duration: 2017 ns; Tx buffer block size: 8192 bytes;
Rx buffer block size: 4096 bytes; Ethernet frame size: 1518 bytes.

TABLE II

LABEL PROCESSING DELAY AT FPGA-BASED TOR AND SWITCH CONTROLLER

Processing Blocks	Delay (nanoseconds)
Label Packet Generating (ToR)	25.6
10G GTH (IP from Xilinx Ultrascale XCVU095) (ToR)	
TX path	79.2
RX path	87.8
10G GTH (IP from Xilinx Virtex VC709) (Switch Controller)	
TX path	145.6
RX path	207.8
Label Packet Receiving (Switch Controller)	25.6
Contention Resolution and SOA Controlling	12.8
Label Packet Generating (Switch Controller)	19.2
Label Packet Receiving and Processing (ToR)	73.5
Total	677.1

ToR. This is summarized in Tab. 1. These values are measured using the traffic analyzer and the intermediate steps with a logic analyzer inside the FPGA. The latency components of the label processing time at the FPGA-based ToR and switch controller are also measured and shown in Tab. 2. To measure the end-to-end latency, the Spirent adds a timestamp at the start of each Ethernet frame and calculates the latency based on this timestamp when it receives this frame. The measured latency is shown in Fig. 6(b) for the various size of Ethernet frames. It is visible that the latency drops quickly when the load is below 0.7 for all frame sizes. This can be explained by frames having to wait the head of line frames at the high load, but the buffers will be mostly empty at the low load and frames can be processed immediately. There are more packets conflicted at the switch node when the traffic load is high, which introduces extra packet retransmission delay required by the OFC technique, increasing the server-to-server latency. This is also the reason for Ethernet frames of 64 bytes suffering the highest transmission latency because more frames need to be retransmitted with respect to the big size frames, once the contention happens.

The electrical buffer blocks store the Ethernet frames in case of retransmission to prevent the packet loss at the switch node. When the buffer is fully occupied, the newly arrived Ethernet frames will be discarded. Thus, the buffer size is one of the most critical hardware dimensions as larger buffer can reduce the buffer overflowed caused packet loss and in principle, would deteriorate the latency performance at higher load. Thus, we investigate the performances of the flow-controlled and clock-distributed switch and control system as a function of the buffer size. The various size of buffer blocks (2048 bytes, 4096 bytes, 8192 bytes and 16384 bytes) are implemented in the FPGA-based ToR. To emulate the real DCN operating environment, the Ethernet frames with random length (between 64 bytes and 1518 bytes with an average size of 792 bytes) are generated and the traffic ratio of intra-rack and inter-rack is set as 50%:50%. As clearly shown in Fig. 6(c)(d), the larger buffer improves the packet loss performance but at the expense of slightly larger latency. The packet loss is unavoidable for load higher than 0.8, and a larger buffer size does not help to significantly improve the performance. This is due to the heavy traffic congestion and full buffer occupation. The latency performance becomes worse for larger buffer because of the longer queuing time and the extra transmission delay caused by the packet retransmission.

The system performance under different intra-/inter-rack traffic ratios are also analysed. The destination of the generated frames can be flexibly configured in the Spirent to set the various traffic ratios. Four traffic cases have been studied: A) 50% intra-rack traffic and 50% inter-rack traffic (case-A); B) 65% intra-rack traffic and 35% inter-rack traffic (case-B). C) 85% intra-rack traffic and 15% inter-rack traffic (case-C); D) 100% intra-rack traffic and 0% inter-rack traffic (case-D). For the other critical hardware parameters, the buffer of 8192 bytes each block is deployed in FPGA-based ToR and the Ethernet frames with random length are generated in the Spirent. Fig. 6(e)(f) shows the packet loss and server end-to-end latency performance as a function of the traffic load. As expected, the packet loss and latency performance degrade at higher load, due to the high traffic congestion and buffer occupancy. The increasing ratio of inter-rack traffic could also deteriorate the system performance because the data packet contention caused retransmission and high buffer occupation ratio. It can be seen that these four different intra-/inter-cluster traffic ratios give stable latency performances when the load is lower than 0.7. Zero packet loss and lower than 3 μs server end-to-end latency are achieved for case-C (85% intra-rack traffic and 15% inter-rack traffic) with a load of 0.5. The latency is mainly contributed by the processing and buffering time at the ToR, propagation delay and the retransmissions for the blocked data packet. Note that the traffics in data centers do not exceed 30% of the maximum network capacity for most of the time [30], [31], the proposed flow-controlled and clock-distributed switch and control system can effectively handle the traffic, featuring the fast switch control and data recovery capability. The performance can be further improved by employing multiple receivers and bundles of parallel fibers on the data channels in order to forward and receive more packets at the same time.

4) TCP Throughput: Transmission Control Protocol (TCP) was designed to ensure delivery of all packets and minimize packet loss. It is a connection-oriented protocol which ensures the quality of service by re-transmitting packets until all packets are received correctly. Thus, TCP throughput is extremely sensitive to network packet loss performance. Since the TCP protocol is widely used to DCN applications, such as video streaming services and mission critical applications, the TCP throughput that dictates the network performance is an ideal parameter to evaluate the capability of running practical applications in the proposed optical switch system.

Four servers (removing the Spirent related connections) installing the *Netdata* as the network performance monitor and *iPerf* as the TCP traffic generator are employed in this setup to investigate the capability of running practical DCN applications (through TCP throughput monitoring). The four servers are connected to four ToRs, as shown in Fig. 3, and the server-generated TCP traffics are aggregated as the optical data packets at each ToR. The flow-controlled and clock-distributed optical switch system is deployed between the FPGA-based ToRs and optical switch node. Server_1, Server_2 and Server_3 work as the traffic source in this setup, delivering the TCP traffic to the destination Server_4 on the common 10.3125Gb/s data channel. *iPerf* deployed in the source servers generates as much TCP traffic as possible to reach the maximum TCP throughput. Because the traffics from these 3 servers share the 10.3125Gb/s data channel, only the traffics with the highest priority can reach the full (maximum) throughput of 3.25 Gb/s based on the theoretical calculating. The priority order of TCP packets in this setup is set as 'Server_1> Server_2> Server_3', where packets from Server_1 have the highest priority. The distributed *Netdata* is used to monitor the real-time traffic trace of both the source servers (Server_1, Server_2 and Server_3) and the destination server (Server_4), respectively. The monitored traffic throughput in 120 seconds is illustrated in Fig. 7. Server_1 has a stable and full throughput of 3.25 Gb/s, in line with the theoretical value. Due to the packet contention, the throughput of Server_2 and Server_3 is not stable, but the average throughput is still high than 2.5 Gb/s for Server_3 with the lowest traffic priority. The throughput of the Server_4 is the throughput sum of the three source servers, indicating no packets loss on the optical links.

5) Scalability and Capacity Discussion: The proposed label control mechanism, OFC technique and clock distribution are implemented on the parallel label channels at the cluster network, being independent of the network scale. As we have investigated, large-scale optical DCN with distributed clusters can be built based on the proposed flow-controlled and clock-distributed optical switch and control system in [32].]. The optical DCN is divided into intra-cluster interconnect network and inter-cluster interconnect network, where the i-th inter-cluster optical switch is employed to connect i-th ToR locating in different clusters. The intra-cluster network and inter-cluster network are independent sub-networks with

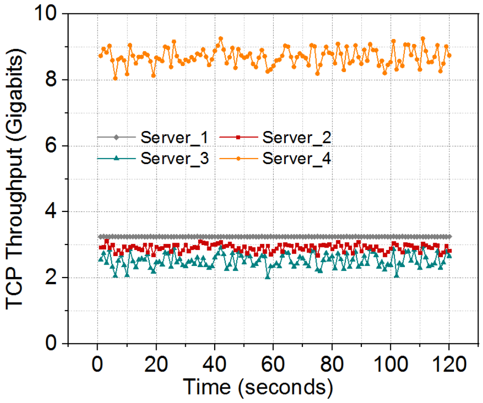

Fig. 7. Real-time TCP traffic monitored at servers for 120 seconds.

their own flow-controlled and clock-distributed optical switch and control system. This enables the full distribution and scalability of the proposed switching architecture and control system to implement a large-scale DCN with a very large number of servers. Scalability investigation of the proposed switching architecture and control system was validated with limited (11%) performance degradation as the network scale from 2560 to 40960 servers [32].

The transparent SOA-based optical switch is independent of the bit-rate and data-format to support high data rate channels in the proposed system. In addition, Wavelength Division Multiplexing (WDM) technology can be employed to boost the switching capacity at a superior power-per-unit bandwidth performance. Thus, the proposed switch and control system can scale to a higher capacity in two ways. 1) Adding more WDM transceivers at the data channels. 2) Deploying high-speed optical transceivers. The label control mechanism and OFC technique are implemented on the independent label channels with lower bandwidth requirements compared with the data channels. Thus, each label channel can control the optical switch to switch the high bandwidth data channel generated by more transceivers or by high-speed transceivers. Thus, even more transceivers are deployed at the data channels, a relatively small number of label channels are sufficient to guarantee the fast switch control and stable packet contention resolution. Moreover, the clock is distributed on the label channel to drive the transceivers of the data channels. This clock can be carried at a lower data rate with respect to the higher data rate transceivers to be deployed at the data channels. At the CDR block, the frequency of the distributed clock will be accordingly multiplicated to adapt to the higher data rate of data channel transceivers. Thus, the existing flow-controlled and clock-distributed optical switch and control system can support the high-speed optical transceivers without equipment updating.

IV. Conclusion

We have developed and experimentally demonstrated for the first time a flow-controlled and clock-distributed optical switch and control system for optical DCNs based on the novel label control mechanism, OFC technique, and the clock distribution mechanism. Optical label channels deliver the label signals for nanoseconds packets forwarding, implement the OFC technique signals for optical packet contention resolution, and also deliver the clock distribution to all ToRs for a nanoseconds CDR locking. Experimental results confirm an overall 43.4 ns optical switch and network control time, 3.1 ns data recovery, and a less than 3.0E-10 packet loss rate due to the OFC based packet contention resolution. Real servers, running the TCP traffic generating and monitoring tools, are connected by the proposed optical switch and control system, validating the capability of running the practical applications in the optical DCNs with full TCP bandwidth. 3 μs server end-to-end latency and zero buffer overflow caused packet loss are achieved at the traffic load of 0.5. Those results enable the practical deployments of high bandwidth optical switches in DCNs.

Disclosures

The authors declare no conflicts of interest.

References

[1] A. Ghiasi, "Large data centers interconnect bottlenecks," *Opt. Exp.*, vol. 23, no. 3, pp. 2085–2090, 2015.

[2] H. J. S. Dorren, E. H. M. Wittebol, R. D. Kluijver, G. G. D. Villota, P. Duan, and O. Raz, "Challenges for optically enabled high-radix switches for data center networks," *J. Lightw. Technol.*, vol. 33, no. 5, pp. 1117–1125, Mar. 1, 2015.

[3] F. Testa and L. Pavesi, *Optical Switching in Next Generation Data Centers*. Springer, Spring 2017.

[4] N. Parsons and N. Calabretta, "Optical switching for data center networks," in *Springer Handbook of Optical Networks*. Springer, Spring 2020, pp. 795–825.

[5] M. Fiorani, S. Aleksic, M. Casoni, L. Wosinska, and J. Chen, "Energy-efficient elastic optical interconnect architecture for data centers," *IEEE Commun. Lett.*, vol. 18, no. 9, pp. 1531–1534, Sep. 2014.

[6] X. Xue, K. Prifti, B. Pan, F. Yan, X. Guo, and N. Calabretta, "Fast dynamic control of optical data center networks based on nanoseconds WDM photonics integrated switches," in *Proc. 24th OptoElectron. Commun. Conf. (OECC) Int. Conf. Photon. Switching Comput. (PSC)*, Jul. 2019, pp. 1–3.

[7] E. N. Lallas, "A survey on all optical label swapping techniques: Comparison and trends," *Opt. Switching Netw.*, vol. 31, pp. 22–38, Jan. 2019.

[8] H. Ballani *et al.*, "Bridging the last mile for optical switching in data centers," in *Proc. Opt. Fiber Commun. Conf. (OFC)*, 2018, pp. 1–3.

[9] X. Xue *et al.*, "Experimental assessments of SDN-enabled optical polling flow control for contention resolution in optical DCNs," *J. Lightw. Technol.*, vol. 39, no. 9, pp. 2652–2660, May 1, 2021.

[10] M. Moralis-Pegios, N. Terzenidis, G. Mourgias-Alexandris, K. Vyrsokinos, and N. Pleros, "A low-latency high-port count optical switch with optical delay line buffering for disaggregated data centers," in *Optical Interconnects XVIII*, vol. 10538. Bellingham, WA, USA: SPIE, 2018, Art. no. 1053805.

[11] R. Farhat, A. Farhat, and M. Menif, "All-optical variable-length packet router with contention resolution based on wavelength conversion," in *Nonlinear Optics and Applications X*, vol. 10228. Bellingham, WA, USA: SPIE, 2017, Art. no. 102280X.

[12] B. Nleya and A. Mutsvangwa, "A node-regulated deflection routing framework for contention minimization," *J. Comput. Netw. Commun.*, vol. 2020, pp. 1–14, Jun. 2020.

[13] P. J. Argibay-Losada, D. Chiaroni, and C. Qiao, "Optical packet switching and optical burst switching," in *Springer Handbook of Optical Networks*. Springer, Spring 2020, pp. 665–701.

[14] K. Clark *et al.*, "Sub-nanosecond clock and data recovery in an optically-switched data centre network," in *Proc. Eur. Conf. Opt. Commun. (ECOC)*, Sep. 2018, pp. 1–3.

[15] F. Yan, X. Xue, and N. Calabretta, "HiFOST: A scalable and low-latency hybrid data center network architecture based on flow-controlled fast optical switches," *J. Opt. Commun. Netw.*, vol. 10, no. 7, pp. 1–14, Jul. 2018.

[16] X. Xue, F. Yan, B. Pan, and N. Calabretta, "Flexibility assessment of the reconfigurable OPSquare for virtualized data center networks under realistic traffics," in *Proc. Eur. Conf. Opt. Commun. (ECOC)*, Sep. 2018, pp. 1–3.

[17] W. Miao, F. Yan, and N. Calabretta, "Towards Petabit/s all-optical flat data center networks based on WDM optical cross-connect switches with flow control," *J. Lightw. Technol.*, vol. 34, no. 17, pp. 4066–4075, Sep. 1, 2016.

[18] Y. Yin, R. Proietti, X. Ye, C. J. Nitta, V. Akella, and S. Yoo, "LIONS: An AWGR-based low-latency optical switch for high-performance computing and data centers," *IEEE J. Sel. Topics Quantum Electron.*, vol. 19, no. 2, Mar./Apr. 2012, Art. no. 3600409.

[19] M. C. Yuang *et al.*, "OPMDC: Architecture design and implementation of a new optical pyramid data center network," *J. Lightw. Technol.*, vol. 33, no. 10, pp. 2019–2031, May 15, 2015.

[20] Z. Zhang, W. Hu, W. Sun, L. Zhao, and K. Zhang, "Elastic optical ring with flexible spectrum ROADMs: An optical switching architecture for future data center networks," *Opt. Switching Netw.*, vol. 19, pp. 1–9, Jan. 2016.

[21] K. Chen *et al.*, "OSA: An optical switching architecture for data center networks with unprecedented flexibility," *IEEE/ACM Trans. Netw.*, vol. 22, no. 2, pp. 498–511, Apr. 2014.

[22] R. Sinha, C. Papadopoulos, and J. Heidemann, "Internet packet size distributions: Some observations," USC/Information Sci. Inst., Marina del Rey, CA, USA, Tech. Rep., ISI-TR-2007, vol. 643, 2007, pp. 1276–1536.

[23] T. Benson, A. Anand, A. Akella, and M. Zhang, "Understanding data center traffic characteristics," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 40, no. 1, pp. 92–99, 2010.

[24] T. Benson, A. Akella, and D. A. Maltz, "Network traffic characteristics of data centers in the wild," in *Proc. 10th Annu. Conf. Internet Meas. (IMC)*, 2010, pp. 267–280.

[25] NETDATA. *Monitor Everything in Real Time*. Accessed: Mar. 2021. [Online]. Available: https://www.netdata.cloud/

[26] Iperf. *What is iPerf*. Accessed: Jun. 2020. [Online]. Available: https://iperf.fr/

[27] A. Y. Takabayashi *et al.*, "Broadband compact single-pole double-throw silicon photonic MEMS switch," *J. Microelectromech. Syst.*, vol. 30, no. 2, pp. 322–329, Apr. 2021.

[28] Y. Liu, J. Liu, B. Yu, and X. Liu, "A compact single-cantilever multicontact RF-MEMS switch with enhanced reliability," *IEEE Microw. Wireless Compon. Lett.*, vol. 28, no. 3, pp. 191–193, Mar. 2018.

[29] N. Cheng *et al.*, "Multi-rate 25/12.5/10-Gb/s burst-mode upstream transmission based on a 10G burst-mode Rosa with digital equalization achieving 20 dB dynamic range and sub-100 ns recovery time," in *Proc. Eur. Conf. Opt. Commun. (ECOC)*, Dec. 2020, pp. 1–3.

[30] A. Roy, H. Zeng, J. Bagga, G. Porter, and A. C. Snoeren, "Inside the social network's (datacenter) network," in *Proc. ACM Conf. Special Interest Group Data Commun.*, Aug. 2015, pp. 123–137.

[31] A. Singh *et al.*, "Jupiter rising: A decade of Clos topologies and centralized control in Google's datacenter network," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 45, no. 4, pp. 183–197, 2015.

[32] X. Xue *et al.*, "ROTOS: A reconfigurable and cost-effective architecture for high-performance optical data center networks," *J. Lightw. Technol.*, vol. 38, no. 13, pp. 3485–3494, Jul. 1, 2020, doi: 10.1109/JLT.2020.3002735.

Xuwei Xue is currently a Post-Doctoral Researcher with the Electro-Optical Communication System Group, Eindhoven University of Technology (TU/e). His research interests include FPGA implementation of scheduling algorithms for low latency control of fast and large port optical switches, develop of optical interconnecting architectures, and develop of the software defined networking (SDN) control techniques for optical networks.

Bitao Pan is currently a Ph.D. Researcher with the Electro-Optical Communication Group, Eindhoven University of Technology. His research interests include flexible optical network systems, software defined networking, network function virtualization, and FPGA-based traffic engineering and optical devices control.

Xiaotao Guo is currently a Ph.D. Researcher with the Electro-Optical Communication Group, Eindhoven University of Technology. His research interests include data center networks, software defined networking, and resource allocation algorithm.

Nicola Calabretta is currently an Associate Professor with the Electro-Optical Communication Systems and a Senior Research Fellow at the Eindhoven University of Technology (TU/e). His expertise is in telecommunications engineering, electrical engineering, and optical engineering. His research focuses on optical signal processing for highly spectral efficient multi-level modulation formats, high-speed electronics for processing of novel labeling techniques, FPGA implementation of scheduling algorithms for low latency control of large port optical switches, optical interconnects, photonic integrated wavelength selector with flexible bandwidth, photonic integrated optical switches, and c optical nodes for metro networks. Applications of the group's work on fast optical switches include performance enhancement for optical metro/access networks and data center networks.